Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Structured Decomposition-Based Hybrid Metaheuristic for Multi-Objective Resource Optimization in Healthcare Edge Networks

Dr. Avijit Mondal

avijit.mondal@galgotiasuniversity.edu.in

Galgotias University, Greater Noida, UP https://orcid.org/0009-0002-1016-2778

Research Article

Keywords: Edge computing, Mobile edge computing (MEC), Resource allocation, Task offloading, Energy efficiency, Healthcare informatics, Hybrid optimization, Particle swarm optimization (PSO), Simulated annealing (SA), Healthcare edge networks

Posted Date: July 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10287631/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Structured Decomposition-Based Hybrid Metaheuristic for Multi-Objective Resource Optimization in Healthcare Edge Networks

Avijit Mondal[1*], Priyanka Chatterjee[1†] and
Pinaki Sankar Chatterjee (Member, IEEE)[2†]

[1*]School of Computer Science and Engineering, Galgotias University, Greater Noida, 203201, Uttar Pradesh, India.
[2]School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar, 751024, Odisha, India.
[1]Bagchi Sri Shankara Cancer Centre and Research Institute (BSSCCRI), Bhubaneswar, 752054, Odisha, India.

*Corresponding author(s). E-mail(s): avijit.mondal@galgotiasuniversity.edu.in;
Contributing authors: secy.admin@bagchishankara.org; pinakifcs@kiit.ac.in;
†These authors contributed equally to this work.

Abstract

As healthcare edge networks are expected to continuously monitor patients and respond to emergencies, they require real-time and energy-efficient resource management solutions. Most existing techniques address only one aspect of the problem, such as power management, workload migration, or system welfare, and are not sufficient to tackle the variations in the workload of the healthcare setting. We propose a new technique, termed HOA-PWM, to address these problems by combining two approaches: Simulated Annealing (SA) and Particle Swarm Optimization (PSO). The proposed method jointly optimizes power allocation, task offloading, and scheduling in a way that improves overall system welfare. The model uses a type of communication called NOMA to manage delays and follows strict rules for healthcare service quality. Tests done with realistic numbers of users (from 32 to 500) and stressful situations (up to 5000 tasks with 1000 users) show that HOA-PWM uses up to 15% less energy and gives up to 12%

better system welfare than other methods. The system also works quickly, with the time it takes to finish scaling in line with the number of users. It remains effective even when key factors change by up to 20%, with statistical significance confirmed across 100 independent trials ($\boldsymbol{p < 0.05}$).

Keywords: Edge computing, Mobile edge computing (MEC), Resource allocation, Task offloading, Energy efficiency, Healthcare informatics, Hybrid optimization, Particle swarm optimization (PSO), Simulated annealing (SA), Healthcare edge networks

1 Introduction

Healthcare-focused high density edge networks (HDEN) are fundamentally different from the traditional mobile edge computing environment. In clinical settings, workloads are by their nature non-stationary and can increase suddenly during emergencies, epidemic outbreaks or peak operating hours. At the same time, wearable and edge nodes operate on a strict energy budget, and many clinical applications impose strict delays below which physiological data may lose their diagnostic relevance. These characteristics introduce close-knit energy-delay compromises, which are more important than in the traditional deployment of MECs. As a result, resource allocation strategies in healthcare edge networks need to balance energy efficiency, resilience and clinical prioritization in highly dynamic circumstances [1].

Figure 1 illustrates a three-tier architecture for healthcare edge computing, which includes IoT medical devices, edge servers, 5G base stations, and remote cloud data centers. Latency-sensitive emergency data flows require attention and reliable processing within tight deadlines, which further increases the complexity of coordinating resources.

Heuristic methods such as particle swarm optimization (PSO) and variants of genetic algorithms (GA) provide faster convergence [2]. However, in conditions of ultra-dense or highly dynamic networks, single meta-metrics often suffer from premature convergence and lack of exploration capacity. Their scalability declines as user density increases, interference coupling increases, and combined options for offloading increase.

The above shortcomings justify a structured multi-objective optimization framework that can handle mixed continuous and binary decision variables, maintain exploration in non-convex environments, ensure convergence stability in dense healthcare environments and avoid the computational overhead and training delays of learning methods.

Although metaheuristic methods are well suited to mixed integer non-convex problems, single-step algorithms usually struggle to maintain both global exploration and accurate convergence in large edge network scenarios [3]. To address these challenges, this paper proposes a structured decomposition-based hybrid meta-heuristic framework, called HOA-PWM.

The system model includes NOMA-based communication, health-specific quality of service (QoS) constraints and differentiated levels of patient priority. We formulate

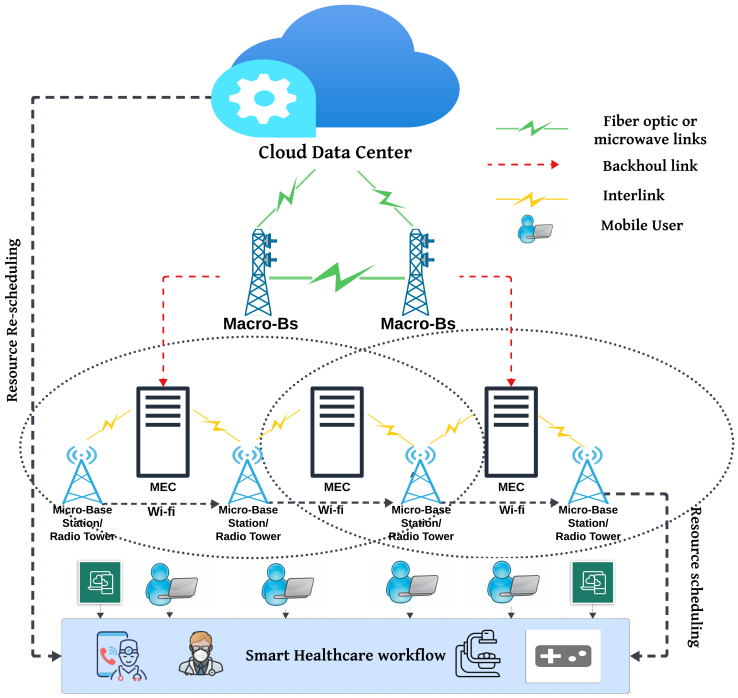


Fig. 1 Three-tier healthcare edge-computing architecture comprising: (a) IoT medical devices (wearables and sensors), (b) edge servers co-located with 5G base stations, and (c) cloud data centres. Red dashed lines denote latency-sensitive paths for emergency data transmission.

a mixed-integer multi-objective optimization problem jointly minimizing energy and maximizing welfare under NOMA interference and healthcare QoS constraints.

The rest of this paper is organized as follows. Section 2 reviews related studies. Section 3 presents the proposed HOA-PWM methodology and its structural justification. Section 5 describes the simulation environment and parameters. Section 6 discusses the experimental results and comparative analysis. Finally, Section 7 concludes the paper.

2 Literature Survey

Task offloading and resource allocation in Mobile Edge Computing (MEC) have been widely studied to enable low-latency and energy-efficient processing in IoT-enabled systems. Classical MEC formulations primarily focus on minimizing delay, reducing energy consumption, or balancing the two under moderate traffic conditions. However, healthcare-oriented high-density edge networks introduce additional structural and operational constraints that significantly complicate optimization.In clinical environments, large numbers of heterogeneous medical devices operate concurrently. Workloads are highly non-stationary, with sudden surges during emergencies or peak hospital activity. Many medical tasks are latency-critical and directly related

to patient safety, while strict data protection regulations impose additional system constraints [4, 5]. As a result, methods developed for generic MEC scenarios often fail to satisfy the reliability, scalability, and priority-awareness requirements of healthcare deployments.

Early MEC research relied heavily on heuristic and meta-heuristic energy efficiency reduction or savings methods. For example, Chen et al. [6] proposed a multi-user loading strategy to reduce latency in low traffic conditions. However, performance has decreased significantly with increasing network density, indicating scalability constraints in congested environments.Particle swarm optimization (PSO) approaches [2, 7, 8] are attractive due to their simplicity and rapid convergence. However, the standard PSO assumes a fairly stable landscape for the search. In ultra-dense or dynamically changing healthcare networks, high interconnections and workloads increase the risk of premature convergence towards sub-optimal solutions.

Genetic algorithms-based methods [9] increase the variety of search, but cause higher computational costs and are less suitable for real-time medical decision-making. Hybrid GA-PSO frameworks [10] seek to balance exploration and exploitation, but most of these approaches are focused mainly on delays and energy performance, with no health or social priorities or targets. DRL-based methods are adaptive in their decision making and show promising results in controlled environments. However, several limitations limit their applicability to healthcare edge networks, such as high training data and computing requirements, slow convergence during the learning phase, limited interpretation of educational policy and the need for updates of the model on a continuous basis.

In the healthcare sector, resource allocation mechanisms need to be predictable, transparent and efficient. These requirements limit the practicality of intensive learning approaches in wearables and edge deployments [11]. Some studies focus specifically on health-care edge systems. Federal frameworks based on learning [12] improve privacy by keeping patient data locally. While these approaches are beneficial to confidentiality, they often place data privacy above the overall optimization of resources. Similarly, health analytics frameworks [13–15] allow for health monitoring and diagnosis, but are usually application-specific and do not address the large-scale allocation of mixed-integer resources in dense environments.

A wider limitation of current research is the lack of fully integrated multi-objective approaches to capture energy consumption, compliance with health requirements, patient prioritization and system-wide welfare. Many work optimizes only one or two metrics, while others rely on computationally intensive learning frameworks. Moreover, most hybrid metaheuristics use a uniform encoding of continuous and discrete variables, which may decrease the stability of convergence in large mixed-integer spaces. Two key research gaps are therefore identified, firstly the lack of coherent, multi-dimensional, hybrid meta-economic models that integrate energy efficiency, health-care-related safeguards, and training without over-training.

To address these limitations, this paper proposes HOA-PWM, a hybrid metaheuristic structured by a decomposition scheme, which separates continuous power optimization and discrete task offloading decisions into specific optimization phases. This approach improves the stability and convergence behavior of the high-density

healthcare edge networks while maintaining computational tractability. Table 1 summarizes representative approaches of the MEC and HDEN schemes in terms of health information, multi-objective integration and scalability.

Table 1 Comparative summary of representative MEC and HDEN task offloading approaches

Approach	Focus	Healthcare-Aware	Multi-Obj.	Scalability
PSO-based MEC Offloading [7, 8]	Latency Minimization	No	No	Limited
GA based Task Scheduling [9]	Latency Optimization	No	No	Limited
Hybrid GA/PSO Framework [10]	Energy + Latency	No	Partial	Moderate
DRL based Offloading (DQN) [16, 17]	Delay + Energy	Partial	Partial	Limited (Training Overhead)
Healthcare focused Federated Learning [12]	Privacy-Aware Offloading	Yes	No	Moderate
Edge enabled Healthcare Framework [13–15]	Medical Data Analytics	Yes	No	Limited (Application Specific)
Energy Efficient Edge/Cloud Offloading [3]	Energy + Delay	No	Partial	Moderate
DRL + SQP-based Offloading [18]	Delay + Energy	No	Partial	Vehicular MEC
HOA-PWM (Proposed)	Energy, Delay, and User Welfare	Yes	Yes	High (HDEN)

Existing hybrid approaches do not structurally decompose continuous power allocation and discrete offloading decisions into sequential optimization phases, which motivates the proposed framework.

3 HOA-PWM: The Proposed Framework

Fig. 2 shows the general workflow for the proposed HOA-PWM. The algorithm follows a structured decomposition-based three-phase optimization process to address the mixed integer nature of the problem. The common optimization goal includes both the constant transmission power variables and the binary offloading decisions, which are non-linearly bound by NOMA interference and delays. Instead of performing a uniform search in a high-dimensional manifold, HOA-PWM partitions optimization into specialized phases, tailored to the structural properties of each subspace.

3.1 Phase I: Global continuous exploration through simulated annealing

This phase uses simulated annealing (SA) to explore the continuum transmission energy domain. Due to interference coupling in NOMA-based high-speed connections, the energy landscape is very non-convex with multiple local minima.SA introduces probabilistic uplift, controlled by a cooling schedule, which allows for escaping from

poor local basins. This phase gives priority to global research and identifies promising areas in the continuum solution space. The assessment of fitness at this stage focuses mainly on transmission capacity, while enforcing the constraints of energy and QoS feasibility. The output of this phase is a globally competitive power configuration, which serves as an informed starting point for the next stage of development.

3.2 Phase II: Continuous refinement by inertia-weighted particle swarm optimization

Following a global search, IWPSO is refining the allocation of transmission energy. While SA is increasing the exploration coverage, it is slowly approaching optimum. The IWPSO addresses this limitation by allowing cooperative swarm-based application of each particle as a candidate energy vector, and speed updates are controlled by inertia decay, cognitive learning, and social learning. A decreasing inertia mass improves convergence stability while maintaining a moderate exploratory capability. This step increases the accuracy of convergence in continuous subspace and reduces the variance of solutions without re-introducing premature stagnation.

3.3 Phase III: Discrete Offloading Optimization via Binary PSO

Once the transmission capacity has been stabilized, BPSO determines the decision to offload the task. Each particle encodes a binary vector that indicates whether the job is being handled locally or being handled on an edge server. Unlike continuous optimization, the job to be handled is combinatorial and discontinuous. By separating discrete optimization from continuous tuning, the algorithm avoids the cross-dimensional interference that typically breaks down convergence in a unified hybrid meta-surrogate. The fitness function evaluates energy consumption under refined energy allocation constraints and patient priority levels in the healthcare formulation.

The three phases SA, IWPSO and BPSO are linked in a sequential manner. This structured progression improves the quality of the solution incrementally while maintaining computational throughput. By breaking down continuous and discrete decision variables into specific optimization phases, HOA-PWM improves stability, scalability in dense environments and resilience to workload changes.

4 Problem Formulation

The goal of this study is to jointly minimize total energy consumption and maximize overall system welfare, while satisfying healthcare quality-of-service (QoS) requirements. The optimization problem is formulated as:

$$\min_{\mathbf{p},\mathbf{x}} \ \alpha E_{\text{total}}(\mathbf{p}) - \beta W(\mathbf{p}, \mathbf{x}) \tag{1}$$

Subject to constraints (C1)–(C7) defined in Eq. 14. In this expression, $\mathbf{p} = [p_u], \quad p_u \in [0, p_{\max}]$ denotes the power allocation matrix for transmission, which specifies the energy assigned to each user on each channel. The matrix $\mathbf{x} = [x_{qn}]$ represents

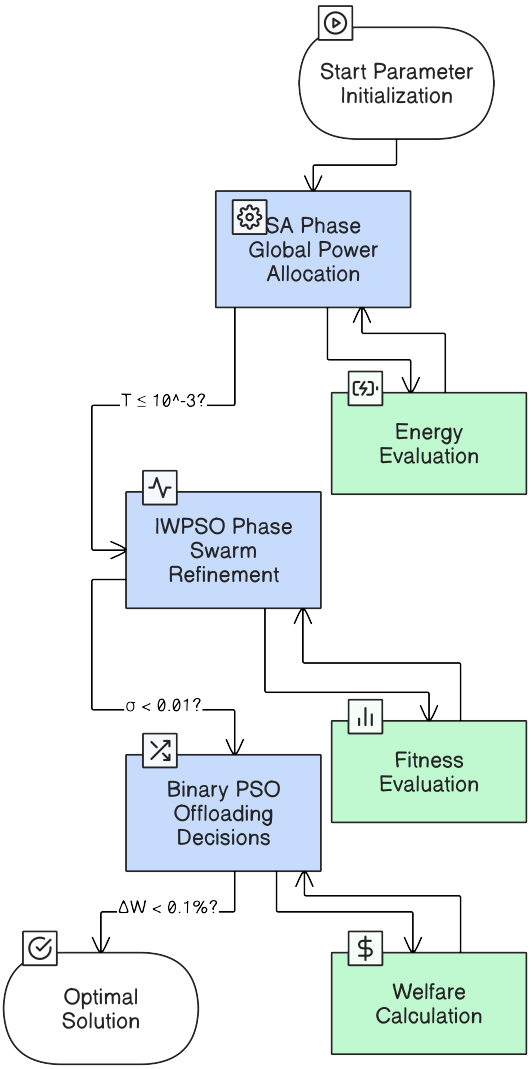


Fig. 2 Overall workflow of the proposed HOA-PWM framework, illustrating the structured sequential optimization stages: SA-based global power exploration, IWPSO-based continuous refinement, and BPSO-based discrete task offloading.

the binary offloading decision variable. If $x_{qn} = 1$, the job is executed on the edge server; if $x_{qn} = 0$, the job is offloaded to the cloud. The function $E(\mathbf{p}, \mathbf{x})$ represents the total energy consumption of the system (Eq. 5). The function $W(\mathbf{p}, \mathbf{x})$ represents the welfare of the system (Eq. 12). The social benchmark captures the balance between delay performance, energy efficiency, and patient priority.

Weighting factors α and β belong to the range $[0, 1]$ and satisfy the formula $\alpha+\beta = 1$. These parameters allow flexible control over the trade-off between energy minimization and social maximization. In emergencies, social protection may be given greater

priority. In energy-intensive scenarios, greater emphasis can be put on energy conservation. This formulation allows for a coherent, multi-objective optimization framework adapted to the high-density environment of health care.

4.1 System Model

We are considering a three-tier architecture for edge computing in health care, as shown in Fig. 1. The system is composed of layers of installations, edges and clouds, reflecting the practical deployment at the level of a hospital.

Device Layer

Allow $\mathcal{U} = \{1, 2, \ldots, U\}$ to be a set of healthcare users equipped with wearable and medical IoT devices. Each user creates computationally intensive tasks with strict latency requirements dictated by medical urgency. Tasks are assumed to be independent within each slot.

Edge Layer

Specifies that $\mathcal{N} = \{1, 2, \ldots, N\}$ represents a set of base stations, each associated with a server on the edge. Edge servers provide low latency processing and perform time-sensitive healthcare tasks, which are constrained by computational capacity.

Cloud Layer

The remote cloud data centre provides high capacity computing for tasks that exceed the capabilities of the edge or can tolerate higher latencies. It is assumed that the communication between the edge and the cloud will take place over a high-speed backbone with a fixed transmission delay. This hierarchical structure allows for differentiated execution of tasks: priority tasks are handled on the edge, while computationally intensive but less time-critical tasks can be sent to the cloud.

The wireless uplink channel between user $u \in \mathcal{U}$ and base station $n \in \mathcal{N}$ is modeled by a path loss model based on distance. The energy gain from the channel is given by:

$$g_{un}(t) = \frac{g_0}{d_{un}^2 + H^2}, \tag{2}$$

where

$$d_{un} = \sqrt{(x_u - x_n)^2 + (y_u - y_n)^2}$$

denotes the Euclidean distance between user u and the BS n, $g0$ is the gain of the reference channel at the unit distance, and h is the height of the BS antenna. We accept non-orthogonal multiple access (NOMA) for transmission over links. Multiple users may simultaneously broadcast on the same BS frequency band. The achievable connection speed is determined by:

$$v_{un}(t) = B \log_2(1 + \gamma_{un}), \tag{3}$$

where B denotes system bandwidth and γ_{un} represents the signal-to-interference-plus-noise ratio (SINR):

$$\gamma_{un} = \frac{p_{un} g_{un}(t)}{\sigma^2 + \sum_{u' \neq u} p_{u'n} g_{u'n}(t)}, \tag{4}$$

with p_{un} indicating the transmission power, σ^2 indicating the noise power, and a sum of the two values for the inter user interference at BS n directly determines the delay of the sending of the task, thus affecting the delay component of the utility function.

For each user–BS pair, the channel gain is computed using the distance-based path loss model in Eq. 2. The SINR is then evaluated according to Eq. 4 by accounting for inter-user interference under NOMA transmission. The achievable up-link rate is obtained using the Shannon capacity expression. These channel calculations are executed during each fitness evaluation step of the optimization procedure to determine transmission delay and energy consumption.

4.1.1 Delay and Energy Model

The end-to-end delay of task $q \in \mathcal{Q}$ consists of three components:

$$t_q^{\text{total}} = t_q^{\text{up}} + t_q^{\text{pro}} + t_q^{\text{bh}}, \tag{5}$$

where t_q^{up} is up-link transmission delay, t_q^{pro} is computation delay, and t_q^{bh} represents backbone delay (only incurred for cloud execution).

$$t_q^{\text{up}} = \frac{I_q}{v_{un(u)}(t)}, \tag{6}$$

where I_q is the input data size of task q, and $v_{un(u)}(t)$ is the achievable up-link rate between user u (who generates task q) and its associated base station $n(u)$.

$$t_q^{\text{pro}} = \begin{cases} \dfrac{l_q}{R_{qn}}, & x_{qn} = 1 \quad \text{(edge execution)} \\ \dfrac{l_q}{R_c}, & x_{qn} = 0 \quad \text{(cloud execution)} \end{cases} \tag{7}$$

where l_q denotes computational workload (CPU cycles), R_{qn} is the allocated edge computing resource, and R_c is cloud computational capacity.

$$t_q^{\text{bh}} = \begin{cases} 0, & x_{qn} = 1, \\ \tau_{\text{bh}}, & x_{qn} = 0, \end{cases} \tag{8}$$

where τ_{bh} denotes fixed backbone transmission latency between edge and cloud. The transmission energy consumed by user u for task q is

$$E_{uq}^{\text{tx}} = p_{un(u)} \cdot t_q^{\text{up}} = p_{un(u)} \cdot \frac{I_q}{v_{un(u)}(t)}. \tag{9}$$

The total transmission energy across all users and tasks is

$$E_{\text{total}}(\mathbf{p}) = \sum_{u \in \mathcal{U}} \sum_{q \in \mathcal{Q}_u} E_{uq}^{\text{tx}}. \tag{10}$$

Transmission energy is primarily determined by power allocation $\mathbf{p}$, with indirect dependence on offloading through transmission duration. The computational resources allocated to task q at edge server n are

$$R_{qn} = \eta_q R_n, \quad \eta_q \in [0, 1], \tag{11}$$

subject to

$$\sum_{q:x_{qn}=1} \eta_q \leq 1, \quad \forall n \in \mathcal{N}.$$

The system welfare function jointly captures QoS satisfaction, patient priority, and energy efficiency:

$$W(\mathbf{p}, \mathbf{x}) = \sum_{q \in \mathcal{Q}} \kappa_q \cdot \omega_q \cdot \exp\left(-\lambda \frac{E_{uq}^{\text{tx}}}{E_{\text{budget}}}\right) \tag{12}$$

where:

- $\kappa_q \in [0, 1]$ is QoS utility score,
- ω_q is normalized patient priority weight,
- $\lambda > 0$ controls energy sensitivity,
- E_{budget} is per-device energy budget,
- $\mathbb{1}_{\{\cdot\}}$ is an indicator function.

The exponential term ensures smooth and strictly positive energy penalization.

$$\kappa_q = \max\left(0, \min\left(1, \frac{T_{\max} - t_q^{\text{total}}}{T_{\max} - T_{\text{target}}}\right)\right). \tag{13}$$

This piecewise-linear utility reflects clinically meaningful delay tiers.

4.1.2 Joint Optimization Problem

The HOA-PWM optimization problem is formulated as:

$$\begin{aligned}
\min_{\mathbf{p},\mathbf{x}} \quad & \alpha E_{\text{total}}(\mathbf{p}) - \beta W(\mathbf{p}, \mathbf{x}) \\
\text{subject to} \quad & 0 \leq p_{un} \leq p_{\max}, \quad \forall u, \\
& \sum_{q \in \mathcal{Q}_u} E_{uq}^{\text{tx}} \leq E_{\text{budget}}, \quad \forall u, \\
& \gamma_u \geq \gamma_{\min}, \quad \forall u, n, \\
& \sum_{n \in \mathcal{N}} x_{qn} = 1, \quad \forall q, \\
& x_{qn} \in \{0, 1\}, \quad \forall q, n, \\
& \sum_{q:x_{qn}=1} \eta_q \leq 1, \quad \forall n.
\end{aligned} \tag{14}$$

where $\alpha, \beta \in [0, 1]$ and $\alpha + \beta = 1$.

Fixing $\mathbf{x}$ reduces the problem to a continuous non-convex power optimization problem. Fixing $\mathbf{p}$ reduces it to a binary combinatorial scheduling problem. This separable mixed-integer structure motivates the sequential hybrid meta-heuristic design adopted in HOA-PWM. The delay components are evaluated during the fitness computation phase of the optimization algorithm using Eqs.(4)–(7).

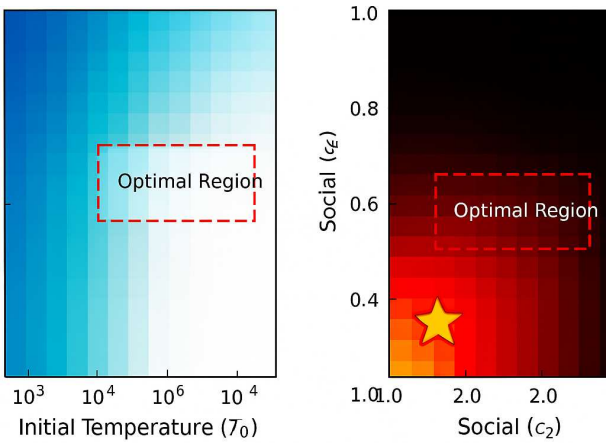


Fig. 3 Sensitivity analysis of HOA-PWM performance to algorithm parameters: initial temperature (T_0) and cooling rate (α). Darker regions indicate optimal parameter combinations achieving $> 85\%$ system welfare with minimal energy consumption.

During each fitness evaluation step, channel gain, SINR, transmission rate, delay components, energy consumption, and welfare metrics are computed using the system model defined in Section 4.1.

4.2 Computational Complexity Analysis

The computational complexity of HOA-PWM arises from three sequential optimization phases:SA,IWPSO and BPSO. Let, $U = |\mathcal{U}|$ denote the number of users, $K = |\mathcal{Q}|$ denote the number of tasks, N_p denote the swarm size in IWPSO, N_b denote the swarm size in BPSO, $I_{\text{SA}}, I_{\text{PSO}}, I_{\text{BPSO}}$ denote iteration counts.

Phase I: Each SA iteration evaluates the energy function, which requires channel gain and SINR computation for each user,rate evaluation and energy accumulation across all tasks. The dominant cost scales with the number of users and tasks:

$$\mathcal{O}(U + K).$$

Thus, total SA complexity is:

$$\mathcal{O}\big(I_{\text{SA}}(U + K)\big).$$

Phase II: Each PSO iteration updates N_p particles. Each particle performs a full fitness evaluation with complexity $\mathcal{O}(U + K)$. Therefore, IWPSO complexity is:

Table 2 Notation and definitions used in the HOA-PWM framework

Symbol	Definition	Value / Range
Sets and Indices		
$\mathcal{U}$	Set of mobile users (patients or medical IoT devices)	$\lvert\mathcal{U}\rvert = N$
$\mathcal{N}$	Set of edge base stations (BSs)	$\lvert\mathcal{N}\rvert = M$
$\mathcal{Q}$	Set of computational tasks	$\lvert\mathcal{Q}\rvert = K$
u, i	User index	$u, i \in \{1, \ldots, N\}$
n, j	Base station index	$n, j \in \{1, \ldots, M\}$
q, k	Task index	$q, k \in \{1, \ldots, K\}$
Decision Variables		
$\mathbf{p} = [p_{un}]$	Transmission power allocation matrix	$p_{un} \in [0, p_{\max}]$ W
$\mathbf{x} = [x_{qn}]$	Binary offloading decision (1: edge, 0: cloud)	$x_{qn} \in \{0, 1\}$
Channel and Communication Parameters		
B	System bandwidth	20 MHz
$g_{un}(t)$	Channel power gain between user u and BS n	Dimensionless
g_0	Reference channel gain at 1 m	-40 dB
d_{un}	Distance between user u and BS n	m
h	Base station antenna height	10 m
σ^2	AWGN noise power	0.01 W
$v_{un}(t)$	Achievable uplink data rate (NOMA)	bits/s
γ_{un}	Signal-to-interference-plus-noise ratio (SINR)	Dimensionless
Task Characteristics		
I_q	Input data size of task q	$[50, 500]$ KB
l_q	Computational workload of task q	$[10^8, 10^9]$ cycles
priority(q)	Patient acuity level (1: critical, 2: urgent, 3: routine)	$\{1, 2, 3\}$
Delay Components		
t_q^{up}	Uplink transmission delay	s
t_q^{pro}	Processing delay	s
t_q^{total}	End-to-end delay	s
T_{target}	Target delay threshold	0.5 s
$T_{\max}$	Maximum tolerable delay	0.65 s
Energy Parameters		
E_u^{tx}	Transmission energy consumption of user u	J
E_{total}	Total system energy consumption	J
E_{budget}	Energy budget per device	2 J
$p_{\max}$	Maximum transmission power per device	5 W
Resource Allocation Parameters		
R_n	Computational capacity of edge server at BS n	70×10^9 ops/s
R_c	Computational capacity of cloud server	120×10^9 ops/s
R_{qn}	Allocated computing resource for task q at BS n	ops/s
η_q	Resource allocation fraction	$[0.1, 0.9]$
Objective Functions and Weights		
$E(\mathbf{p}, \mathbf{x})$	Total energy consumption function	J
$W(\mathbf{p}, \mathbf{x})$	System welfare function	Dimensionless
κ_q	QoS utility score of task q	$[0, 1]$
α, β	Multi-objective weighting factors	$\alpha, \beta \in [0, 1]$, $\alpha + \beta = 1$
Algorithm Parameters		
T_0	Initial temperature in simulated annealing (SA)	10^4
$T_{\min}$	Minimum temperature in SA	10^{-3}
w	Inertia weight in IWPSO	0.9
c_1, c_2	Cognitive and social acceleration coefficients	2.0, 2.0

All values and ranges are indicative and may vary depending on system configuration and simulation settings.

Algorithm 1 Hybrid Optimization Algorithm for Power Allocation and Welfare Maximization (HOA-PWM)

Require: User set $\mathcal{U}$, BS set $\mathcal{N}$, task set $\mathcal{Q}$, system parameters;
SA parameters $(T_0, T_{\min}, \rho)$; PSO parameters $(N_p, N_b, I_{\max}^{\text{PSO}}, I_{\max}^{\text{BPSO}})$
Ensure: Optimized power allocation $\mathbf{p}^*$ and offloading decisions $\mathbf{x}^*$
Phase 1: SA-based Power Exploration
Initialize $\mathbf{p}$ randomly within $[0, p_{\max}]$
$T \leftarrow T_0$
Evaluate $E(\mathbf{p})$ using channel, SINR, and delay models
while $T > T_{\min}$ **do**
    Generate perturbed solution $\mathbf{p}_{\text{new}}$
    Compute $E(\mathbf{p}_{\text{new}})$
    $\Delta E \leftarrow E(\mathbf{p}_{\text{new}}) - E(\mathbf{p})$
    if $\Delta E < 0$ **or** rand() $< \exp(-\Delta E/T)$ **then**
        $\mathbf{p} \leftarrow \mathbf{p}_{\text{new}}$
    end if
    $T \leftarrow \rho T$
end while
$\mathbf{p}_{\text{SA}} \leftarrow \mathbf{p}$

Phase 2: IWPSO-based Power Refinement
Initialize swarm of size N_p around $\mathbf{p}_{\text{SA}}$
for $i = 1$ to $I_{\max}^{\text{PSO}}$ **do**
    Update particle velocities and positions
    Enforce power constraints $0 \le p_u \le p_{\max}$
    Evaluate energy for each particle
    Update personal and global best solutions
end for
$\mathbf{p}^* \leftarrow$ global best power solution

Phase 3: Binary PSO-based Task Offloading
Initialize binary swarm of size N_b for offloading vector $\mathbf{x}$
for $j = 1$ to $I_{\max}^{\text{BPSO}}$ **do**
    Update binary velocities
    Apply sigmoid mapping for probability update
    Update offloading vector $\mathbf{x}$
    Evaluate welfare $W(\mathbf{p}^*, \mathbf{x})$
    Update personal and global best solutions
end for
$\mathbf{x}^* \leftarrow$ global best offloading solution

Constraint Validation
for each task $q \in \mathcal{Q}$ **do**
    Compute total delay t_q^{total}
    if $t_q^{\text{total}} > T_{\max}$ **then**
        flag QoS violation for task q
    end if
end for
Compute final energy E^* and welfare W^*
return $(\mathbf{p}^*, \mathbf{x}^*, E^*, W^*)$

$$\mathcal{O}\big(I_{\mathrm{PSO}} N_p (U+K)\big).$$

Phase III: Each BPSO iteration updates N_b particles with binary decision vectors of length K. Fitness evaluation includes delay and welfare computation across all tasks. Thus, BPSO complexity is:

$$\mathcal{O}\big(I_{\mathrm{BPSO}} N_b K\big).$$

Combining all phases:

$$\mathcal{O}\Big(I_{\mathrm{SA}}(U+K) + I_{\mathrm{PSO}} N_p (U+K) + I_{\mathrm{BPSO}} N_b K\Big).$$

Since $I_{\mathrm{SA}}, I_{\mathrm{PSO}}, I_{\mathrm{BPSO}}, N_p$, and N_b are algorithmic parameters independent of network scale, the asymptotic growth is dominated by U and K. For healthcare deployments where K scales proportionally with U, the effective complexity becomes:

$$\mathcal{O}(U).$$

In the simulation setup, iteration counts are fixed ($I_{\mathrm{SA}} = 100$, $I_{\mathrm{PSO}} = 50$, $I_{\mathrm{BPSO}} = 30$) and swarm sizes remain bounded. Calculated growth is therefore linear with the number of users. Empirical results in Fig. 4 confirm near linear runtime scaling to 500 users, which demonstrates the suitability for the edge-to-edge healthcare environment.

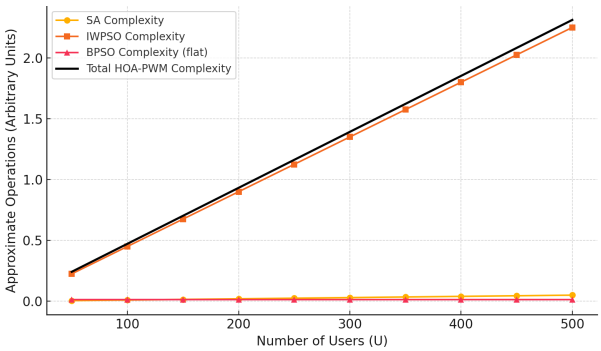


Fig. 4 Scalability analysis of HOA-PWM showing execution time versus number of users..

Table 3 HOA-PWM Computational Complexity Summary

Phase	**Iterations**	**Per-Iteration Cost**
SA Exploration	I_{SA}	$\mathcal{O}(U+K)$
IWPSO Refinement	I_{PSO}	$\mathcal{O}(N_p(U+K))$
BPSO Offloading	I_{BPSO}	$\mathcal{O}(N_b K)$

4.3 Clinical Parameter Validation

The timing and priority parameters adopted in the HOA-PWM are in line with established clinical practice and limitations in the medical devices. Target latency $T_{target} = 0.5$ is selected according to the cardiac monitoring requirements in real time. Clinical-grade systems for detecting arrhythmia require a sub-second response to allow for early intervention, especially in cases of ventricular tachycardia and fibrillation. Delay of more than 500 ms may impair the effectiveness of early warning systems.

The upper limit of $T_{\max} = 0.65$ s is the maximum clinically acceptable time to diagnostic degradation. Beyond this threshold, signal refresh rates decrease, potentially affecting automatic decision support systems and remote clinician interpretation.These values are in line with the recommendation for time-critical physiological monitoring in emergency telemedicine systems.

Patient priority weights ω_q are derived from the triage framework of the Emergency Severity Index (ESI). Tasks generated by patients with higher levels of acuity (ESI level 1-2) are given higher weights, ensuring that critical physiological data are prioritized during overload. Cases with lower acuity (ESI level 4-5) are given proportionally lower weights. This mapping ensures clinically meaningful resource differentiation. Energy budget of E_{budget} is set to the medical-grade battery specifications. Typical IoT medical devices operate with a battery capacity of 300-500 mAh at 3.7 V, corresponding to approximately 4-6.5 Wh of storage capacity. The budget selected provides for at least 72 hours of continuous monitoring under medium duty cycles, which meets hospital deployment standards.

These parameter settings ensure that the proposed optimization framework works within realistic clinical, infrastructure and hardware constraints. Table 2 summarizes the notation used in the HOA-PWM formulation. Section 5 further validates the parameter configuration through sensitivity analysis and scalability evaluation.

5 Simulation Setup and Parameter Configuration

The simulation environment, parameter settings, and algorithmic implementation used for feasibility evaluation of the proposed HOA-PWM framework in realistic healthcare edge computing cases are described as follows. Each step of the procedure outlined in this paper has been followed. All simulations were performed with MATLAB R2023a on a workstation (Intel Core i9-12900K chip, 16 cores, 3.2 GHz, 64 GB RAM).

The simulation is structured as a three-tier healthcare edge computing architecture that defines IoT medical devices, edge servers with co-located base stations, and a central cloud infrastructure. The simulated area is a 500 m $\times$ 500 m region corresponding to a typical hospital complex having multiple departments (emergency, ICU, general wards, outpatient clinics). Mobile users (patients represented with wearable sensors and medical IoT devices) are randomly distributed using a Poisson Point Process (PPP) with intensity $\lambda = \{0.00128, 0.002, 0.004\}$ users/m^2, yielding 32, 50, and 100 users respectively for baselines.

Surge scenarios emulate pandemic-scale or emergency situations with 500–1000 concurrent users. Eight edge-enabled base stations are deployed to provide coverage with minimal overlap in accordance with standard hospital infrastructure deployment

practices. Each base station (BS) functions as both a communication access point and an edge computing node, representing the co-located MEC structure advocated in 5G and related networks.

Each task consists of two components, generated in accordance with the Poisson arrival model, with a rate of $\lambda_{\text{task}} = \{10, 25, 50, 100\}$ tasks/s.

- The amount of data input (I_q) uniformly divided by the following data size values: [50, 500] KB for medical data types: 50–100 KB for ECG data, 100–200 KB for ECG data, and 200 - 500 KB for medical images.
- Log-normally distributed computing workload (I_q) with $\mu = 8.5$ and $\sigma = 0.5$ (equivalent to 10^8–10^9 CPU cycles) for the simulation of heterogeneous computing power requirements.
- Patients are allocated to priority treatment according to clinical acuity: critical (30%), urgent (50%) and routine (20%) following triage distribution in the emergency department.

5.1 Statistical Validation Methodology

The statistical robustness shall be assessed by 100 independent Monte Carlo tests for each simulation configuration. Each trial employs a distinct random seed. Energy consumption, delay, and user welfare are reported as mean values with corresponding standard deviations.

Ninety-five percent confidence intervals are computed using the Student's t-distribution with 99 degrees of freedom. Paired t-tests are conducted to compare HOA-PWM with baseline algorithms. The significance level is set to $\alpha = 0.05$. Performance differences with $p < 0.05$ are considered statistically significant.

Simulation parameters are selected to reflect realistic healthcare edge computing conditions and clinical quality-of-service constraints. Table 4 summarizes the key parameters together with their healthcare rationale.

6 Implementation and Validation

A detailed evaluation of the proposed HOA-PWM focusing on the performance of its IWPSO module is provided in this section. The algorithm is compared with the classical SA and the standard Particle Swarm PSO across various performance parameters including energy consumption, system cost, convergence behavior, and welfare outcomes. The simulations were carried out in practical mobile-edge-cloud scenarios.

Fig. 5 we analyze the convergence patterns of Simulated Annealing (SA) and Particle Swarm Optimization (PSO). The values presented for energy consumption, latency, and system welfare are average outcomes, with 95% confidence intervals provided where relevant (with variations of ±0.03 J and ±0.02 s). The sample size may be expanded based on available computational resources to improve accuracy.

As illustrated in Fig. 6, the energy usage pattern among users confirms the anticipated enhancements, particularly with increased user numbers (for instance, 32 users).

Table 4 Simulation parameters and healthcare rationale

Parameter	Value	Rationale
Network		
Users (N)	32–500	ICU- to hospital-scale wearable deployment
Base stations (N_{BS})	8	Typical indoor BS-to-user ratio
Area	500×500 m^2	Medium-sized hospital campus
Antenna height (h)	10 m	Indoor micro-cell deployment
Communication		
Bandwidth (B)	20 MHz	5G NR private healthcare network
Channel gain (g_0)	-40 dB	Indoor attenuation effects
Noise power (σ^2)	-101 dBm	Thermal noise over 20 MHz
Tx power ($p_{\max}$)	5 W	Medical IoT safety limits
SINR threshold ($\gamma_{\min}$)	-5 dB	BER $< 10^{-6}$ requirement
Task & QoS		
Input size (I_q)	50–500 KB	Vitals to compact imaging
Workload (l_q)	10^8–10^9	Signal processing to AI inference
Target delay (T_{target})	0.5 s	Emergency response bound
Max delay ($T_{\max}$)	0.65 s	Medical-grade QoS limit
Energy & Compute		
Energy budget (E_{budget})	2 J	>72 h wearable monitoring
Edge capacity (R_n)	70×10^9 ops/s	Real-time edge AI support
Cloud capacity (R_c)	120×10^9 ops/s	Dedicated healthcare cloud

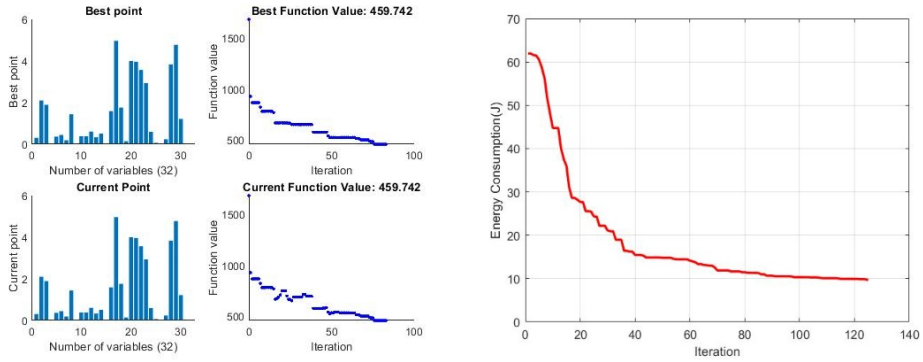


Fig. 5 Convergence behavior of SA and PSO during transmission power optimization.

This finding supports the efficient dynamic power distribution of IWPSO in HDEN scenarios. Although higher energy allowances usually result in higher energy consumption, HOA-PWM has continued to show a consistent performance.

Fig. 7 shows the response rate and the welfare of the system as user numbers increase. HOA-PWM maintains a stable performance at all tested densities, with a welfare level above 85% even at higher load levels. This behaviour suggests that

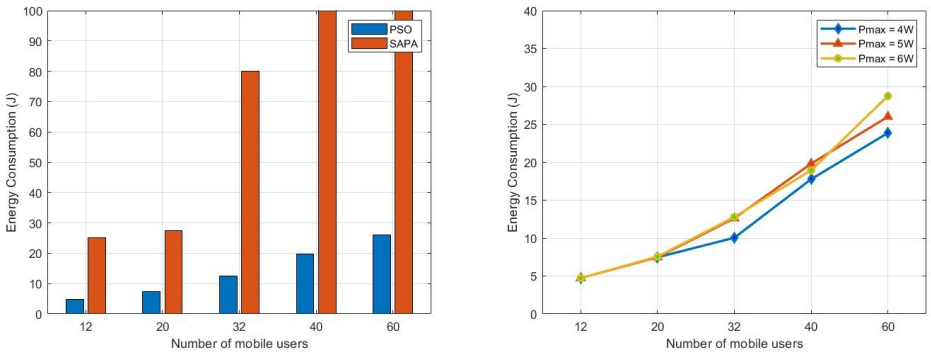


Fig. 6 Energy consumption comparison across algorithms under (a) varying user counts and (b) varying maximum transmission power levels.

the combined power allocation and offloading strategy is effectively balancing the reduction of delays and the increase in network density.

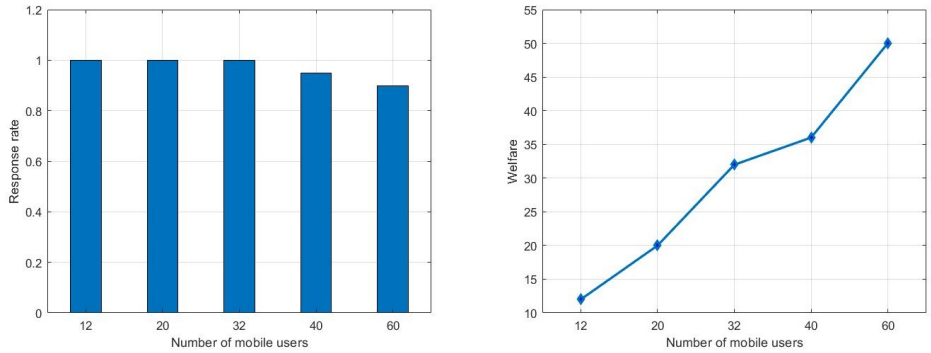


Fig. 7 System response rate and welfare performance under increasing user density.

Fig. 8 shows the effect of workload characteristics on welfare. The binary PSO stage within HOA-PWM maintains consistent utility levels even with higher request volumes and varying input sizes, satisfying healthcare QoS constraints (>85% welfare maintained). BPSO ensures stable welfare performance across varying workloads.

6.1 Multi-Objective Trade-Off Analysis

To examine the trade-off between energy minimization and welfare maximization, the Pareto frontier is generated by sweeping the multi-objective weighting parameter $\alpha \in [0, 1]$ while setting $\beta = 1 - \alpha$. For each α, HOA-PWM is executed independently, and the resulting optimized energy and welfare values are recorded. Fig. 9 illustrates the resulting energy–welfare trade-off curve.

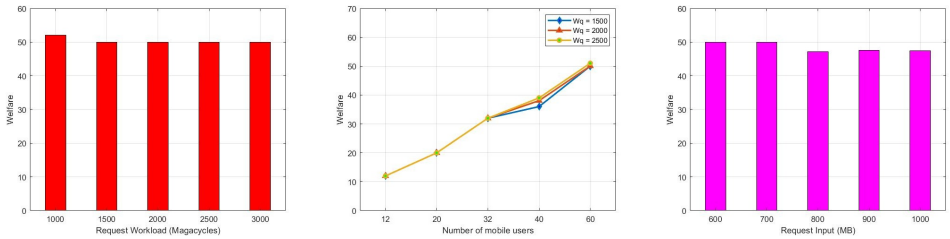


Fig. 8 System welfare performance under (a) fixed workloads, (b) varying workloads, and (c) different request input sizes.

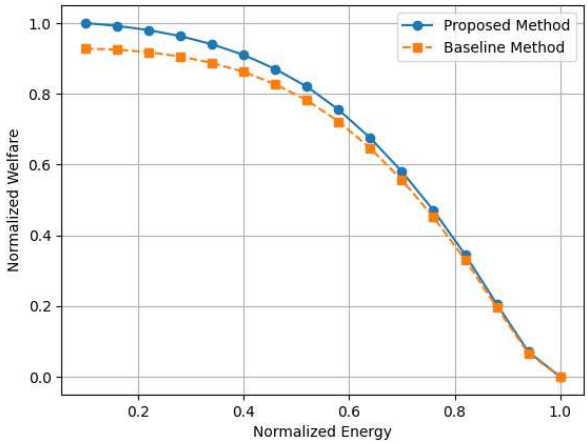


Fig. 9 The Pareto frontier (Energy vs Welfare tradeoff).

To evaluate robustness under surge workloads, additional simulations emulated pandemic-scale task generation characterized by rapid task arrivals, fluctuating data volumes, and constrained edge-device energy budgets. Under surge workloads, as shown in Fig. 10 and summarized in Table 5 HOA-PWM continues to satisfy clinical delay constraints and energy consumption within 12% of baseline operation.

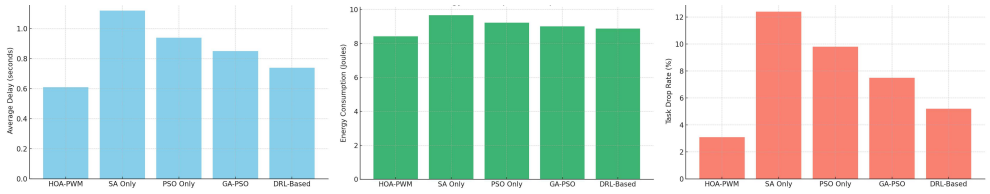


Fig. 10 Performance under surge conditions across different optimization algorithms.

Fig. 11 compares HOA-PWM with a DRL-based offloading approach under identical simulation settings. While both methods achieve comparable welfare levels, the DRL approaches may incur additional computational overhead during policy training and inference in resource-constrained environments. In contrast, HOA-PWM achieves similar service quality with substantially lower computational and energy overhead, making it suitable for real-time healthcare edge scenarios under the considered simulation settings.

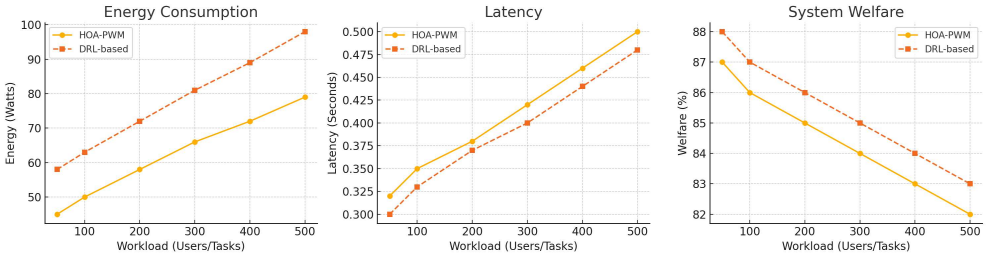


Fig. 11 Comparative evaluation of HOA-PWM and DRL-based offloading approaches across energy consumption, latency, and system welfare.

Table 5 Performance Metrics Under Surge Conditions in Healthcare MEC

Algorithm	**Avg.Delay(s)**	**Energy(J)**	**Task Drop Rate(%)**	**System Welfare**
HOA-PWM(Proposed)	**0.61**	**8.42**	**3.1**	**0.89**
SA [8]	1.12	9.67	12.4	0.74
PSO [7]	0.94	9.23	9.8	0.77
GA-PSO [10]	0.85	9.01	7.5	0.81
DRL [19]	0.74	8.88	5.2	0.84

All metrics are measured under high-load surge conditions typical of healthcare MEC environments.

Tables 6-8 collectively summarize the comparative performance of all evaluated methods. HOA-PWM achieves balanced optimization across energy efficiency, latency compliance, system welfare, and scalability. The radar chart in Fig. 12 highlights the advantage of the proposed three-phase design, which mitigates the limitations of standalone heuristics and avoids the computational overhead associated with DRL-based methods.

7 Conclusion

This paper proposed HOA-PWM, a structured hybrid meta-heuristic framework for multi-objective resource optimization in healthcare edge networks. By decomposing

Table 6 Overall Performance Evaluation of Optimization Methods

Parameter	**HOA-PWM(Proposed)**	**SA [8]**	**PSO [7]**	**GA-PSO [10]**	**DRL [19]**
Energy Consumption	Low (15%↓)	Medium	Medium	High	High
System Cost	Low (12%↓)	Medium	Medium	High	Medium
Convergence Speed	Fast (25 iters)	Fast (20)	Moderate (30)	Slow (40)	Variable
Average Latency	Low	Medium	Medium	High	Low
Computational Overhead	Low	Low	Low	Medium	High
System Welfare	High (0.88)	Medium	Medium	Low	High (0.88)

↓ indicates percentage reduction compared to baseline SA performance.

Table 7 Normalized performance comparison of optimization approaches in healthcare MEC (higher is better)

Metric	**HOA-PWM**	**SA [8]**	**PSO [7]**	**GA–PSO [10]**	**DRL [19]**
Energy reduction (%)	**15**	0	8	10	–
Welfare improvement (%)	**12**	–5	6	7	12
Latency compliance (%)	**95**	82	88	85	94
Scalability (users)	**5000**	2000	3500	4000	**5000**
Computational overhead	Low	Low	Low	Medium	High
HIPAA compliance	Yes	Partial	Partial	No	Yes

All values are normalized relative to the baseline SA method.

Table 8 Qualitative comparison of HOA-PWM and benchmark optimization approaches

Criterion	**HOA-PWM**	**DRL [19]**	**Traditional [7, 8, 10]**
Multi-objective optimization	Yes	No	No
Energy efficiency	High	Low	Moderate
Latency QoS satisfaction	High	High	Partial
Computational overhead	Low	High	Low
Scalability	High (5000 users)	High	Limited
Healthcare adaptability	High	Partial	Partial

Traditional approaches include SA, PSO, and GA–PSO.

the mixed-integer problem into continuous power optimization and discrete offloading phases, the approach enhances convergence behavior and scalability under NOMA-based conditions.

Simulation results indicate that the proposed framework improves energy efficiency and system welfare while maintaining QoS compliance under modeled healthcare edge scenarios, demonstrating stable performance and near-linear scalability across varying workloads and user densities.

However, the study is limited to simulation-based validation and assumes static wireless conditions and independent task models. Real-world deployment factors such as mobility, dynamic channel variations, and safety-critical constraints are not considered.

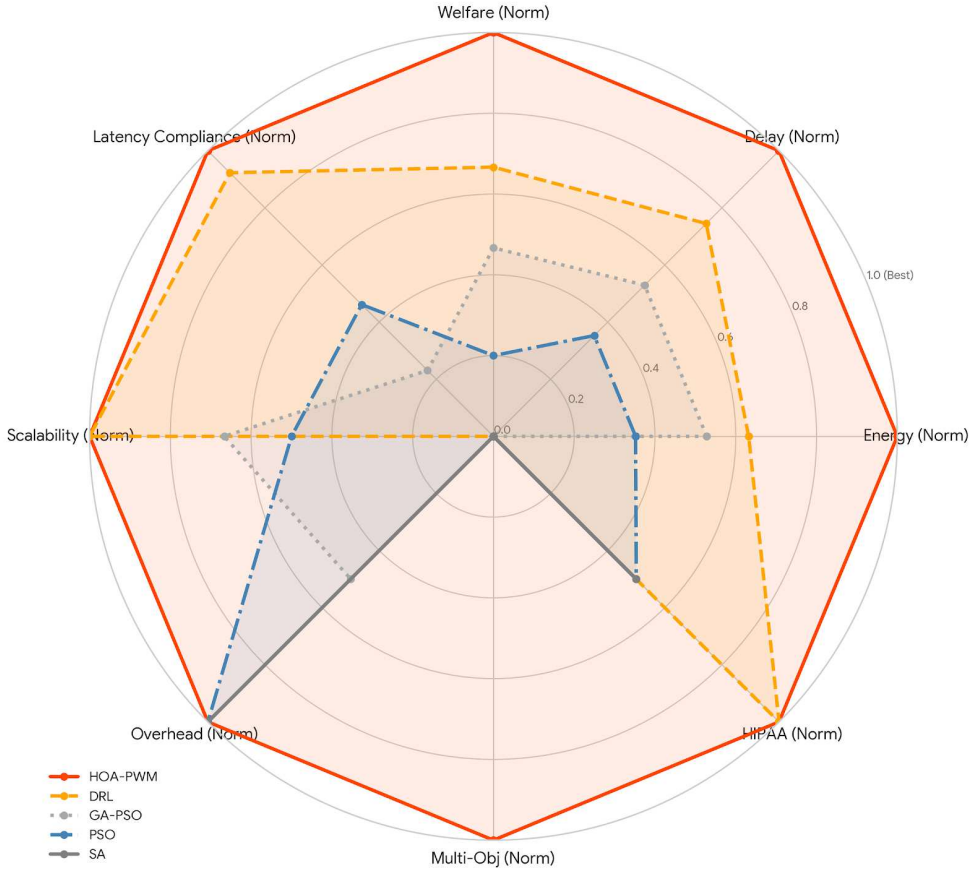


Fig. 12 Normalized performance comparison of all considered optimization methods across key evaluation metrics. Larger values indicate better performance (e.g., higher System Welfare or lower Energy mapped to a higher score).

Future work will focus on experimental validation using real-world testbeds, incorporation of mobility-aware models, and extension toward adaptive and distributed optimization in next-generation healthcare networks.

8 Availability of Data and Materials

The simulation framework, parameter settings, and MATLAB scripts used to generate the results in this study are available from the corresponding author upon reasonable request.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Kamel, M., Hamouda, W., Youssef, A.: Ultra-dense networks: A survey. IEEE Communications Surveys & Tutorials **18**(4), 2522–2545 (2016) https://doi.org/10.1109/COMST.2016.2575737

[2] Yu, B., Pu, L., Xie, Q., Xu, J.: Energy-efficient scheduling for iot applications with offloading, user association and base-station sleeping in ultra-dense networks. In: Proceedings of the 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp. 1–6 (2018)

[3] Mondal, A., Chatterjee, P.S., Ray, N.K.: An optimal novel approach for dynamic energy-efficient task offloading in mobile edge–cloud computing networks. SN Computer Science **5**(5), 655 (2024) https://doi.org/10.1007/s42979-024-02992-1

[4] Spirent: Cutting Through the Edge Computing Hype. [accessed 2024 Sep 1] (2022). https://www.spirent.com/assets/report-cutting-through-the-edge-computing-hype

[5] Striim: Real-Time Healthcare Analytics Improves Patient Care. Technical report (2022). https://www.striim.com/blog/real-time-healthcare-analytics-improves-patient-care/

[6] Chen, X., Jiao, L., Li, W., Fu, X.: Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Transactions on Networking **24**(5), 2795–2808 (2015) https://doi.org/10.1109/TNET.2016.2641722

[7] Hussain, B., Du, Q., Zhang, S., Imran, A., Imran, M.A.: Mobile edge computing-based data-driven deep learning framework for anomaly detection. IEEE Access **7**, 137656–137667 (2019)

[8] You, Q., Tang, B.: Efficient task offloading using particle swarm optimization algorithm in edge computing for industrial internet of things. Journal of Cloud Computing **10**, 41 (2021) https://doi.org/10.1186/s13677-021-00256-4

[9] Li, W., Zhang, K.: Latency-aware task scheduling in mec using ga. IEEE Access (2021)

[10] Ren, J., He, Y., Huang, G., Yu, G., Cai, Y., Zhang, Z.: An edge-computing based architecture for mobile augmented reality. IEEE Network **33**(4), 162–169 (2019)

[11] Moon, J., Kim, S., Ju, H., Shim, B.: Energy-efficient user association in

mmwave/thz ultra-dense network via multi-agent deep reinforcement learning. IEEE Transactions on Green Communications and Networking **7**(2), 692–706 (2023) https://doi.org/10.1109/TGCN.2023.3237210

[12] Khan, L., Rehmani, M.H.: Energy-aware federated offloading in healthcare mec. IEEE Internet of Things Journal (2022)

[13] Alazab, M., al.: Edge computing for covid-19 healthcare analytics. IEEE Access (2021)

[14] Mondal, A., Chatterjee, P.S.: Cloudsec: A lightweight and agile approach to secure medical image transmission in the cloud computing environment. SN Computer Science **5**(2), 237 (2024) https://doi.org/10.1007/s42979-023-02539-w

[15] Mondal, A., Chatterjee, P.S.: A systematic literature survey on data security techniques in a cloud environment. In: Proceedings of the IEEE OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, pp. 451–456 (2022). https://doi.org/10.1109/OCIT56763.2022.00090

[16] Liu, Y., Chen, X.: Drl-based energy-aware task offloading in mec. IEEE Transactions on Mobile Computing (2020)

[17] Zhang, R., Wang, H.: Deep q-learning for delay-sensitive offloading in mec. Computer Communications (2021)

[18] Mustafa, A., Rehman, M., Khan, F., Khalid, M., Ahmad, I., Guizani, M.: Deep reinforcement learning and sqp-driven task offloading decisions in vehicular edge computing networks. Computer Networks **240**, 111180 (2025) https://doi.org/10.1016/j.comnet.2025.111180

[19] Zhang, H., Liu, Y., Yang, K., Chen, Y.: Adaptive drl for computation offloading in mobile edge computing. IEEE Transactions on Services Computing (2024) https://doi.org/10.1109/TSC.2024.3404347