Quantum-Inspired Simulated Annealing with Neural Guidance for Hospital Scheduling

Nourhen Kachroudi[1], Faiza Ajmi[2], Sarah Ben Othman[3], Juliette Therasse[4], Robert Caiazzo[4], and Slim Hammadi[1]

[1] CRIStAL UMR CNRS 9189, Ecole Centrale of Lille, Scientific City, 59650 Villeneuve d'Ascq, France
[2] ICL, Junia, Catholic University of Lille, LITL, F-59000 Lille, France
[3] CRIStAL UMR CNRS 9189, Polytech Lille, Scientific City, 59650 Villeneuve d'Ascq, France
[4] CHU Lille - Hôpital Claude Huriez, Rue Michel Polonowski, 59000 Lille

Abstract. Hospital scheduling requires the coordinated allocation of operating rooms, inpatient beds, and healthcare staff under strict constraints and interacting performance objectives. Decisions at the surgical level propagate through downstream care units, while congestion in inpatient services may restrict surgical activity, resulting in a large-scale and tightly coupled optimization problem.

This paper proposes an integrated framework based on a unified Quadratic Unconstrained Binary Optimization (QUBO) formulation that jointly models operating room scheduling, bed management across clinical phases, and human resource allocation. Non-linear congestion effects are captured using a Choquet integral–based aggregation of soft criteria, explicitly modeling interactions between delays, bed shortages, staff overload, and cancellations. The resulting non-convex QUBO is solved using a neural-guided quantum-inspired simulated annealing algorithm.

Experiments on realistic synthetic instances demonstrate clear improvements over classical simulated annealing and unguided quantum-inspired methods in terms of feasibility, solution quality, and convergence speed.

Keywords: Hospital resource optimization · Operating room scheduling · Bed management · Quadratic unconstrained binary optimization (QUBO) · Choquet integral · Quantum-inspired optimization · Neural-guided metaheuristics

1 Introduction

Healthcare systems operate under increasing demand, limited capacity, and growing organizational complexity. Efficient coordination of operating rooms, inpatient beds, and healthcare staff is therefore central to hospital performance and sustainability [1, 2]. Poor coordination may lead to prolonged waiting times, resource underutilization, or cascading congestion across care units.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

Hospital scheduling problems are inherently combinatorial and multi-objective. Traditional optimization approaches, including mixed-integer programming and constraint-based models, provide strong guarantees for small instances but face scalability limitations in realistic multi-phase and resource-constrained environments [3, 4].

A further challenge lies in the nonlinear nature of congestion effects. Operational criteria such as delays, bed shortages, and staff overload are often treated independently through linear aggregation. However, empirical evidence suggests that their simultaneous occurrence may amplify system instability, a phenomenon insufficiently captured by additive models [5, 6].

Recent advances in combinatorial optimization have highlighted the potential of quadratic binary formulations [7], non-additive multi-criteria aggregation [8], and learning-guided metaheuristics [9, 10]. Yet, these research directions have largely evolved separately in the context of hospital resource allocation.

This work builds upon these developments by integrating structured binary modeling, interaction-aware aggregation, and learning-guided stochastic search within a unified framework. The remainder of the paper details the problem formulation, methodological components, and empirical evaluation.

2 Related Work

Hospital scheduling and resource allocation have been extensively studied due to their economic impact and operational complexity. These problems are large-scale, combinatorial, and multi-objective, limiting the scalability of exact optimization methods in realistic settings [1, 11]. Classical approaches rely on mixed-integer programming, constraint programming, and rule-based heuristics, which perform well on small instances but degrade in multi-phase and resource-constrained environments [3]. Consequently, metaheuristics such as simulated annealing, tabu search, genetic algorithms, and hybrid strategies have been widely adopted to improve scalability [4].

QUBO has emerged as a unified modeling framework for encoding constraints and objectives into a single quadratic formulation [7]. Annealing-based and related techniques have shown promising results for complex scheduling and allocation problems [12], particularly in settings characterized by rugged energy landscapes.

Recent advances in learning-based combinatorial optimization leverage neural networks to guide search procedures and learn problem-specific heuristics [9]. Hybrid neural–metaheuristic approaches improve convergence and robustness, especially in high-dimensional QUBO problems where decision variables have heterogeneous impact [10, 13].

Hospital planning also involves interacting performance criteria. Linear weighted aggregation assumes independence and may fail to capture nonlinear interactions. The Choquet integral provides a flexible framework for modeling such interactions and has been applied to complex decision-making systems, including healthcare applications [14]. Building on these research directions, this

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

work integrates QUBO modeling, Choquet-based aggregation, and neural-guided quantum-inspired optimization within a unified framework for hospital resource allocation.

3 Problem Description

We consider an integrated hospital scheduling problem involving operating rooms, inpatient beds, and healthcare staff over a discrete planning horizon.

3.1 Sets and Parameters

The horizon is discretized as $T = \{1, \ldots, H\}$.

We define the sets:

$$P \text{ (patients)},\ B \text{ (operating rooms)},\ L \text{ (beds)},\ R \text{ (staff)},\ PH \text{ (clinical phases)}.$$

Each patient $p \in P$ is characterized by surgical duration d_p^{op}, phase durations $d_{p,ph}^{ph}$, planned start time t_p^{plan}, and priority weight ω_p.

Each phase $ph \in PH$ has capacity C_{ph}, and each resource $r \in R$ has workload limit $W_r^{\max}$. Compatibility between patient p and resource r is given by $\kappa_{p,r} \in \{0, 1\}$.

3.2 Decision Variables

All decision variables are binary:

$$x_{p,b,t}^{op} = \begin{cases} 1 & \text{if patient } p \text{ undergoes a surgical operation} \\ & \text{in operating room } b \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

$$x_{p,l,ph,t} = \begin{cases} 1 & \text{if patient } p \text{ occupies bed } l \\ & \text{in phase } ph \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

$$x_{p,r,t} = \begin{cases} 1 & \text{if staff } r \text{ is assigned to patient } p \\ & \text{at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

3.3 Hard Constraints

The schedule must satisfy:

- Each patient is operated exactly once.
- Each operating room hosts at most one patient per time slot.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

- Surgeries last d_p^{op} consecutive slots.
- Each bed hosts at most one patient at a time.
- Phase occupancy does not exceed C_{ph}.
- A patient occupies at most one phase at any time.
- Staff workload does not exceed $W_r^{\max}$.
- Staff assignments satisfy compatibility $\kappa_{p,r}$.

3.4 Objectives

Beyond feasibility, the objective is to minimize surgical delays, bed congestion, staff overload, and cancellations. The problem is combinatorial and NP-hard, motivating advanced reformulation and metaheuristic solution strategies.

To efficiently address this computational complexity, the integrated optimization problem is reformulated into a quadratic unconstrained framework. This transformation enables the simultaneous handling of feasibility constraints and performance objectives within a unified energy-based representation, paving the way for advanced solution approaches.

4 QUBO Modeling with Choquet Integral

The integrated hospital resource allocation problem is reformulated as a QUBO, encoding all constraints and objectives into a single quadratic energy function over binary variables.

4.1 Global Binary Representation

Let $x \in \{0,1\}^n$ denote the global binary vector obtained by concatenating:

- $x_{p,b,t}^{op}$: operating room assignment variables,
- $x_{p,l,ph,t}$: bed assignment variables,
- $x_{p,r,t}$: human resource assignment variables,
- $s_k^{ph,t} \in \{0,1\}$: Binary slack variable used in the binary encoding of unused capacity for phase ph at time t (bit position k).
- $u_k^{r,t} \in \{0,1\}$: Binary slack variable used in the binary encoding of unused capacity of resource r at time t (bit position k).
- $z_{p,t} \in \{0,1\}$: Binary variable equal to 1 if the surgery of patient p starts at time t, and 0 otherwise.
- $\delta_p \in \{0,1\}$: Binary variable equal to 1 if patient p is cancelled (i.e., no surgery is scheduled), and 0 otherwise.
- $e_{ph,t} \geq 0$: Non-negative excess variable representing bed congestion in phase ph at time t, i.e., the amount by which occupancy exceeds capacity.
- $h_{r,t} \geq 0$: Non-negative overload variable representing workload excess of resource r at time t, i.e., the amount by which workload exceeds its maximum capacity.

All variables are binary unless otherwise stated.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

4.2 QUBO Formulation and Constraint Handling

The optimization problem is:

$$\min_{x\in\{0,1\}^n} H(x) = x^\top Q x, \tag{1}$$

where $Q \in \mathbb{R}^{n\times n}$ is symmetric.

The energy function is decomposed as:

$$H(x) = H_{\text{hard}}(x) + \lambda H_{\text{soft}}(x), \tag{2}$$

where $\lambda > 0$ ensures feasibility dominance.

$H_{\text{hard}}(x)$ encodes all constraints as quadratic penalties (zero if satisfied, positive otherwise), while $H_{\text{soft}}(x)$ represents the optimization objectives.

A solution is feasible if $H_{\text{hard}}(x) = 0$ (or $\leq \varepsilon$). The number of violations is:

$$V(x) = \sum_{c\in\mathcal{C}} \mathbb{I}(H_c(x) > 0), \tag{3}$$

with $H_c(x)$ the penalty of constraint c. Penalty coefficients and λ are set to prioritize feasibility, driving the search toward valid solutions.

4.3 Hard Constraint Encoding

All inequality constraints of the form $\sum_i x_i \leq C$ are transformed into equality constraints using binary slack variables.

Operating Room Capacity

$$\sum_{p\in P} x^{op}_{p,b,t} \leq 1. \tag{4}$$

QUBO encoding:

$$H_{\text{OR}} = A_{\text{OR}} \sum_{b,t} \left(\sum_p x^{op}_{p,b,t} \right) \left(\sum_p x^{op}_{p,b,t} - 1 \right). \tag{5}$$

This expression equals zero if at most one patient is assigned and is strictly positive otherwise.

Bed Uniqueness

$$\sum_{p\in P} x_{p,l,ph,t} \leq 1. \tag{6}$$

Encoded as:

$$H_{\text{bed-unique}} = A_{\text{bed-unique}} \sum_{l,ph,t} \left(\sum_p x_{p,l,ph,t} \right) \left(\sum_p x_{p,l,ph,t} - 1 \right). \tag{7}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

Phase Capacity

$$\sum_{p,l} x_{p,l,ph,t} \leq C_{ph}. \tag{8}$$

Let

$$K_{ph} = \lceil \log_2(C_{ph} + 1) \rceil.$$

Introduce binary slack encoding:

$$s_{ph,t} = \sum_{k=0}^{K_{ph}-1} 2^k s_k^{ph,t}. \tag{9}$$

The inequality is transformed into:

$$\sum_{p,l} x_{p,l,ph,t} + s_{ph,t} = C_{ph}. \tag{10}$$

Penalty:

$$H_{\text{phase-cap}} = A_{\text{phase-cap}} \sum_{ph,t} \left(\sum_{p,l} x_{p,l,ph,t} + \sum_k 2^k s_k^{ph,t} - C_{ph} \right)^2. \tag{11}$$

The slack variable $s_{ph,t}$ represents unused capacity.

Human Resource Capacity

$$\sum_{p} x_{p,r,t} \leq W_r^{\max}. \tag{12}$$

Binary slack:

$$u_{r,t} = \sum_{k=0}^{K_r-1} 2^k u_k^{r,t}, \quad K_r = \lceil \log_2(W_r^{\max} + 1) \rceil. \tag{13}$$

Equality form:

$$\sum_{p} x_{p,r,t} + u_{r,t} = W_r^{\max}. \tag{14}$$

Penalty:

$$H_{\text{HR-cap}} = A_{\text{HR-cap}} \sum_{r,t} \left(\sum_{p} x_{p,r,t} + \sum_k 2^k u_k^{r,t} - W_r^{\max} \right)^2. \tag{15}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

Skill Compatibility

$$x_{p,r,t} \leq \kappa_{p,r}. \tag{16}$$

Penalty:

$$H_{\text{skill}} = A_{\text{skill}} \sum_{p,r,t} (1 - \kappa_{p,r}) x_{p,r,t}. \tag{17}$$

4.4 Soft Performance Criteria

Once feasibility is enforced through hard constraints, the objective focuses on improving schedule quality. While hard constraint violations are unacceptable, soft degradations—such as delays, congestion, overload, and cancellations—should be minimized.

Since QUBO cannot directly encode non-linear operators such as $\max(0, \cdot)$, each criterion is reformulated using non-negative auxiliary variables that capture the corresponding positive excess.

Surgical Delay Let $z_{p,t} \in \{0, 1\}$ indicate whether the surgery of patient p starts at time t, with:

$$\sum_{t \in T} z_{p,t} = 1. \tag{18}$$

The effective start time is:

$$t_p^{start} = \sum_{t \in T} t \, z_{p,t}. \tag{19}$$

To measure delay relative to the planned time t_p^{plan}, we introduce a non-negative auxiliary variable d_p such that:

$$d_p \geq t_p^{start} - t_p^{plan}. \tag{20}$$

The delay cost is defined as:

$$f_1(x) = \sum_{p \in P} \omega_p d_p^2, \tag{21}$$

where ω_p reflects patient priority. Energy minimization ensures:

$$d_p = \max(0, t_p^{start} - t_p^{plan}),$$

meaning only actual delays are penalized.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

Bed Congestion Let the occupancy of phase ph at time t be:

$$O_{ph,t} = \sum_{p \in P} \sum_{l \in L} x_{p,l,ph,t}. \tag{22}$$

To capture capacity violations, we introduce a non-negative excess variable $e_{ph,t}$ such that:

$$e_{ph,t} \geq O_{ph,t} - C_{ph}. \tag{23}$$

The congestion cost is:

$$f_2(x) = \sum_{ph \in PH} \sum_{t \in T} e_{ph,t}^2. \tag{24}$$

Minimization ensures:

$$e_{ph,t} = \max(0, O_{ph,t} - C_{ph}),$$

thus penalizing only actual overcrowding.

Human Resource Overload The workload of resource r at time t is:

$$W_{r,t} = \sum_{p \in P} x_{p,r,t}. \tag{25}$$

We introduce a non-negative overload variable $h_{r,t}$ such that:

$$h_{r,t} \geq W_{r,t} - W_r^{\max}. \tag{26}$$

The overload cost is defined as:

$$f_3(x) = \sum_{r \in R} \sum_{t \in T} h_{r,t}^2. \tag{27}$$

Minimization guarantees:

$$h_{r,t} = \max(0, W_{r,t} - W_r^{\max}),$$

so that only excessive workload is penalized.

Cancellations Let $\delta_p \in \{0, 1\}$ denote whether patient p is cancelled. We impose:

$$\delta_p \geq 1 - \sum_{b \in B} \sum_{t \in T} x_{p,b,t}^{op}. \tag{28}$$

The cancellation cost is:

$$f_4(x) = \sum_{p \in P} \omega_p \delta_p. \tag{29}$$

Energy minimization ensures that $\delta_p = 1$ only if no surgery is scheduled.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

Normalization The four criteria $f_k(x)$ have different physical scales (time, occupancy units, workload, binary events). To ensure comparability, each component is normalized:

$$\tilde{f}_k(x) = \frac{f_k(x)}{f_k^{\max}}, \quad k = 1, \ldots, 4, \tag{30}$$

where $f_k^{\max}$ is a scaling constant representing the maximum admissible magnitude of criterion k.

This normalization prevents any single criterion from dominating the objective purely due to scale differences.

Choquet Integral Aggregation A classical aggregation uses a weighted sum,

$$\sum_{k=1}^{n} \alpha_k \tilde{f}_k(x),$$

where $\alpha_k \geq 0$ denotes the relative importance weight assigned to criterion k, satisfying

$$\sum_{k=1}^{n} \alpha_k = 1.$$

This linear formulation assumes mutual independence between criteria. In the present study, $n = 4$, corresponding to surgical delay, bed congestion, human resource overload, and cancellations.

However, in hospital systems, simultaneous degradations (e.g., congestion and overload) may generate amplified systemic effects, making linear aggregation inadequate.

We therefore adopt a Choquet integral defined by a monotone capacity

$$\mu : 2^{\{1,\ldots,n\}} \to [0, 1], \tag{31}$$

with $\mu(\emptyset) = 0$ and $\mu(\{1, \ldots, n\}) = 1$, allowing interaction modeling between criteria.

Since the sorting-based Choquet expression is not directly quadratic, it is approximated by

$$H_{\text{Choquet}}(x) \approx x^\top Q_{\text{Choquet}} x, \tag{32}$$

preserving QUBO compatibility while capturing nonlinear effects.

This section presents a unified QUBO formulation embedding all constraints and interactions into a single energy function. The resulting large, non-convex landscape makes exact optimization intractable, motivating the use of advanced metaheuristics.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

5 Neural-Guided Quantum-Inspired Optimization

The resulting QUBO is large-scale, non-convex, and NP-hard, making exact optimization impractical and classical heuristics prone to local minima. To address this, we extend Simulated Annealing to a quantum-inspired variant (QISA) and further enhance it with neural guidance (QISA+NN) for efficient exploration.

5.1 Classical Simulated Annealing

SA is a stochastic metaheuristic inspired by thermodynamic cooling. At iteration k, a candidate solution x' is generated and accepted with probability:

$$P(\text{accept}) = \exp\left(-\frac{H(x') - H(x)}{T_k}\right), \tag{33}$$

where T_k is the temperature parameter. High temperatures promote exploration by allowing uphill moves, while decreasing T_k gradually enforces exploitation. In high-dimensional QUBO problems, however, SA may struggle to escape deep local minima.

Algorithm 1 Classical Simulated Annealing (SA)

```
Initialize x and temperature T_0
for k = 1 to K_max do
    Generate neighbor x'
    Δ ← H(x') − H(x)
    if Δ ≤ 0 or rand() < exp(−Δ/T_k) then
        x ← x'
    end if
    Update T_k
end for
return best solution
```

5.2 Quantum Annealing Intuition

Quantum Annealing introduces quantum tunneling, allowing transitions across energy barriers without thermal activation. Rather than climbing over barriers, the system may probabilistically traverse them. We adopt a classical approximation that mimics this behavior through collective probabilistic updates.

5.3 Quantum-Inspired Simulated Annealing (QISA)

QISA replaces a single solution by a probability vector:

$$\psi_i = \mathbb{P}(x_i = 1), \quad i = 1, \ldots, n. \tag{34}$$

At each iteration, a solution is sampled independently:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

$$x \sim \prod_{i=1}^{n} \psi_i^{x_i} (1-\psi_i)^{1-x_i}. \tag{35}$$

The best solution x^* is retained, and probabilities are updated as:

$$\psi_i^{(k+1)} = \psi_i^{(k)} + \eta(k)\big(x_i^* - \psi_i^{(k)}\big) + \gamma(k)\big(\xi_i^{(k)} - \psi_i^{(k)}\big), \tag{36}$$

where $\eta(k)$ promotes exploitation, $\gamma(k)$ controls exploration, and $\xi_i^{(k)} \sim \mathcal{U}(0,1)$. Collective probability shifts enable multi-bit transitions, producing tunneling-like effects.

Algorithm 2 Quantum-Inspired Simulated Annealing (QISA)

```
Initialize ψ_i = 0.5 and best solution x*
for k = 1 to K_max do
    Sample x from Bernoulli(ψ)
    Evaluate H(x) and update x*
    for each i do
        Draw ξ_i ~ U(0,1)
        Update ψ_i
    end for
end for
return x*
```

5.4 Limitations of Pure QISA

Although QISA enhances global exploration, it treats all variables uniformly. In structured hospital scheduling problems, some decisions have much larger impact than others. Uniform exploration may therefore reduce efficiency and slow convergence.

5.5 Neural-Guided QISA (QISA+NN)

To address this limitation, we introduce a neural network that estimates the relative importance of decision variables.

Given the current solution and system-level indicators (e.g., congestion, occupancy, workload), the neural network outputs:

$$w = (w_1, \ldots, w_n), \quad w_i \in [0,1], \tag{37}$$

where w_i approximates the local sensitivity:

$$w_i \approx \left| \frac{\partial H(x)}{\partial x_i} \right|. \tag{38}$$

These weights guide both sampling and tunneling dynamics. To further reduce infeasible solutions, neural guidance prioritizes variables contributing most to constraint violations, accelerating convergence toward feasible regions.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

5.6 Guided Update Dynamics

The probability update becomes:

$$\psi_i^{(k+1)} = \psi_i^{(k)} + \eta(k)\big(x_i^* - \psi_i^{(k)}\big) + \gamma(k) w_i \big(\xi_i^{(k)} - \psi_i^{(k)}\big). \tag{39}$$

High-impact variables receive stronger perturbations, focusing exploration on critical decisions.

The parameters follow adaptive schedules:

$$\alpha(k) = 1 - e^{-k/\tau}, \quad \gamma(k) = \gamma_0 e^{-k/\tau}, \quad \eta(k) = \eta_0(1 - e^{-k/\tau}). \tag{40}$$

Algorithm 3 Neural-Guided QISA (QISA+NN)

```
Initialize ψ_i = 0.5
Initialize x*
for k = 1 to K_max do
    Sample x from Bernoulli(ψ)
    Evaluate H(x)
    if H(x) < H(x*) then
        x* ← x
    end if
    Compute neural weights w = NN(x)
    for each variable i do
        Draw ξ_i ~ U(0, 1)
        Update ψ_i using guided rule
    end for
end for
return x*
```

6 Experimental Evaluation

6.1 Experimental Setup

We evaluate the QUBO–Choquet–QISA+NN framework on synthetic instances representing a medium-sized surgical center.

Each instance includes 20 patients, 4 operating rooms, 15 beds across clinical phases, and 8 staff members over 48 half-hour time slots. Patients follow multi-phase pathways with heterogeneous durations and priorities.

The resulting QUBO contains 65.280 binary variables ($2^{65.280}$ search space), integrating 8 hard constraint classes and 4 interacting soft criteria. Clinical parameters are sampled from realistic distributions (ICU probability: 40%). Results are averaged over five random seeds.

We compare three methods under identical budgets ($K_{\max} = 500$): **SA**, **QISA**, and **QISA+NN**. Metrics include total energy, hard violations, Choquet soft cost, allocation rate, and feasibility.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

6.2 Results

Table 1 reports mean results over five runs.

Table 1. Comparative performance (mean ± std)

Method	Energy	Hard Viol.	Choquet	Allocation	Feasible
SA	2143.7 ± 178.4	687.2	0.743	0.72	1/5
QISA	1582.4 ± 121.6	361.8	0.628	0.84	3/5
QISA+NN	$\mathbf{1238.9 \pm 93.7}$	**88.4**	**0.547**	**0.91**	**5/5**

QISA+NN achieves the lowest energy, minimal violations, highest allocation rate, and full feasibility. It converges faster than both SA and QISA, typically within 200 iterations.

Interaction effects represent 29% of the final soft cost, confirming that linear aggregation underestimates cascading congestion. Removing either neural guidance or Choquet aggregation degrades performance, demonstrating their complementary impact.

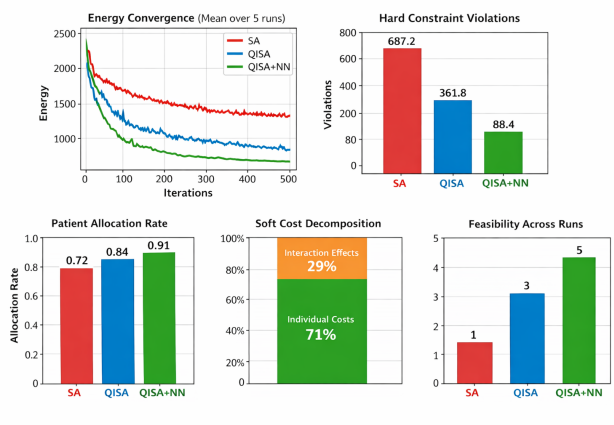

Fig. 1. Comparison of SA, QISA, and QISA+NN: energy convergence, final violations, allocation rate, Choquet interaction share, and feasibility. QISA+NN shows faster convergence and greater robustness.

6.3 Discussion

The experimental results provide several important insights into both the modeling and optimization components of the proposed framework.

First, the unified QUBO formulation successfully captures the strong interdependencies that naturally arise in hospital systems. Operating rooms, beds,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

and human resources cannot be managed independently: a delay in surgery propagates to bed occupancy, which in turn impacts staff workload. By embedding all constraints and criteria into a single quadratic energy function, the model internalizes these cascading effects rather than treating them sequentially. This explains why purely local improvements (as often performed by classical heuristics) are insufficient to reach high-quality global solutions.

Second, the Choquet aggregation plays a crucial role in representing nonlinear interactions between performance criteria. A simple weighted sum assumes that criteria contribute independently to the global objective. However, hospital congestion is rarely additive in practice. For example, simultaneous bed saturation and staff overload can amplify delays beyond what a linear model would predict. The observed 29% interaction contribution confirms that these nonlinear effects are not marginal but structurally significant. Removing the Choquet component leads to systematically higher congestion levels, highlighting the importance of interaction-aware aggregation.

Third, the optimization strategy itself strongly influences feasibility attainment. Standard SA struggles to escape poor local minima in such a high-dimensional binary landscape. The quantum-inspired extension (QISA) improves exploration through enhanced diversification mechanisms, reducing violations and improving allocation rates. However, the most significant improvement comes from neural guidance (QISA+NN), which biases the search toward structurally promising regions of the solution space. This hybridization accelerates convergence and stabilizes feasibility, achieving 100% feasible solutions within the computational budget.

From a practical perspective, these results suggest that large-scale integrated hospital scheduling problems require both expressive modeling and intelligent search mechanisms. A powerful model without adaptive optimization may remain computationally intractable, while a strong heuristic without interaction-aware modeling may overlook critical systemic effects.

Overall, the combination of QUBO reformulation, Choquet-based interaction modeling, and neural-guided quantum-inspired optimization provides a coherent and scalable framework capable of handling realistic congestion scenarios in medium-sized surgical centers.

7 Limitations and Future Work

The framework has several limitations. Experiments rely on synthetic data and do not fully capture real-world uncertainty (e.g., emergencies or variable durations), motivating validation on real hospital data and stochastic extensions.

The model assumes fixed clinical pathways, limiting flexibility for adaptive care processes. Moreover, the Choquet capacity is manually defined; learning these components and exploring hybrid classical–quantum implementations are promising future directions.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15

8 Conclusion

This paper introduced an integrated framework for hospital resource allocation based on a unified QUBO formulation, jointly optimizing operating rooms, beds, and staff under complex operational constraints.

A Choquet integral was used to model nonlinear interactions between performance criteria, and a neural-guided quantum-inspired simulated annealing algorithm was developed to efficiently solve the resulting large-scale problem.

Experiments on realistic synthetic instances show improved feasibility, solution quality, and convergence compared to classical and unguided approaches. Overall, the proposed framework demonstrates the value of combining QUBO modeling, non-additive aggregation, and learning-guided optimization for complex scheduling problems.

References

1. Guerriero, F., Guido, R.: Surgery planning and scheduling: A literature review. *European Journal of Operational Research* **262**(2), 2017.
2. Topaloglu, U., Brown, L.: Data-driven surgical planning and control. *Health Systems* **9**(2), 2020.
3. Liu, X., Wang, S., Chu, C.: Integrated operating room and inpatient bed scheduling. *Computers & Industrial Engineering* **154**, 2021.
4. Akpinar, S., Baykasoglu, A.: Hybrid metaheuristics for hospital resource scheduling. *Computers & Operations Research* **132**, 2021.
5. Yang, H., Chou, M.: Modeling congestion propagation in healthcare systems. *Health Care Management Science* **22**(4), 2019.
6. Zhang, Y., Sun, L.: System-level interactions in hospital operations. *European Journal of Industrial Engineering* **16**(3), 2022.
7. Lucas, A.: Ising formulations of many NP problems. *Frontiers in Physics* **2**, 2014.
8. Li, X., Chen, Y.: Learning Choquet integrals for interacting criteria. *Information Sciences* **547**, 2021.
9. Bengio, Y., Lodi, A., Prouvost, A.: Machine learning for combinatorial optimization: A methodological tour. *European Journal of Operational Research* **290**(2), 2021.
10. Li, P., Chen, W.: Learning to guide combinatorial search. *Journal of Artificial Intelligence Research* **73**, 2022.
11. Zhang, Y., Li, H., Sun, L.: Integrated optimization of hospital systems: A system-level review. *European Journal of Operational Research* **311**(1), 2023.
12. Kanamaru, S., Tanaka, S.: Advances in simulated bifurcation for large-scale QUBO optimization. *Physical Review Applied* **20**(4), 2023.
13. Zhou, Z., Wu, Y., Bengio, Y.: Neural guidance for large-scale combinatorial optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
14. Li, X., Zhao, Y., Chen, Y.: Learning non-additive Choquet capacities for interacting criteria. *Information Sciences* **654**, 2024.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_15