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.

Trust-Constrained Adaptive Relay Selection for Attack-Resilient Wireless Body Area Networks

Sam Nguyen-Xuan

samnx2@fe.edu.vn

FPT University

Research Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10908796/v1

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

Additional Declarations: No competing interests reported.

Trust-Constrained Adaptive Relay Selection for Attack-Resilient Wireless Body Area Networks

Sam Nguyen-Xuan[1]

Received: date / Accepted: date

Abstract This work proposes a trust-constrained adaptive relay-selection framework for wireless body area networks that separates relay eligibility from relay preference. Authentication, behavioural trust, anomaly evidence, and residual energy define the admissible relay set, after which a reinforcement-learning policy ranks only eligible relays. A channel-conditioned statistical trust update distinguishes likely channel losses from anomalous forwarding behaviour, while a separate anomaly state captures replay, injection, and abnormal resource use. Evaluation is driven by measured body-channel traces with a calibrated link-success model, while direct transmission is treated as a modelled reference. Across seven adversary models and 30 paired seeds, the proposed controller maintains 97.52–97.86% post-onset delivery under black-hole, selective-forwarding, on–off, and threshold-aware attacks, and 98.03–98.63% on three attacks withheld from policy training. Ablations show that lightweight constrained controllers can match or exceed recurrent proximal policy optimization, indicating that hard security admissibility, rather than policy complexity, provides the main source of attack resilience.

Keywords Wireless Body Area Networks · Secure Relay Selection · Trust Management · Reinforcement Learning · Attack Resilience

1 Introduction

Wireless body area networks (WBANs) support continuous health monitoring, activity recognition, rehabilitation, and context-aware computing using resource-constrained sensors placed on or around the human body, with sensed data transmitted to a coordinator such as a smartphone. Limited battery capacity, changing posture, mobility, and interference all degrade reliability and operational lifetime [1]. Because the human body obstructs propagation, the direct sensor-to-coordinator link is frequently unreliable, and relay-assisted forwarding has been adopted to improve delivery [2–5]. Relay selection is dynamic by necessity: a forwarding node that is effective at one instant may be poorly placed shortly afterwards, and repeatedly selecting the same relay depletes its battery. Bilandi *et al.* accordingly formulated WBAN relay selection as a multi-criteria decision problem over residual energy, traffic load, node density, signal-to-noise ratio, distance, and node criticality [6].

Communication and energy criteria are not sufficient, however, when data traverses an intermediate node. A relay occupies a privileged position and can discard packets, forward selectively, modify or inject readings, replay earlier traffic, impersonate a legitimate node, or exhaust network resources. Trust management has therefore been identified as an important complement to cryptographic protection in WBANs [7], and recent routing work has begun to treat security as a routing-layer concern rather than a separate network function [8–11].

A limitation persists across both lines of work. Relay-selection methods optimise channel quality, energy, or predicted conditions while assuming cooperative candidates. Trust-aware methods identify malicious nodes but generally incorporate trust as one weighted criterion among many. The second choice has a specific consequence: a poorly trusted relay remains selectable whenever its channel or energy characteristics are sufficiently favourable, which is precisely the condition an adversary is able to arrange. A relay with substantial residual energy and an excellent link may nonetheless be compromised.

Sam Nguyen-Xuan
samnx2@fe.edu.vn

1 Department of Computer Science, Swinburne Vietnam, FPT University, A35 Bach Dang, Ho Chi Minh City, 70000, Vietnam

This paper addresses this limitation through a *trust-constrained adaptive relay-selection* framework. The central design principle is to separate *security admissibility* from *communication preference*. Candidate relays are first assessed using authentication state, behavioural trust, anomaly evidence, and a residual-energy survivability floor. A relay that fails these requirements is removed from the admissible action set and cannot be selected regardless of its channel quality, available energy, or predicted utility. Only after this security screening does an adaptive controller determine preference among the remaining candidates using communication and resource state.

A reinforcement-learning controller is used to investigate whether temporal decision making provides an additional benefit after security admissibility has been enforced. Its observations include channel quality, forwarding behaviour, residual energy, traffic load, trust state, anomaly evidence, eligibility, channel confidence, and the age of relay observations. However, security is not represented as a favourable reward that can trade against communication performance. Instead, the admissibility constraint is applied directly to the action space. This distinction is important because reward-weighted trust can still permit selection of a risky relay when another objective provides a sufficiently large compensating benefit.

Energy is likewise treated conservatively. The framework does not claim that two-hop relaying reduces total communication energy relative to direct transmission. Under the adopted on-body radio model, relay reception introduces additional energy cost that can exceed the amplifier saving obtained from shorter transmission distances. Residual energy is therefore used as a survivability floor and a relay-preference variable, while energy consumption and delay are reported explicitly as operating costs of secure relaying. The main contributions are:

- A security-first relay-selection formulation that separates relay eligibility from relay preference. Authentication, behavioural trust, anomaly state, and survivability define a state-dependent admissible action set, ensuring that favourable communication conditions cannot compensate for an unacceptable security state.
- A channel-conditioned statistical trust mechanism that evaluates observed forwarding performance relative to expected benign link reliability. This reduces penalties for losses consistent with body-channel degradation while penalising statistically significant forwarding shortfalls, thereby reducing false exclusion of legitimate relays.
- A complementary anomaly mechanism for behaviours that cannot be inferred reliably from packet delivery alone, including replay, injection, abnormal resource consumption, and unauthorised relay participation.
- An adaptive relay-preference controller operating only over the admissible set, together with recurrent, feed-forward, permutation-sensitive, best-channel, and posterior sampling variants that separate the contribution of security admissibility from that of policy architecture and temporal learning.
- A comprehensive evaluation under seven adversary models using paired experiments over 30 seeds. Three adversary classes are withheld entirely from policy training, permitting explicit evaluation of attack generalisation, while false exclusion, attack-resilient delivery, delay, throughput, energy cost, and direct-link fallback are reported separately.
- An empirical finding that the hard security constraint, rather than increased policy complexity, accounts for most of the observed attack resilience. This result identifies lightweight constrained controllers as a practical deployment alternative when recurrent learning provides no statistically significant advantage.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the system and threat models, Section 4 the proposed framework, and Section 5 the experimental design. Section 6 reports results and Section 7 concludes.

2 Related Work

2.1 Relay Selection in WBANs

Nguyen *et al.* proposed an adaptive relay transmission scheme in which candidate relays are formed from nodes that successfully decode the source packet and the candidate with the highest channel gain retransmits when direct transmission fails [12]. The scheme separates channel-related from collision-related failures, establishing that relay participation should follow current conditions rather than a fixed assignment. Subsequent work turned to lifetime objectives. Zhang *et al.* formulated relay selection under the IEEE 802.15.6 two-hop topology as network-lifetime maximisation over energy-consumption rate and residual-energy disparity, solved heuristically [13], and later proposed an adaptive mechanism that reforms the relay topology as node energy and mobility change [14]. Bilandi *et al.* ranked candidates by hybrid AHP–TOPSIS over six communication and resource criteria [6], and Singh and Bilandi extended this to neutrosophic AHP–TOPSIS in ARAP [15]. These

criteria characterise communication quality and resource availability, but none determines whether a candidate forwards honestly.

2.2 Learning-Based Relay and Routing Decisions

Mohammadi and Shirmohammadi modelled active-node-group selection in energy-harvesting WBANs as a Markov decision process solved by Q-learning, with active nodes able to serve as relays [16]. Pan *et al.* combined deep incremental human-body channel prediction with relay selection over predicted channel quality and energy [17], advancing beyond schemes that react only to instantaneous measurements. In the broader wireless-sensor setting, Zhou *et al.* proposed DQN- and PPO-based relay-selection schemes, demonstrating that relay selection admits a sequential-decision formulation [18]. This class of method suits WBANs because body-channel quality, battery depletion, traffic demand, and relay behaviour all have temporal dependencies that a single observation does not reveal. It does not, however, transfer directly to security-critical settings: all of these methods assume candidate relays are legitimate, and a relay with favourable predicted conditions may still be compromised.

2.3 Trust Management and Security-Aware Routing

Authentication establishes identity, but an authenticated relay may subsequently be compromised. Ayed *et al.* surveyed WBAN trust management and identified behavioural trust, derived from direct observation, forwarding performance, or reputation, as a necessary complement [7]. Forwarding is particularly exposed to black-hole, grey-hole, and selective-forwarding attacks, and to replay, injection, impersonation, resource exhaustion, and on–off behaviour, the last of which is difficult to detect because the node behaves correctly for extended periods. Dass *et al.* addressed malicious-node avoidance alongside energy, delivery, and delay in body-area sensor networks [9], Fu *et al.* combined energy-aware routing with two-way trust evaluation [10], Fei *et al.* incorporated trust and QoS factors into secure WSN routing [11]. More recently, Gupta *et al.* integrated DQN-based routing, trust management, intrusion detection, and lightweight cryptography for secure WSN routing [19]. That work is important for positioning the present study: it shows that trust, security, and deep reinforcement learning have already been combined in general WSN routing, so the contribution here is not that combination but the specific role assigned to trust, a hard admissibility constraint rather than a weighted criterion, and its formulation for body-centric links.

2.4 Positioning

Four directions are relevant: adaptive cooperative selection [12], energy- and lifetime-aware selection [6, 13–15], learning-based and predictive adaptation [16–18], and trust- and security-aware routing [7, 9, 10, 19]. Their intersection remains insufficiently addressed for WBAN relay selection.

Predictive WBAN mechanisms optimise channel quality and energy, trust-aware schemes target broader routing problems without relay-level behavioural evidence on body-centric links, and where trust does appear, it appears as an optimisation criterion that favourable channel conditions can outweigh. The proposed framework instead makes authentication, behavioural trust, and anomaly state determine *eligibility*, and leaves channel, energy, reliability, and load to determine *preference* among eligible candidates.

This remains the prevailing formulation in recent work. Arafat *et al.* apply Q-learning to QoS-aware routing in IoMT-enabled WBANs, selecting paths over combined link-reliability, priority, and energy-balancing criteria [20]. Shyja *et al.* compute adaptive trust with a recurrent network and use the resulting score to rank candidate cluster heads [21]. In both cases trust contributes to a ranking that other criteria can offset. The distinction drawn in the present work is therefore not that trust is estimated more accurately, but that it determines admissibility rather than rank. Table 1 positions the framework against representative methods.

3 System and Threat Models

3.1 Network Model and Notation

The network comprises body sensors $\mathcal{S} = \{S_1, \ldots, S_N\}$, candidate wearable relays $\mathcal{R} = \{R_1, \ldots, R_M\}$, and a single coordinator C, typically a smartphone carried by the user. Time is slotted, and t denotes the current relay-selection epoch. All selection decisions are made by C, which observes channel conditions and forwarding outcomes and therefore acts as the trust authority (Figure 1).

For each candidate relay R_i, the coordinator maintains a ten-dimensional observation vector

$$\mathbf{x}_i(t) = \big[E_i(t),\, Q_i(t),\, F_i(t),\, L_i(t),\, T_i(t),\, A_i(t),\, c_i(t),\, m_i(t),\, s_i(t),\, v_i\big]^\top, \tag{1}$$

Table 1 Comparison of representative relay-selection and secure-routing approaches. "Elig." denotes a hard admissibility constraint, as distinct from a weighted trust criterion.

Study	WBAN	Adaptive	Pred./ML	Energy	Trust	Attack	Elig.
Nguyen *et al.* [12]	Yes	Yes	No	No	No	No	No
Zhang *et al.* [13]	Yes	Limited	No	Yes	No	No	No
Zhang *et al.* [14]	Yes	Yes	No	Yes	No	No	No
Bilandi *et al.* [6]	Yes	Limited	No	Yes	No	No	No
Mohammadi *et al.* [16]	Yes	Yes	RL	Yes	No	No	No
Pan *et al.* [17]	Yes	Yes	Deep/Incr.	Yes	No	No	No
Singh and Bilandi [15]	Yes	Yes	MCDM	Yes	No	No	No
Zhou *et al.* [18]	No	Yes	DQN/PPO	Yes	No	No	No
Gupta *et al.* [19]	No	Yes	DQN	Yes	Yes	Yes	Limited

where $E_i(t)$ denotes normalised residual energy, $Q_i(t)$ link quality, $F_i(t)$ forwarding reliability, $L_i(t)$ relay load, $T_i(t)$ behavioural trust, and $A_i(t)$ the anomaly score. The additional variables $c_i(t)$, $m_i(t)$, and $s_i(t)$ denote benign channel-success confidence, the current eligibility indicator, and normalised observation staleness, respectively. The validity indicator $v_i \in \{0,1\}$ distinguishes a real relay entry from a zero-padded position when a policy is evaluated with fewer than the maximum number of candidate relays.

Staleness is included because trust is updated only when a relay provides new forwarding evidence. Consequently, the trust state of an unselected relay may remain unchanged for several epochs even though the underlying information becomes progressively older. This observation age is therefore relevant to the preference controller.

In addition to the relay-specific states, the coordinator observes the global context

$$\mathbf{g}(t) = \left[\frac{t}{H}, \frac{|\mathcal{R}_{\text{safe}}(t)|}{M}, d_{\text{fb}}(t)\right]^{\top}, \tag{2}$$

where H is the episode horizon and $d_{\text{fb}}(t) \in \{0,1\}$ indicates that no admissible relay remains and direct-link fallback is therefore active. The complete policy observation is

$$\mathbf{o}(t) = \left(\{\mathbf{x}_i(t)\}_{i=1}^{M}, \mathbf{g}(t)\right). \tag{3}$$

All state variables are time dependent because body posture, channel conditions, relay load, residual energy, and observed forwarding behaviour change during operation. Table 2 summarises the principal notation.

Table 2 Principal notation.

Symbol	Meaning
$\mathcal{S}, \mathcal{R}, C$	sensors, relays, coordinator
N, M	numbers of sensors and relays
E_i, Q_i	residual energy, link quality
F_i, L_i	forwarding reliability, relay load
T_i, A_i	trust, anomaly score
c_i	benign channel confidence
m_i	eligibility indicator
s_i	observation staleness
v_i	valid-relay indicator
$\mathbf{x}_i$	relay observation vector
$\mathbf{g}$	global context vector
d_{fb}	direct-fallback indicator
Auth_i	authentication state
$\mathcal{R}_{\text{safe}}$	admissible relay set
n_i, k_i	assigned/delivered packets
ψ_i	forwarding shortfall score
h_t, a_t	hidden state, action
$R^*(t)$	selected relay
$T_{\min}, E_{\min}, A_{\max}$	eligibility thresholds
α, β	trust recovery/penalty rates
$z_{\text{tol}}, z_{\text{sat}}$	shortfall thresholds
λ	anomaly memory factor
H	episode horizon

3.2 Threat Model

Sensors and the coordinator are assumed to be trusted, while candidate wearable relays constitute the potentially untrusted forwarding layer. A relay may be compromised after successful authentication and may then drop all assigned packets, forward selectively, alternate between cooperative and malicious periods, replay stale packets, inject unauthorised traffic, or generate abnormal activity intended to consume communication resources. A rogue relay may also attempt to participate without a valid authentication state.

The adversary may compromise one or more candidate relays and coordinate their behaviour over time. It is assumed to know the relay-selection procedure, trust-update rule, eligibility thresholds, and policy structure. In particular, an adaptive relay may regulate its defection rate or alternate between benign and malicious phases in an attempt to remain eligible. The coordinator does not observe the latent compromise state directly, it observes only authentication state, link conditions, forwarding outcomes, and independently detectable security events.

The adversary cannot compromise the coordinator, modify the trust state stored at the coordinator, forge a valid authentication state, or directly manipulate the coordinator's measurements. Sensors are also assumed uncompromised. The study therefore addresses malicious or misbehaving forwarding relays rather than end-device or coordinator compromise.

This distinction is important in WBANs because missing packets are not necessarily malicious: posture changes, body shadowing, mobility, and interference can reduce delivery even when a relay behaves correctly. The security mechanism must therefore distinguish unexplained forwarding degradation from loss consistent with the measured channel state. Physical layer jamming, wideband denial-of-service, sensor compromise, coordinator compromise, cryptographic key extraction, and attacks on the authentication protocol itself are outside the present scope. The seven evaluated adversary behaviours are specified in Section 5.3.

4 Proposed Approach

4.1 Channel-Conditioned Statistical Trust and Anomaly Estimation

Packet loss in a WBAN is not, by itself, evidence of malicious forwarding. Shadowing, posture changes, mobility, and interference can reduce delivery even when a relay behaves correctly. The trust mechanism therefore evaluates each selected relay relative to the benign delivery rate expected from the current body-channel condition.

Suppose that $n_i(t)$ packets are assigned to relay R_i during epoch t and that $k_i(t)$ are successfully delivered. Let $c_i(t) = g(Q_i(t)) \in (0, 1)$ denote the estimated probability of successful delivery under benign operation at the observed link quality. Under the benign null model,

$$K_i(t) \sim \text{Binomial}(n_i(t), c_i(t)) \,. \tag{4}$$

Its mean and standard deviation are

$$\mu_i(t) = n_i(t)c_i(t), \qquad \sigma_i(t) = \sqrt{n_i(t)c_i(t)\,[1 - c_i(t)]}. \tag{5}$$

The standardised delivery shortfall is

$$z_i(t) = \frac{\mu_i(t) - k_i(t)}{\max(\sigma_i(t), \varepsilon)}, \tag{6}$$

where $\varepsilon > 0$ prevents numerical instability. The statistical evidence is converted to a graded shortfall score

$$\psi_i(t) = \text{clip}\left(\frac{z_i(t) - z_{\text{tol}}}{z_{\text{sat}} - z_{\text{tol}}}, 0, 1\right), \tag{7}$$

where z_{tol} is a tolerance region for ordinary channel variation and z_{sat} is the point at which the deficit is treated as fully significant.

Trust is updated only for a selected relay when forwarding evidence is available:

$$T_i(t{+}1) = \begin{cases} \min(1,\ T_i(t) + \alpha[1 - T_i(t)])\,, & \psi_i(t) = 0, \\ \max(0,\ T_i(t) - \beta\, c_i(t)\psi_i(t)T_i(t))\,, & \psi_i(t) > 0. \end{cases} \tag{8}$$

If R_i is not selected, its trust value is unchanged. The asymmetric rates $0 < \alpha < \beta < 1$ give gradual recovery when forwarding is statistically consistent with the benign channel model and faster degradation when a deficit is difficult to explain by channel variation. The additional factor $c_i(t)$ reduces the penalty when the underlying body channel is itself unreliable.

For interpretation and adversary construction, consider the corresponding mean-drift approximation of an asymmetric binary update under constant channel confidence c. Its reference defection boundary is

$$p^*(c) = \frac{\alpha(1 - T_{\min})}{\alpha(1 - T_{\min}) + \beta c T_{\min}}. \tag{9}$$

Equation (9) is not an exact exclusion theorem for the statistical detector of Eqs. (6)–(8). It is used only as a mean-drift reference to parameterise the threshold-aware adversary in Section 5.3, actual exclusion is determined by the statistical shortfall and the accumulated trust state.

Delivery-based trust does not capture security events that may occur while legitimate forwarding continues. Let $u_i(t)$ denote the number of independently observable security events associated with R_i during epoch t, including replay indicators, unexpected source or sequence behaviour, and abnormal activity associated with resource-exhaustion attempts. The anomaly state is updated as

$$A_i(t + 1) = \lambda A_i(t) + (1 - \lambda) \min\left(1, \frac{u_i(t)}{u_{\max}}\right), \tag{10}$$

where $\lambda \in (0, 1)$ controls estimator memory. Authentication is maintained as a separate state. An unauthenticated rogue relay has $\text{Auth}_i(t) = 0$, while a high anomaly score excludes a relay through the eligibility rule without being interpreted as automatic credential revocation.

The channel-confidence mapping $g(\cdot)$ is fitted on the calibration partition only and frozen before training and final evaluation. Because the available channel file does not contain verified packet-level ACK ground truth, the final experiment uses a monotonic, calibration-shaped logistic mapping whose mean predicted benign

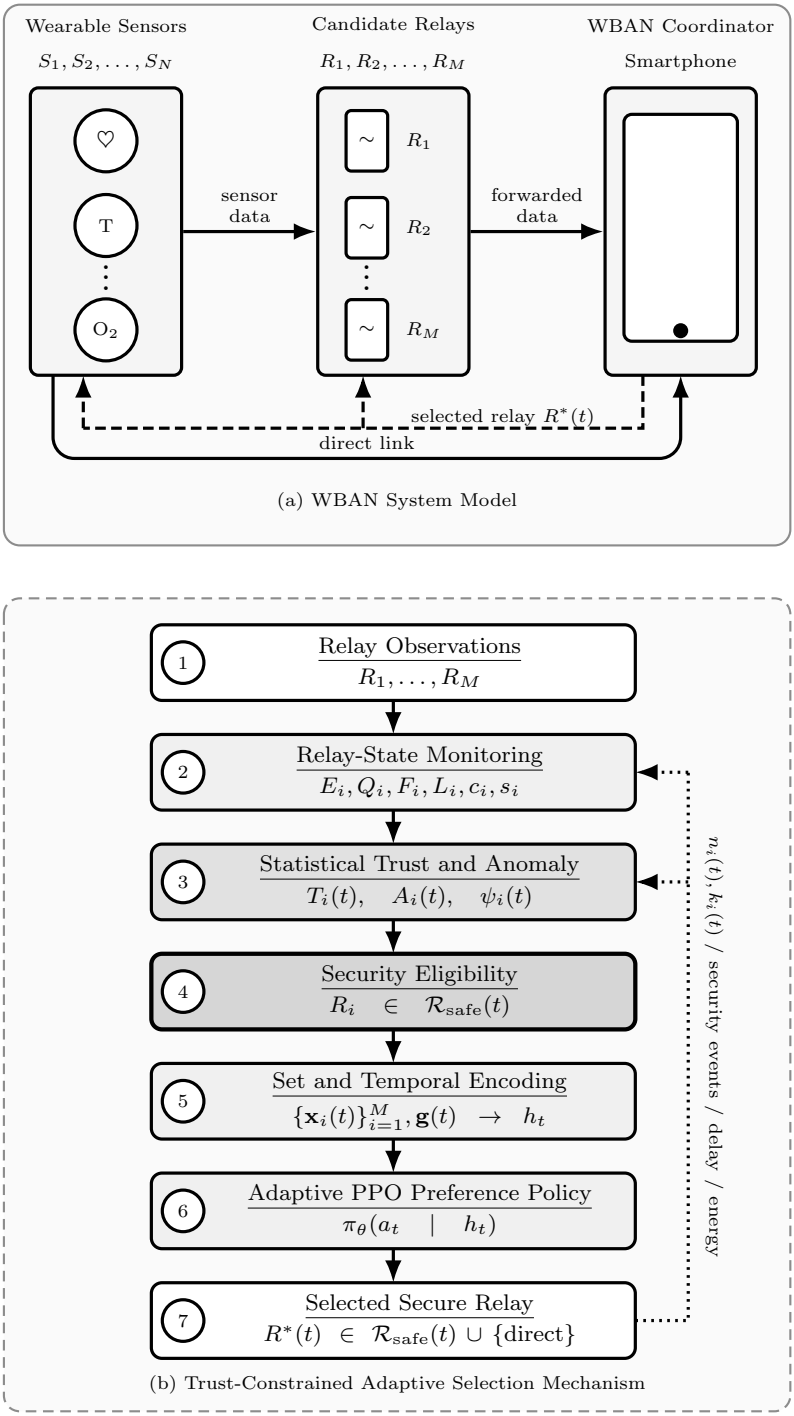


Fig. 1 Secure WBAN relay selection. (a) System model: sensors transmit either directly to the coordinator or through candidate relays, and the coordinator determines $R^*(t)$. (b) Coordinator-side mechanism. Line style carries meaning: solid arrows are payload and pipeline flow, dashed lines the control bus, dotted lines the feedback bus. The heavier rule on stage 4 marks the hard eligibility constraint.

Algorithm 1: Statistical trust and anomaly update at the coordinator

Input: $n_i(t)$, $k_i(t)$, $c_i(t)$, $u_i(t)$, $\alpha, \beta, \lambda, z_{\mathrm{tol}}, z_{\mathrm{sat}}$
Output: $T_i(t+1)$, $A_i(t+1)$

foreach $R_i \in \mathcal{R}$ **do**
  if R_i *was selected and* $n_i(t) > 0$ **then**
    $\mu_i \leftarrow n_i(t)c_i(t)$;
    $\sigma_i \leftarrow \sqrt{\max(\mu_i[1-c_i(t)], \varepsilon)}$;
    $z_i \leftarrow [\mu_i - k_i(t)]/\sigma_i$;
    $\psi_i \leftarrow \mathrm{clip}((z_i - z_{\mathrm{tol}})/(z_{\mathrm{sat}} - z_{\mathrm{tol}}), 0, 1)$;
    if $\psi_i = 0$ **then**
      $T_i(t+1) \leftarrow \min(1, T_i(t) + \alpha[1 - T_i(t)])$;
    else
      $T_i(t+1) \leftarrow \max(0, T_i(t) - \beta c_i(t)\psi_i T_i(t))$;
    $A_i(t+1) \leftarrow$
     $\lambda A_i(t) + (1-\lambda)\min(1, u_i(t)/u_{\max})$;
  else
    $T_i(t+1) \leftarrow T_i(t)$;
    $A_i(t+1) \leftarrow A_i(t)$;

success on the calibration segment is fixed at 0.95. This modelling assumption is reported explicitly and is examined as a limitation in Section 7.

4.2 Security Eligibility

Eligibility is evaluated before and independently of communication preference:

$$\mathcal{R}_{\mathrm{safe}}(t) = \Big\{ R_i \in \mathcal{R} \;\Big|\; \mathrm{Auth}_i(t) = 1,\ T_i(t) \geq T_{\min},\ A_i(t) \leq A_{\max},\ E_i(t) \geq E_{\min} \Big\}. \tag{11}$$

The first three conditions represent security admissibility. The energy condition is a survivability floor rather than a security test: a relay near exhaustion should not be assigned additional traffic even if it remains trustworthy. Energy otherwise enters the preference decision as a graded resource variable.

The key distinction is that a relay outside $\mathcal{R}_{\mathrm{safe}}(t)$ is not assigned a low score, it is removed from the admissible action set. Consequently, favourable channel quality, residual energy, or low delay cannot compensate for failure of an authentication, trust, or anomaly requirement.

The implementation uses direct transmission only as a fallback. If at least one admissible relay exists, the action mask contains only those relays. If $\mathcal{R}_{\mathrm{safe}}(t) = \emptyset$, relay actions are disabled and the direct sensor-to-coordinator action is enabled. The direct-fallback rate is reported because excessive fallback would indicate that security is being obtained by avoiding relaying rather than by selecting trustworthy relays.

4.3 Relay Selection as a Constrained Decision Process

The coordinator does not observe the latent compromise state directly, so relay preference is modelled as a partially observable sequential decision problem. The admissible action set is state dependent:

$$\mathcal{A}(t) = \begin{cases} \{\, i : R_i \in \mathcal{R}_{\mathrm{safe}}(t) \,\}, & \mathcal{R}_{\mathrm{safe}}(t) \neq \emptyset, \\ \{0\}, & \mathcal{R}_{\mathrm{safe}}(t) = \emptyset, \end{cases} \tag{12}$$

where action 0 denotes direct transmission and action $i \geq 1$ denotes relay R_i. Thus, direct transmission is a fallback rather than a competing action when an admissible relay is available.

The proposed learned preference controller uses a permutation-equivariant set representation. Each relay observation $\mathbf{x}_i(t)$ is encoded by a shared multilayer perceptron. Mean and maximum pooling over valid relay embeddings are concatenated with the global context $\mathbf{g}(t)$ to produce a set summary. This avoids learning arbitrary relay-index order and permits the same network to operate with different numbers of valid relay entries.

To test whether temporal memory provides additional value after the trust state has already summarised forwarding history, the proposed policy augments the set representation with a gated recurrent unit:

$$\begin{aligned} h_t &= f_\phi(h_{t-1}, \mathbf{o}(t), \mathrm{onehot}(a_{t-1}), r_{t-1}), \\ a_t &\sim \pi_\theta(\cdot \mid h_t). \end{aligned} \tag{13}$$

Recurrence is therefore an evaluated architectural choice rather than an assumed source of security. A feedforward ablation uses the same set encoder without recurrent memory.

For the proposed controller, trust does not appear as a positive reward term. Security is imposed by Eq. (12). The epoch reward used for offline policy training is

$$\begin{aligned} r_t = w_{\mathrm{del}} D_t - w_{\mathrm{energy}} \bar{e}_t - w_{\mathrm{delay}} \bar{d}_t \\ - w_{\mathrm{load}} \bar{L}_t - w_{\mathrm{defect}} D_t^{\mathrm{adv}} + \gamma \Phi(t+1) - \Phi(t), \end{aligned} \tag{14}$$

where D_t is the fraction of generated packets delivered during the epoch, $\bar{e}_t$ and $\bar{d}_t$ are normalised energy and delay, $\bar{L}_t$ is the selected-relay load, and D_t^{adv} is the fraction of packet losses caused by the simulated forwarding adversary. The potential

$$\Phi(t) = w_{\mathrm{life}} \min_i E_i(t) \tag{15}$$

provides potential-based shaping on the worst residual relay energy.

The adversarial-loss term is a privileged simulator-training signal: it is available while policies are trained

Algorithm 2: Trust-constrained relay selection at one epoch

Input: $\{\mathbf{x}_i(t)\}$, $\mathbf{g}(t)$, policy π_θ, hidden state h_{t-1}
Output: selected route, updated hidden state h_t

$\mathcal{R}_{\text{safe}}(t) \leftarrow \{R_i : \text{Auth}_i = 1,\ T_i \geq T_{\min},\ A_i \leq A_{\max},\ E_i \geq E_{\min}\}$;
if $\mathcal{R}_{\text{safe}}(t) = \emptyset$ **then**
  select direct transmission;
else
  encode relay set and global context;
  $h_t \leftarrow f_\phi(h_{t-1}, \mathbf{o}(t), a_{t-1}, r_{t-1})$;
  mask policy logits to $\mathcal{R}_{\text{safe}}(t)$;
  sample/select $a_t \sim \pi_\theta(\cdot \mid h_t)$;
  select $R^*(t) \leftarrow R_{a_t}$;
transmit, observe delivery, delay, energy, channel state, and security events;
update T_i and A_i using Algorithm 1;
return *selected route,* h_t

offline in the controlled attack environment but is not included in the execution-time observation. At deployment, selection uses only the observable state in Eqs. (1)–(3) and the hard eligibility mask. A separate *reward-trust* ablation removes the trust/anomaly hard mask and adds a positive trust term to the reward. It is included specifically to distinguish soft reward shaping from non-compensable admissibility.

Action masking is an established mechanism for state-dependent action validity in policy-gradient methods. Huang and Ontañón showed that masking invalid actions yields a valid policy gradient rather than a biased approximation, and that it scales substantially better than penalty-based handling as the proportion of invalid actions grows [22]. The safe-reinforcement-learning literature frames the same design choice in general terms: unsafe behaviour may either be penalised within the objective or excluded from the action space, and only the latter makes an unsafe action unreachable rather than merely unattractive [23]. The present work adopts the second formulation, with the security state of Eq. (11) supplying the validity condition at each epoch. The *reward-trust* ablation of Section 5.4 instantiates the first, which is what makes the two directly comparable in Section 6.

4.4 Policy Optimisation and Complexity

The learned controllers are trained with proximal policy optimisation (PPO). For the recurrent variant, trajectories are preserved as ordered sequences rather than shuffled transitions so that the hidden state is reconstructed consistently during optimisation. Generalised advantage estimation is used, and the clipped policy objective is combined with a clipped value loss and an entropy bonus:

$$\mathcal{L}(\theta) = \mathbb{E}_t\Big[\min\Big(\rho_t(\theta)\hat{A}_t,\ \text{clip}\big(\rho_t(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_t\Big)\Big] - c_1\mathcal{L}^V(\theta) + c_2\mathcal{H}[\pi_\theta]. \tag{16}$$

where $\rho_t(\theta) = \pi_\theta(a_t \mid h_t)/\pi_{\theta_{\text{old}}}(a_t \mid h_t)$. Action masks are stored with the trajectory and reapplied during optimisation, so inadmissible actions remain impossible during both training and execution. The learning rate and entropy coefficient are annealed during training, and the final checkpoint is selected using held-out validation performance rather than final-test results.

The permutation-equivariant actor–critic encodes each relay with the same two-layer multilayer perceptron. Masked mean and maximum pooling create the set-level context, while a relay-specific head combines each relay embedding with the shared context state to produce one logit per candidate. A separate head represents direct fallback. With hidden width d, relay encoding and pooling scale linearly with M, while the recurrent update contributes $O(d^2)$. The resulting per-decision complexity is $O(Md + d^2)$. All inference executes at the coordinator. Candidate relays perform ordinary forwarding only and incur no additional learning computation.

5 Experimental Design

5.1 Data Sources

The radio environment and adversarial process are modelled separately. Link-quality evolution is driven by the body-area channel measurements of Smith *et al.* [24], collected using the wearable channel-sounding platform described in [25]. The prepared channel bank contains 60,001 samples for four candidate-relay traces and is replayed chronologically so that temporal channel dependence is preserved.

The available channel file does not provide verified packet-level ACK labels. Accordingly, the recorded "success" field is not treated as delivery ground truth. Instead, the calibration segment is used to construct the monotonic channel-success mapping $c_i(t) = g(Q_i(t))$, where $Q_i(t)$ is the measured channel-quality state of relay i. The final experiment uses the calibration-shaped logistic mapping described in Section 4.1, with the mean benign link-success probability set to 0.95. The mapping is frozen before policy training and final evaluation.

PAMAP2 [26] provides activity-regime information used to modulate traffic intensity and is not treated as a radio-channel dataset.

The prepared channel bank is divided chronologically into calibration, training, validation, and final-test segments using 15%, 55%, 15%, and 15% of the samples, respectively. No policy, trust parameter, or model hyperparameter is selected using the final-test segment. Because these partitions are contiguous segments of one prepared recording bank rather than independent recordings, the split reduces direct temporal overlap but does not constitute recording-level external validation. This limitation is discussed further in Section 7.

The measured channel bank contains candidate-relay traces but does not provide a matched sensor-to-coordinator direct-link trace. Direct transmission is therefore treated as a controlled reference scenario rather than as measured direct-link ground truth. Rather than inferring direct-link quality from the concurrent relay traces, its packet-success probability is parameterised explicitly as

$$P\big(D_t^{\mathrm{dir}} = 1\big) = p_{\mathrm{d}}, \tag{17}$$

where p_{d} denotes the assumed direct-link packet-success probability.

To determine when relay-assisted forwarding is beneficial, the direct-link reference is evaluated over

$$\mathcal{P}_{\mathrm{d}} = \{0.70,\, 0.80,\, 0.90,\, 0.95,\, 0.975,\, 0.985,\, 0.99\}. \tag{18}$$

All other traffic, attack, and random-seed conditions are held fixed across this sensitivity analysis. This formulation does not assume that relaying is preferable to direct communication. Instead, it identifies the operating region in which degradation of the direct body-area link is sufficient to justify the additional energy and delay cost of relay-assisted forwarding. When the direct link is sufficiently reliable, direct transmission remains the preferred operating mode.

5.2 Simulation Environment

A discrete-event simulator implements the sensor–relay–coordinator topology of Section 3.1. The nominal configuration contains six sensors, four candidate relays, and one trusted coordinator. Relay selection occurs once per second and each evaluation episode spans 3600 epochs. Each sensor has a nominal generation rate of one 64-byte packet per second. The epoch-level packet count is Poisson distributed around the activity modulated aggregate rate, while PAMAP2 context scales the traffic multiplier between 0.6 and 1.6.

Sensors and relays start with 0.5 J. The source–relay, relay–coordinator, and modelled direct distances are 0.60, 0.80, and 1.60 m, respectively. The radio-energy parameters follow the Nordic nRF2401 transceiver model reported for short-range body-area communication [27, 28]. The transmitter and receiver electronics consume $E_{\mathrm{TX,elec}} = 16.7$ nJ/bit and $E_{\mathrm{RX,elec}} = 36.1$ nJ/bit, respectively. Transmission over distance d is modelled as

$$E_{\mathrm{TX}}(b, d) = bE_{\mathrm{TX,elec}} + bE_{\mathrm{amp}}d^n, \tag{19}$$

where $E_{\mathrm{amp}} = 1.97$ nJ/bit/m^2 and the present energy-accounting model uses the quadratic distance exponent $n = 2$.

Trust parameters are selected on validation data only. Grid search over $\beta \in \{0.10, 0.20, 0.30, 0.40\}$ and $T_{\min} \in \{0.45, 0.50, 0.60, 0.70\}$ selects $\beta = 0.10$ and $T_{\min} = 0.45$ for the final test. The remaining statistical-trust parameters are $\alpha = 0.02$, $z_{\mathrm{tol}} = 1.5$, and $z_{\mathrm{sat}} = 3.0$. The anomaly parameters are $\lambda = 0.90$, $A_{\max} = 0.60$, and $u_{\max} = 5$.

Table 3 summarises the nominal configuration. Additional experiments vary the trust penalty and threshold, and a relay-count sweep is performed for $M \in \{2, 4\}$.

Table 3 Nominal simulation and learning parameters.

Parameter	Value
Sensors N	6
Candidate relays M	4
Decision interval	1 s
Episode length	3600
Packet size	64 B
Nominal rate	1 pkt/s/sensor
Initial energy	0.5 J
Energy floor $E_{\min}$	0.10
Target benign success	0.95
Trust recovery α	0.02
Trust penalty β	0.10
Trust threshold $T_{\min}$	0.45
$z_{\mathrm{tol}}/z_{\mathrm{sat}}$	1.5 / 3.0
Anomaly memory λ	0.90
Anomaly threshold $A_{\max}$	0.60
$E_{\mathrm{TX,elec}}$	16.7 nJ/bit
$E_{\mathrm{RX,elec}}$	36.1 nJ/bit
E_{amp}	1.97 nJ/bit/m^2
Path-loss exponent n	2
GRU hidden width	64
PPO clip ϵ	0.20
Discount γ	0.99
GAE λ_{GAE}	0.95
Learning rate	3×10^{-4}
Train / validation episodes	3000 / 300
PPO rollout / sequence	512 / 128
PPO epochs/update	4
Entropy coefficient	$0.02 \rightarrow 0.002$
$w_{\mathrm{del}}/w_{\mathrm{energy}}$	2.0 / 0.8
$w_{\mathrm{delay}}/w_{\mathrm{load}}$	0.5 / 0.25
$w_{\mathrm{defect}}/w_{\mathrm{life}}$	2.0 / 1.0
Attack onset	20%
Final paired seeds	30

5.3 Adversary Models

WBAN relays may be exposed to forwarding, authentication, availability, and trust-manipulation attacks because wearable forwarding devices operate over an open medium and may be physically compromised [29–31]. Seven adversary models are evaluated:

- **Black-hole:** a compromised relay drops every packet assigned to it after attack onset [29, 31].
- **Selective forwarding:** the relay independently drops assigned packets with probability p. The nominal final-test value is $p = 0.30$, while training randomises p over $[0.10, 0.60]$.
- **On–off:** the relay alternates cooperative and malicious periods. The nominal final test uses 40 cooperative epochs followed by 10 malicious epochs, while training randomises both durations.
- **Threshold-aware:** the adversary knows the trust parameters and current channel confidence. Its packet-level defection probability is set to $0.95p^*(c_i(t))$, where p^* is the mean-drift reference of Eq. (9). This deliberately operates near, but slightly below, the reference admissibility boundary.
- **Replay/injection:** a compromised relay continues legitimate forwarding while generating replay or injection events inconsistent with the expected source, sequence, or freshness state [29, 30].
- **Resource exhaustion:** the relay generates abnormal communication activity that increases relay energy expenditure [29, 31].
- **Rogue relay:** an unauthenticated candidate attempts to participate. Its authentication state is zero and therefore fails the hard eligibility rule [30].

Training uses benign, black-hole, selective-forwarding, on–off, and threshold-aware episodes with domain randomisation over attack severity, compromised fraction, attack onset, on–off periods, and initial relay energy. Replay/injection, resource exhaustion, and rogue-relay participation are withheld entirely from learned-policy training and are introduced only during final evaluation. The nominal final attack begins after 20% of the episode, and 50% of candidate relays are compromised where the scenario permits. This separation is used to test whether the eligibility structure transfers to attack mechanisms not encountered by PPO during training.

5.4 Baselines, Ablations, and Protocol

The comparison set includes direct transmission, adaptive best-channel relay selection [12], a lifetime-oriented relay selector [13], AHP–TOPSIS [6], a predictive channel-based selector inspired by Pan *et al.* [17], an ARAP-style multi-criteria selector [15], and a PPO-RSS-style learned baseline [18]. Where the published description is insufficient for exact reproduction, the manuscript labels the implementation explicitly as "-like" or "-approx" rather than presenting it as the authors' original code.

Two lightweight controllers use the proposed hard eligibility rule without the learned recurrent preference policy. *Secure best-channel* selects the best currently admissible channel. *Secure Thompson* maintains a Beta–Bernoulli posterior over forwarding success within the admissible set and uses posterior sampling with an energy term. These controllers test whether learning policy complexity is necessary once risky relays have been removed.

Three ablations isolate model components. *Reward-trust* removes the trust/anomaly hard mask and instead adds trust as a positive reward term. *Feedforward hard-mask* retains the proposed set encoder and hard eligibility but removes recurrence. *Flat recurrent hard-mask* retains recurrence and hard eligibility but replaces the permutation-equivariant set encoder with a flat network over the padded observation.

Trust parameters and learned checkpoints are selected from validation data only. PPO checkpoints are evaluated periodically on held-out validation episodes and frozen before final testing. Within each final seed, all methods receive the same channel segment, traffic context, initial-energy realisation, and attacker schedule, so comparisons are paired. No trust, PPO, or baseline parameter is modified after observing final-test performance.

5.5 Metrics and Statistical Analysis

Security is the primary evaluation axis. *Attack-resilient delivery* is the fraction of packets delivered from attack onset to the end of the episode, thereby avoiding dilution by the pre-attack period. Additional security metrics include malicious-relay selection rate, malicious-relay exclusion rate, residual packet loss caused by the adversary, direct-fallback rate, and detection latency. A compromised relay that is never excluded contributes the entire remaining post-onset horizon to the latency statistic rather than being dropped from the average.

False exclusion is evaluated separately under benign conditions. Two measures are distinguished: (i) the fraction of benign relay-epochs for which a legitimate relay is outside the admissible set, and (ii) the fraction of benign relays that are excluded at least once during an episode. Methods that do not implement trust/anomaly exclusion are reported as not applicable rather than being credited with a zero false-exclusion rate.

Communication metrics are packet-delivery ratio (PDR) throughput, mean and 95th-percentile delay, total communication energy, and energy per delivered packet. Time to first depletion and time to 25% node depletion are also recorded. If no depletion occurs within the one-hour episode, these lifetime values are right-censored at the episode horizon and are not interpreted as measured lifetime.

Every final configuration is evaluated over 30 paired seeds, with the seed treated as the statistical unit. Summary metrics are reported as arithmetic means with nonparametric 95% percentile-bootstrap confidence intervals, obtained by resampling the 30 seed-level observations with replacement using 10,000 bootstrap replicates. This procedure preserves the natural bounds of rate metrics and avoids invalid negative confidence limits for quantities such as false-exclusion rate. Pairwise comparisons use the Wilcoxon signed-rank test because each method experiences the same seed-specific environment. Family-wise multiplicity is controlled using Holm correction, and matched-pairs rank-biserial correlation is reported as an effect size.

6 Results and Discussion

6.1 Security Effect of Trust-Aware Relay Selection

Table 4 reports attack-resilient delivery across seven adversary models. The largest differences occur under forwarding-oriented attacks. Under black-hole, selective-forwarding, on–off, and threshold-aware attacks, several conventional and unconstrained methods degrade substantially, whereas the proposed controller maintains 97.52–97.86% delivery.

Figure 2 highlights these four attacks using four representative design stages: AHP–TOPSIS, PPO-RSS-like, Reward-trust, and the proposed hard-constrained controller. Relative to PPO-RSS-like, the proposed controller improves attack-resilient delivery by 44.61, 12.80, 7.95, and 7.85 percentage points, respectively. This indicates that learning without explicit security eligibility is insufficient when malicious relays remain selectable.

Reward-trust also performs strongly, achieving 97.25%, 97.74%, 97.20%, and 97.79% under the four forwarding-oriented attacks, respectively, compared with 97.61%, 97.74%, 97.52%, and 97.86% for the proposed controller. Paired Wilcoxon tests with Holm correction do not establish a significant advantage of the proposed controller over Reward-trust; for example, the black-hole comparison changes from $p = 0.040$ before correction to $p = 0.162$ after correction.

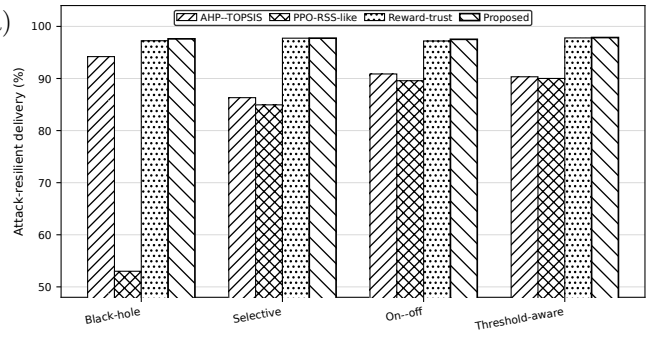


Fig. 2 Attack-resilient delivery under the four forwarding-oriented adversaries, averaged over 30 paired seeds. AHP–TOPSIS represents conventional multi-criteria selection, PPO-RSS-like represents learning without an explicit security constraint, Reward-trust incorporates trust through reward shaping, and Proposed enforces hard security eligibility before adaptive relay preference. The complete comparison is reported in Table 4.

Thus, hard eligibility should not be interpreted as necessarily improving mean delivery. Its main benefit is structural: Reward-trust can still select a relay whose trust penalty is outweighed by favourable communication attributes, whereas a relay excluded by the hard security rule is no longer selectable. The proposed formulation therefore makes security non-compensatory through the feasible action set.

Direct transmission is excluded from the relay-selection ranking because it does not traverse a potentially compromised forwarding relay. Its performance is instead evaluated separately across the assumed direct-link reliability range in Section 6.4.

6.2 Policy Architecture and Ablation

The previous results indicate that security admissibility is more influential than the complexity of the policy operating within the admissible set. To separate architectural effects, Figure 3 compares the feedforward hard-mask controller, the permutation-sensitive flat recurrent controller, and the proposed permutation-equivariant recurrent controller. The feedforward hard-mask controller performs comparably to, and in several conditions slightly better than, the proposed recurrent controller. Under on–off attack, it achieves 97.63% attack-resilient delivery compared with 97.52% for the proposed controller. The paired difference is statistically significant after Holm correction ($p_{\mathrm{Holm}} = 0.012$, rank-biserial effect size $= -0.63$). Consequently, the present experiments do not support a claim that recurrence is necessary for attack resilience.

This result is consistent with the role of the trust state. The trust estimator already integrates forwarding evidence across time, so part of the temporal structure

Table 4 Complete attack-resilient delivery (%) over 30 paired seeds for relay-selection methods. Attacks marked † were withheld from policy training. Direct transmission is not included in this ranking because the forwarding attacks target relay behaviour; the modelled direct link is evaluated separately as a function of p_{d} in Section 6.4. Figure 2 shows the representative comparison.

Method	Black-hole	Selective	On–off	Thresh.-aware	Replay†	Exhaust.†	Rogue†
Best-channel	54.40	85.89	90.42	90.95	99.43	99.36	97.82
Lifetime-max	49.31	83.91	89.20	89.47	98.89	98.88	98.11
AHP–TOPSIS	94.20	86.32	90.87	90.33	98.93	98.92	97.89
ARAP-like	83.88	87.23	91.11	90.58	98.40	98.38	97.06
Predictive-approx	52.59	85.29	90.10	90.58	99.34	99.33	98.46
PPO-RSS-like	53.00	84.94	89.57	90.01	98.58	98.43	96.11
Reward-trust	97.25	97.74	97.20	97.79	99.05	99.01	97.80
Secure best-channel	97.80	98.10	97.86	98.13	99.42	98.37	97.82
Secure Thompson	98.04	98.35	98.05	98.13	99.27	98.36	98.49
Feedforward hard-mask	97.73	97.96	97.63	97.84	98.87	98.21	98.06
Flat recurrent hard-mask	96.88	97.03	97.04	96.95	98.58	97.18	96.27
Proposed	97.61	97.74	97.52	97.86	98.63	98.10	98.03

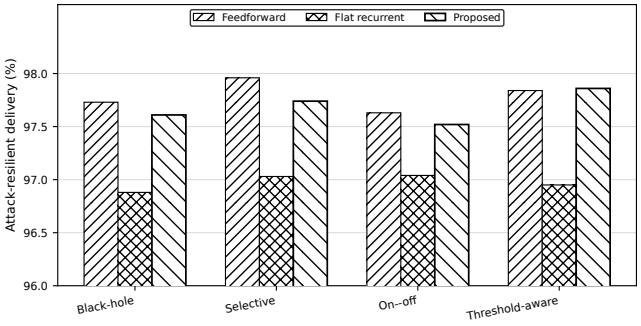


Fig. 3 Architecture ablation under the forwarding-oriented attacks. All three controllers employ the hard eligibility mechanism. Feedforward hard-mask removes recurrence while retaining the permutation-equivariant set representation, Flat recurrent retains recurrence but operates on the padded flat relay-state vector, and Proposed combines the set representation with recurrent state.

that a recurrent policy might otherwise need to learn is explicitly encoded in the state representation. Once unsafe relays have been removed, preference among the remaining candidates may therefore be close to a short-horizon decision problem. This interpretation is consistent with the placement of temporal modelling in related designs, where recurrent structure is applied to the trust estimator itself in order to capture how node reliability evolves [21], rather than to the policy that consumes the resulting state. Where temporal evidence is already accumulated in the trust variable, a recurrent preference policy has correspondingly less residual structure to exploit.

The set representation nevertheless appears useful. The Flat recurrent hard-mask controller, which processes a fixed padded relay ordering rather than using shared relay encoding and permutation-invariant pooling, falls below 97% under black-hole, threshold-aware, and rogue attacks. Its benign relay-epoch false-exclusion rate is also 5.99%, compared with 1.46% for the proposed controller. These results suggest that representing candidate relays as an exchangeable set contributes more consistently to robustness than adding recurrence alone.

The strongest constrained comparator is Secure Thompson. Averaged across the seven attacks, it achieves approximately 98.38% attack-resilient delivery, compared with approximately 97.93% for the proposed recurrent controller. Secure best-channel also achieves approximately 98.21%. We therefore do not claim that recurrent PPO provides the highest delivery ratio. Instead, these comparisons reinforce the principal finding: once the same hard eligibility mechanism is imposed, even lightweight preference controllers can provide strong attack resilience. The learned controller should therefore be viewed as an adaptive preference mechanism within the security constraint rather than as the primary source of the security gain. To isolate the effect of the preference controller after security admissibility has been enforced, Figure 4 compares three methods operating over the same hard-constrained relay set. Secure Thompson achieves slightly higher attack-resilient delivery than the proposed recurrent controller across most adversary models, while Secure best-channel is also competitive. The differences are small compared with the large gap between constrained and unconstrained relay-selection methods. This supports the conclusion that explicit security eligibility, rather than controller complexity, is the principal source of attack resilience.

6.3 Benign Operating Cost and False Exclusion

Table 5 evaluates normal operation in the absence of an adversary. The proposed controller obtains a packet-

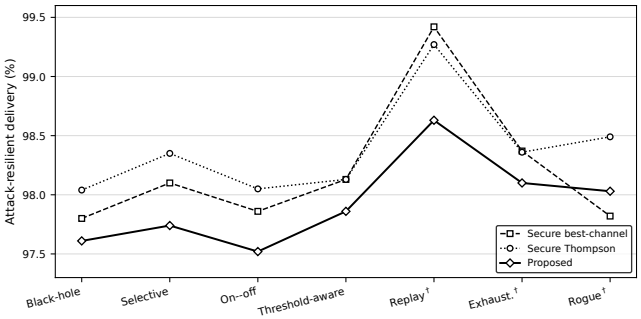


Fig. 4 Attack-resilient delivery of representative controllers operating under the same hard security-eligibility mechanism. Secure best-channel, Secure Thompson, and the proposed recurrent controller all restrict relay selection to the admissible set. Their relatively small performance differences indicate that the hard eligibility mechanism accounts for most of the attack resilience, while the preference controller has a secondary effect.

delivery ratio of $98.60 \pm 0.04\%$, mean delay of 19.86 ± 0.05 ms, and approximately 0.04 mJ per delivered packet. These values are similar to those of the other relay-assisted methods.

Two pairs exhibit identical benign communication summaries at the reported precision. PPO-RSS-like and Flat recurrent hard-mask produce the same PDR, delay, and energy values, while Best-channel and Secure best-channel likewise produce identical benign communication metrics despite a non-zero false-exclusion rate for the secure variant. For Secure best-channel, this is possible because the hard eligibility rule removes candidates before applying the same best-channel preference rule; a legitimate relay may therefore be falsely excluded without affecting communication performance when it would not have been the highest-quality candidate at that epoch. False exclusion thus measures the restriction imposed on the feasible relay set and does not necessarily imply an immediate packet-delivery or delay penalty. Under adversarial conditions, however, hard eligibility can alter the selected relay by removing compromised or otherwise inadmissible candidates.

Relay-assisted methods consume approximately 0.04 mJ per delivered packet and incur the additional delay associated with relay reception and a second transmission. The purpose of the proposed framework is therefore not to claim that two-hop relaying reduces communication energy. Residual energy is instead used as an eligibility and preference variable to avoid selecting relays approaching the survivability threshold and to distribute forwarding load among admissible candidates. The operating-cost comparison with modelled direct transmission is examined separately in Section 6.4.

No node reaches depletion during the 3600-s evaluation horizon. Time to first depletion is therefore right-censored at the episode boundary, and the current experiment does not provide evidence of improved network lifetime. The energy results are consequently interpreted as operating-cost characterisation rather than evidence of lifetime extension.

False exclusion provides a direct measure of the cost introduced by hard security eligibility. Under benign operation, the proposed controller excludes legitimate relays in an average of 1.46% of relay-epochs, compared with 1.01% for the feedforward hard-mask controller and 5.99% for the Flat recurrent controller. Secure Thompson produces no false exclusions in the reported benign runs. The corresponding 95% bootstrap confidence intervals are reported in Table 5. These results show that hard eligibility should be evaluated jointly with the evidence estimator used to determine whether a relay remains admissible.

6.4 Sensitivity to Direct-Link Reliability

The available channel traces cover candidate-relay links but not a matched sensor-to-coordinator direct link. Direct transmission is therefore modelled as a controlled reference with packet-success probability p_{d}, evaluated over $p_{\mathrm{d}} \in \{0.70, 0.80, 0.90, 0.95, 0.975, 0.985, 0.99\}$.

Figure 5 compares direct transmission with representative controllers enforcing hard security eligibility. Direct delivery increases with p_{d}, whereas the constrained relay controllers remain comparatively stable. Their curves intersect only in the high-reliability direct-link regime. Thus, under the present model, relay-assisted operation provides a delivery advantage when direct-link reliability falls below approximately 97%. Above this high-reliability regime, direct transmission becomes preferable. Because the study does not contain a matched measured direct link, this value should be interpreted as an operating boundary of the model rather than as a measured deployment threshold.

This result clarifies the intended operating regime of the framework. Relay-assisted communication is not claimed to be preferable when the direct body-area link is already highly reliable. Rather, security-constrained relay selection is relevant when direct-link reliability degrades, allowing relay assistance while preventing inadmissible relays from entering the feasible action set.

6.5 Held-Out Attacks and Relay-Set Transfer

Replay/injection, resource-exhaustion, and rogue-relay attacks are withheld from policy training and used only during final evaluation. The proposed controller achieves an average attack-resilient delivery of 97.68% on attacks represented during training and 98.25% on the three

Table 5 Benign-condition operating performance over 30 paired seeds. Values are arithmetic means with nonparametric 95% percentile-bootstrap confidence intervals shown in brackets, based on 10,000 bootstrap resamples of the seed-level observations. FE denotes the percentage of benign relay-epochs in which a legitimate relay is falsely excluded; "–" indicates not applicable. Methods below the horizontal rule enforce hard security eligibility.

Method	PDR (%)	Delay (ms)	Energy (mJ/pkt)	FE (%)
Best-channel	99.43 ± 0.03	19.06 ± 0.07	0.04	–
AHP–TOPSIS	98.98 ± 0.03	19.35 ± 0.05	0.04	–
PPO-RSS-like	98.72 ± 0.08	20.35 ± 0.08	0.04	–
Reward-trust	99.03 ± 0.03	19.17 ± 0.06	0.04	–
Secure best-channel	99.43 ± 0.03	19.06 ± 0.07	0.04	1.34 ± 0.24
Secure Thompson	99.30 ± 0.03	19.34 ± 0.08	0.04	0.00 ± 0.00
Feedforward hard-mask	98.85 ± 0.03	19.49 ± 0.04	0.04	1.01 ± 1.57
Flat recurrent hard-mask	98.72 ± 0.08	20.35 ± 0.08	0.04	5.99 ± 0.12
Proposed	98.60 ± 0.04	19.86 ± 0.05	0.04	1.46 ± 1.63

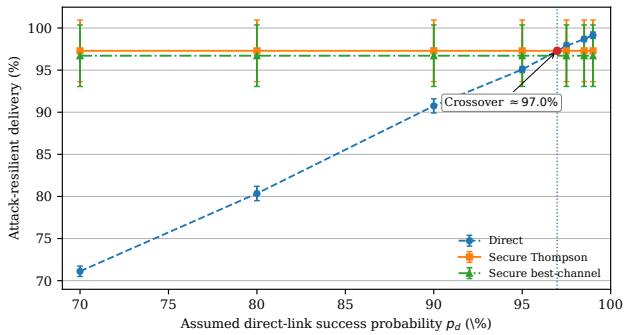


Fig. 5 Attack-resilient delivery versus the assumed direct-link packet-success probability p_{d}. Direct transmission is a modelled reference rather than a measured direct-link result. Points show means over 30 paired seeds; error bars denote 95% confidence intervals.

held-out attacks, corresponding to an absolute difference of 0.57 percentage points.

The higher mean on the held-out set should not be interpreted as evidence that unseen attacks are intrinsically easier to generalise to. Replay/injection, resource exhaustion, and rogue participation affect the delivery metric differently from black-hole and selective-forwarding attacks. The important observation is therefore the absence of substantial deterioration when the policy encounters attack classes excluded from training.

The mechanism supporting this behaviour is partly structural rather than purely learned. Replay/injection and resource-exhaustion events contribute to the anomaly state, while rogue participation is rejected through the authentication condition. These security signals enter the eligibility mechanism independently of the identity of the attack encountered during policy training. Secure Thompson and Secure best-channel exhibit similarly small seen/held-out differences, further supporting the interpretation that the hard eligibility mechanism, rather than attack-specific policy memorisation, is primarily responsible for robustness to these withheld conditions.

Finally, the permutation-equivariant architecture permits evaluation with a different number of candidate relays without retraining the policy. In the relay-count sweep, the proposed controller obtains mean attack-resilient delivery of 93.03% with $M = 2$ and 97.47% with $M = 4$. Secure Thompson achieves 93.03% and 97.71%, respectively, while Secure best-channel achieves 92.96% and 97.28%. Evaluation with $M = 6$ is not reported because the available channel trace provides only four relay-channel columns. The experiment therefore demonstrates transfer between two and four relay candidates, but does not establish scalability beyond the available measurement dimensionality.

6.6 Discussion

The combined results lead to three main observations. First, explicitly removing relays with unacceptable authentication, trust, anomaly, or residual energy states prevents severe degradation under forwarding attacks. The security gain is particularly clear relative to conventional and unconstrained learning-based relay-selection methods.

Second, using trust as a hard admissibility condition and using trust as a reward signal are not equivalent formulations. Reward shaping can produce similar mean delivery in the evaluated scenarios, but it does not provide the same non-compensatory decision rule: a relay failing the hard criterion is unselectable regardless of its communication utility.

Third, policy complexity is secondary once admissibility has been enforced. The recurrent PPO controller does not consistently outperform the feedforward or posterior-sampling alternatives. This finding shifts the practical design priority from increasingly complex preference policies toward accurate, channel-aware evidence

estimation and reliable security eligibility. For resource-constrained WBAN coordinators, a lightweight constrained controller may therefore be preferable unless a deployment exhibits temporal partial-observability that produces a measurable advantage from recurrent learning.

7 Conclusion

This paper presented a trust-constrained adaptive relay-selection framework for wireless body area networks that separates *relay eligibility* from *relay preference*. Authentication state, statistical behavioural trust, anomaly evidence, and residual energy define the admissible relay set, while communication and resource variables are used only to rank eligible candidates. Trust is updated relative to the delivery expected under the current channel condition, and direct transmission is used only when no admissible relay remains.

Across black-hole, selective-forwarding, on–off, and threshold-aware attacks, the proposed recurrent controller maintains 97.52–97.86% post-onset delivery and remains near 98% on three attack types withheld from PPO training. However, the ablations show that recurrent PPO does not consistently outperform feedforward or lightweight constrained controllers. The principal finding is therefore that hard security eligibility, rather than policy complexity, provides the main source of attack resilience. The permutation-equivariant set representation nevertheless supports stable selection across relay identities and different candidate-set sizes.

The results also show that secure relaying introduces an operating cost. Relayed transmission consumes more energy and incurs greater delay than the modelled direct reference, and no node depletes during the one-hour evaluation. Accordingly, this study does not claim energy saving or network-lifetime extension.

The evaluation remains limited by the use of four relay-channel traces from one prepared recording set, a modelled direct link, calibrated rather than measured packet-success probabilities, and simulation-based energy and attack observations. Physical-layer jamming, sensor compromise, and coordinator compromise are also outside the current scope. Future work should therefore evaluate the framework using independent body-channel recordings, measured direct links and packet acknowledgements, physical wearable testbeds, and training procedures that do not rely on privileged attack-loss information. The present results also motivate lightweight implementations that combine the same statistical eligibility mechanism with feedforward or posterior-sampling relay preference.

Declarations

Funding
This research was funded by FPT University.

Competing Interests
The author declares no competing interests.

Ethics Approval
No experiments involving human participants or animals were conducted by the author.

Consent to Participate
Not applicable. No new participants were recruited for this study.

Consent for Publication
Not applicable.

Data Availability
The external datasets used to construct the experimental inputs are cited in the manuscript. The experimental results are available from the corresponding author upon reasonable request, subject to the redistribution conditions of the original data sources.

Code Availability
The implementation code used in this study is publicly available at `https://github.com/samnx070878/Wireless_Networks_TCARS`.

Author Contributions
Sam Nguyen-Xuan conceived the study, developed the methodology and implementation, conducted the experiments and data analysis, interpreted the results, and prepared and revised the manuscript.

References

1. D. M. G. Preethichandra, Lasitha Piyathilaka, Umer Izhar, Rohan Samarasinghe, and Liyanage C. De Silva. Wireless body area networks and their applications—a review. *IEEE Access*, 11:9202–9220, 2023. doi: 10.1109/ACCESS.2023.3239008.
2. Hai Wang and Abraham O. Fapojuwo. A survey of enabling technologies of low power and long range machine-to-machine communications. *IEEE Communications Surveys & Tutorials*, 19(4):2621–2639, 2017. doi: 10.1109/COMST.2017.2721379.
3. Guangzhu Chen, Qingchun Meng, and Lei Zhang. Chain-type wireless sensor network node scheduling strategy. *Journal of Systems Engineering and Electronics*, 25(2):203–210, 2014. doi: 10.1109/JSEE.2014.00024.

4. Chun-Hao Liao, Guibing Zhu, Daiki Kuwabara, Makoto Suzuki, and Hiroyuki Morikawa. Multi-hop LoRa networks enabled by concurrent transmission. *IEEE Access*, 5:21430–21446, 2017. doi: 10.1109/ACCESS.2017.2755858.
5. Daniel Lundell, Anders Hedberg, Christian Nyberg, and Emma Fitzgerald. A routing protocol for LoRa mesh networks. In *Proceedings of the 2018 IEEE 19th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, pages 14–19, Chania, Greece, 2018. IEEE. doi: 10.1109/WoWMoM.2018.8449743.
6. Naveen Bilandi, Harsh Kumar Verma, and Renu Dhir. Energy-efficient relay node selection scheme for sustainable wireless body area networks. *Sustainable Computing: Informatics and Systems*, 30: 100516, 2021. doi: 10.1016/j.suscom.2021.100516.
7. Samiha Ayed, Lamia Chaari, and Amina Fares. A survey on trust management for WBAN: Investigations and future directions. *Sensors*, 20(21):6041, 2020. doi: 10.3390/s20216041.
8. A. Roshini and K. V. D. Kiran. Hierarchical energy efficient secure routing protocol for optimal route selection in wireless body area networks. *International Journal of Intelligent Networks*, 4:19–28, 2023. doi: 10.1016/j.ijin.2022.11.006.
9. Ruby Dass, Manikandan Narayanan, Gayathri Ananthakrishnan, Tamilarasi Kathirvel Murugan, Musiri Kailasanathan Nallakaruppan, Siva Rama Krishnan Somayaji, Kannan Arputharaj, Surbhi Bhatia Khan, and Ahlam Almusharraf. A cluster-based energy-efficient secure optimal path-routing protocol for wireless body-area sensor networks. *Sensors*, 23(14):6274, 2023. doi: 10.3390/s23146274.
10. Tingxuan Fu, Sijia Hao, Qiming Chen, Zihan Yan, Huawei Liu, and Amin Rezaeipanah. An energy-aware secure routing scheme in Internet of Things networks via two-way trust evaluation. *Pervasive and Mobile Computing*, 105:101995, 2024. doi: 10.1016/j.pmcj.2024.101995.
11. Hongmei Fei, Dingyi Jia, Baitao Zhang, Chaoqun Li, Yao Zhang, Tao Luo, and Jie Zhou. A novel energy efficient QoS secure routing algorithm for WSNs. *Scientific Reports*, 14(1):25969, 2024. doi: 10.1038/s41598-024-77686-y.
12. Xuan Sam Nguyen, Daehee Kim, and Sunshin An. An adaptive relay transmission scheme for reliable data forwarding in wireless body area networks. *IEICE Transactions on Information and Systems*, E99-D(2):415–423, 2016. doi: 10.1587/transinf.2015EDP7286.
13. Yu Zhang, Bing Zhang, and Shi Zhang. A lifetime maximization relay selection scheme in wireless body area networks. *Sensors*, 17(6):1267, 2017. doi: 10.3390/s17061267.
14. Yu Zhang, Bing Zhang, and Shi Zhang. An adaptive energy-aware relay mechanism for IEEE 802.15.6 wireless body area networks. *Wireless Personal Communications*, 115(3):2363–2389, 2020. doi: 10.1007/s11277-020-07686-4.
15. Surender Singh and Naveen Bilandi. Adaptive relay-assisted WBAN protocol: Enhancing energy efficiency and QoS through advanced multi-criteria decision-making. *Computer Modeling in Engineering & Sciences*, 144(1):489–509, 2025. doi: 10.32604/cmes.2025.065101.
16. Razieh Mohammadi and Zahra Shirmohammadi. RLS^2: An energy efficient reinforcement learning-based sleep scheduling for energy harvesting WBANs. *Computer Networks*, 229:109781, 2023. doi: 10.1016/j.comnet.2023.109781.
17. Yunlin Pan, Yaxuan Ren, and Jiasong Mu. Deep incremental learning-driven human body channel prediction with adaptive relay selection for enhanced WBAN performance. *IEEE Access*, 12: 186145–186152, 2024. doi: 10.1109/ACCESS.2024.3502797.
18. Dongmei Zhou, Baowan Yan, Cuiran Li, Aihuan Wang, and Haixia Wei. Relay selection scheme based on deep reinforcement learning in wireless sensor networks. *Physical Communication*, 54: 101799, 2022. doi: 10.1016/j.phycom.2022.101799.
19. Ramji Gupta, Shashi Kant Gupta, Swati B. Singh, Ramesh Kumar, and R. Muthukkumar. Adaptive deep reinforcement learning-based secure routing for wireless sensor networks. *Scientific Reports*, 16(1):XXXXX, 2026. doi: 10.1038/s41598-026-62519-x.
20. Muhammad Yeasir Arafat, Sungbum Pan, and EunSang Bak. QQAR: A Q-learning-based QoS-aware routing for IoMT-enabled wireless body area networks for smart healthcare. *Internet of Things*, 26: 101151, 2024. doi: 10.1016/j.iot.2024.101151.
21. V. Irine Shyja, G. Ranganathan, P. Chandrakanth, G. Sindhu Priya, and Dawit Tafesse Gebreyohannes. Optimal cluster-based energy efficient routing scheme for QoS aware IoT-enabled wireless body area network. *Scientific Reports*, 16(1): 6689, 2026. doi: 10.1038/s41598-026-37344-x.
22. Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. *The International FLAIRS Conference Proceedings*, 35, 2022. doi: 10.32473/flairs.v35i.130584.

23. Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42): 1437–1480, 2015.
24. David B. Smith, Leif W. Hanlen, David Rodda, Ben Gilbert, Jie Dong, and Vasanta Chaganti. Body area network radio channel measurement set, 2017. URL `https://data.csiro.au/collection/csiro:18350`. Dataset, version 1. Licensed under CC BY-SA 4.0.
25. Leif W. Hanlen, Vasanta Chaganti, Ben Gilbert, David Rodda, Tharaka Lamahewa, and David B. Smith. Open-source testbed for body area networks: 200 sample/sec, 12 hrs continuous measurement. In *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 66–71, Istanbul, Turkey, 2010. IEEE.
26. Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of the 16th International Symposium on Wearable Computers (ISWC)*, pages 108–109, Newcastle, UK, 2012. IEEE. doi: 10.1109/ISWC.2012.13.
27. Bart Braem, Benoît Latré, Ingrid Moerman, Chris Blondia, Elisabeth Reusens, Wout Joseph, Luc Martens, and Piet Demeester. The need for cooperation and relaying in short-range high path loss sensor networks. In *2007 International Conference on Sensor Technologies and Applications (SENSORCOMM 2007)*, pages 566–571, Valencia, Spain, 2007. IEEE. doi: 10.1109/SENSORCOMM.2007.95.
28. Elisabeth Reusens, Wout Joseph, Benoît Latré, Bart Braem, Günter Vermeeren, Emmeric Tanghe, Luc Martens, Ingrid Moerman, and Chris Blondia. Characterization of on-body communication channel and energy efficient topology design for wireless body area networks. *IEEE Transactions on Information Technology in Biomedicine*, 13(6):933–945, 2009. doi: 10.1109/TITB.2009.2033054.
29. Muhammad Shadi Hajar, M. Omar Al-Kadri, and Harsha Kumara Kalutarage. A survey on wireless body area networks: Architecture, security challenges and research opportunities. *Computers & Security*, 104:102211, 2021. doi: 10.1016/j.cose.2021.102211.
30. Bhawna Narwal and Amar Kumar Mohapatra. A survey on security and authentication in wireless body area networks. *Journal of Systems Architecture*, 113:101883, 2021. doi: 10.1016/j.sysarc.2020.101883.
31. Mohammad Yaghoubi, Khandakar Ahmed, and Yuan Miao. Wireless body area network (wban): A survey on architecture, technologies, energy consumption, and security challenges. *Journal of Sensor and Actuator Networks*, 11(4):67, 2022. doi: 10.3390/jsan11040067.



Sam Nguyen-Xuan received the B.Eng. degree in Electronics and Telecommunications from the Posts and Telecommunications Institute of Technology, Hanoi, Vietnam, in 2002, the M.Sc. degree in Information and Communications from Andong National University, Republic of Korea, in 2009, and the Ph.D. degree in Computer Engineering from Korea University, Republic of Korea, in 2016. He is currently the Dean of the Department of Computer Science at Swinburne Vietnam, Ho Chi Minh City, FPT University. His research interests include natural language processing, machine learning for the Internet of Things, wireless and wearable networks, and cybersecurity.