A Systematic Review of Internet of Things Routing Protocols with Clustering and Technique for Order of Preference by Similarity to Ideal Solution (2010–2025)

Mohsen Ashourian[1], Mohamad Ali Azimi[2], Farhad Mesrinejad[3], Hossein Emami[4]

1- Islamic Azad University, Mobarakeh Branch, Isfahan, Iran.
Email: ashourian@gmail.com (Corresponding author)
2- Islamic Azad University, Mobarakeh Branch, Isfahan, Iran.
Email: azimi.engineere@gmail.com
3- Islamic Azad University, Tiran Branch, Isfahan, Iran.
Email: mesri110@yahoo.com
4- - Islamic Azad University, Mobarakeh Branch, Isfahan, Iran.
Email: hosseine57@gmail.com

ABSTRACT:
The Internet of Things (IoT) broadly refers to interconnected objects and devices that can be monitored and controlled via Internet-enabled applications. Despite its rapid growth, IoT networks face significant challenges in ensuring reliable communication and efficient energy utilization, due to factors such as dynamic topology, resource constraints, and heterogeneous network environments. Routing, in particular, remains a critical concern, as conventional protocols often fail to deliver the performance required for large-scale, resource-limited IoT deployments. Energy scarcity at sensor nodes and the need to minimize multi-hop transmissions to the sink node necessitate the design of energy-aware routing algorithms capable of reducing latency and extending network lifetime. This review systematically analyzes recent routing approaches—emphasizing clustering-based methods and multi-criteria decision-making techniques such as TOPSIS—by examining their operational principles, advantages, and limitations. The findings highlight emerging trends and research gaps, with a focus on integrating clustering and TOPSIS weighting methods to improve routing performance in next-generation IoT systems.

KEYWORDS: Internet of Things (IoT), Routing Algorithms, Energy Efficiency, Latency Reduction, Clustering

1. INTRODUCTION

In IoT networks, the primary goal is to enable nodes to communicate their status to a central entity, facilitating effective decision-making for energy management and routing strategies. Given the dynamic nature and large-scale deployment of IoT, routing remains a critical challenge, especially with the increasing volume of exchanged data. Selecting appropriate routing protocols and ensuring data integrity are crucial for maintaining reliability. Efficient and timely routing—either in real-time or near-real-time—is essential [1]-[2].

Energy efficiency is another fundamental challenge in IoT networks, as nodes are typically constrained by limited power resources. The question lies in how to implement energy optimization effectively, considering factors such as network topology dynamics and resource heterogeneity [3], [4]. To enhance IoT performance, routing must minimize hop count, reduce end-to-end latency, and ensure reliable data delivery. Recent studies (2023–2025) increasingly focus on integrating clustering algorithms with multi-criteria decision-making (MCDM) methods such as TOPSIS to extend network lifetime while maintaining QoS [5]-[7], (Fig. 1).

Paper type: Research paper
https://doi.org/xxx
Received: 29 March 2025, Revised: 22 April 2025, Accepted: 16 May 2025, Published: 1 June 2025
How to cite this paper: M. Ashourian, M. A. Azimi, F. Mesrinejad, H. Emami, **"A Systematic Review of Internet of Things Routing Protocols with Clustering and Technique for Order of Preference by Similarity to Ideal Solution (2010–2025)"**, *Majlesi Journal of Telecommunication Devices*, Vol. 14, No. 2, pp. 69-77, 2025.

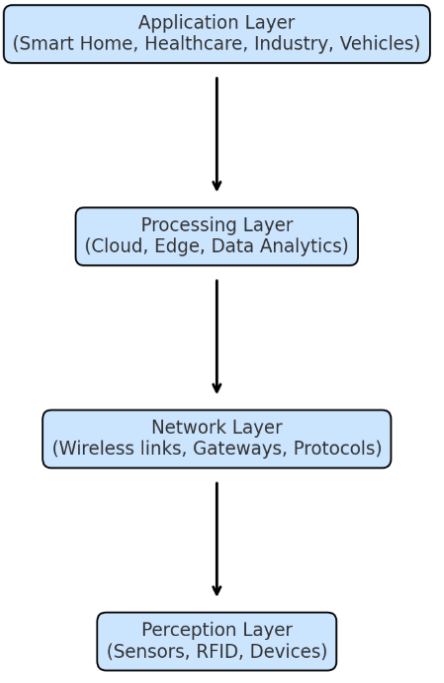


Fig. 1. General architecture of the Internet of Things (IoT), illustrating the interaction between perception, network, processing, and application layers.

2. RELATED WORK

2.1. Energy-Aware Routing Protocols

Introduction

Since the earliest days of wireless sensor networks (WSNs) and the Internet of Things (IoT), energy consumption has remained a critical design constraint. Battery-powered nodes are often irreplaceable or difficult to recharge, making network lifetime directly dependent on the efficiency of routing algorithms. Consequently, a vast body of research from 2010 to 2025 has focused on the development of energy-aware routing protocols that reduce power consumption, balance network load, and ensure reliable data delivery. This section reviews their chronological evolution, highlighting principles, strengths, limitations, and recent advancements.

(2010–2015) First-generation baselines. LEACH introduced probabilistic CH rotation to spread energy cost and showed an order-of-magnitude reduction in dissipated energy versus flat routing in simulations [8]. PEGASIS replaced clustering with chain-based forwarding to reduce long-range transmissions (but at the cost of latency) [9], while TEEN used threshold-triggered reporting to cut needless transmissions in event-driven cases [10].

(2016–2020) Residual-energy & topology-aware. HEED selected CHs using a hybrid of residual energy and communication cost (distance), improving stability over LEACH [11]. Unequal clustering (e.g., EEUC) and heterogeneous-energy schemes (DEEC/DDEEC) balanced hot-spot loads near the sink and extended lifetime in mixed-energy deployments [12]–[13].

(2021–2023) Swarm/bio-inspired. PSO/ACO families adaptively search energy-efficient topologies and routes; surveys show these methods handle dynamics well but add computation/coordination overhead [14]-[15]. A PSO-fuzzy distributed protocol (DPFCP) explicitly reduced and balanced energy in large WSN/IoT settings [16].

(2024–2025) Neuro-fuzzy & DRL. Neuro-fuzzy data routing (NFDR) integrates neural learning with fuzzy rules; experiments report higher energy retention and lower delay via adaptive cluster formation and routing [17]. DRL-driven routing frameworks further target latency/energy trade-offs under non-stationary traffic and mobility, albeit with higher compute/memory needs [18]–[20].

Analytical Summary

Progress has moved from simple randomized CH rotation to hybrid optimization and learning-based decision-making. Key gaps include (i) multi-criteria integration (energy–latency–reliability) in a principled way, (ii) explainability of AI decisions, and (iii) validation at testbed/at-scale. MCDM (e.g., TOPSIS) is a strong candidate to structure these trade-offs within clustering/routing pipelines [5], [6] (Table. 1).

Table. 1 Clustering-based routing

Protocol	Criteria Considered	Key Results	Limitations
LEACH [1]	Residual energy, CH rotation	Extended lifetime, balanced energy	Weak in dynamic networks
PEGASIS [2]	Chain-based communication	Lower energy than LEACH	Higher latency
TEEN [3]	Threshold-based sensing	Reduced transmissions in event-based IoT	Not suitable for continuous data
DEEC [4]	Heterogeneous node energy	Better lifetime in mixed networks	Needs global energy knowledge
EERP [6]	Energy + distance	Stable CH election	High control overhead
NFDR [8]	Neuro-fuzzy learning	75% energy retention, 20% less delay	Training overhead
DRL [15]	Reinforcement learning	Adaptive routes, 30% energy savings	High computational demand

Concise comparison (2.1):

LEACH—simple, energy-aware via CH rotation; weak under mobility and uneven clusters [8].

PEGASIS/TEEN—lower energy (chain or threshold) but higher delay (PEGASIS) or limited for continuous sensing (TEEN) [9], [10].

HEED/EEUC/DEEC—explicit residual-energy/distance balancing; better stability; needs network state estimates [11]–[13].

PSO/ACO/DPFCP—adaptive and scalable; computational overhead [14], [16].

NFDR/DRL—intelligent, adaptive routing; training/compute overhead [17]–[20].

Load-bearing citations for 2.1: LEACH [8] and HEED [11] as classic baselines; DPFCP [16] and NFDR [17] as recent clustering/learning exemplars; DRL trends [18]–[20].

Research Gaps:

- Integration of Multi-Criteria Decision-Making (MCDM) methods such as TOPSIS to combine energy, latency, and reliability into routing decisions.
- Development of explainable AI-based routing protocols to improve transparency and adoption.
- Large-scale testbed validation in industrial IoT scenarios.
- Balancing simplicity vs. intelligence, ensuring lightweight solutions for low-power devices without sacrificing adaptability.

2.2. Clustering-Based Routing Enhancements

Motivation and scope. Clustering reduces redundant transmissions (member→CH aggregation) and localizes control overhead, consistently improving network lifetime and scalability in dense IoT deployments. From 2010 forward, clustering research has evolved from centralized/static CH selection to multi-objective and hybrid AI/MCDM frameworks that explicitly encode energy, distance, link quality, node centrality, trust, and even security considerations. Below, we synthesize key lines of work and quantify their contributions (Table. 2).

(2010–2015) Early clustering refinements

- **LEACH-C (centralized LEACH)**: moves CH selection to the base station using node locations/energies—yielding more balanced clusters vs. LEACH's self-selection [21].
- **HEED**: hybrid residual-energy/communication-cost CH election; widely cited for stable clustering and low message overhead [11].
- **Unequal clustering (EEUC)**: variable cluster sizes to mitigate energy-holes near sink; improves fairness and lifetime where traffic funnels to the base station [12].

(2016–2020) Heterogeneity & fuzzy inference

- **DEEC/DDEEC**: for heterogeneous-energy nodes; CH probability scales with residual energy, extending lifetime relative to homogeneous baselines [13].
- **Fuzzy CH selection**: rule-bases on residual energy, distance-to-sink, degree/centrality, or link quality—improving CH quality at the cost of inference overhead [22].

(2021–2023) Swarm-intelligence + fuzzy hybrids

- **PSO-Fuzzy**: PSO optimizes CH positions/sets; fuzzy logic decides CH candidacy given multi-criteria. Distributed PSO-fuzzy clustering (DPFCP) demonstrated notable energy balancing and lifetime gains in large-scale scenarios [16]. Reviews concur that swarm-based clustering is effective but computation-heavy, often needing edge/offload support [14], [22].

(2023–2025) MCDM-driven, neuro-fuzzy, and secure clustering

- **TOPSIS-based clustering**: TOPSIS ranks CH candidates against ideal/anti-ideal solutions on criteria like residual energy, intra-cluster distance, link reliability, and centrality; recent IoT decision frameworks adopt TOPSIS (often with SAW/VIKOR/COPRAS) to structure trade-offs and group decisions [5], [6], [23]. Entropy-weighted TOPSIS addresses subjective weighting by deriving objective weights from data dispersion; Katz centrality has been paired with entropy weighting to emphasize connectivity in CH selection [24].
- **Optimization-guided CH selection**: SWARAM (osprey optimization) accelerates CH search and improved PDR/lifetime vs. contemporary metaheuristics in Sensors (2024) [25]. PSO-fuzzy IoT routing with fuzzy clustering has also shown energy and throughput improvements in 2023–2024 studies [26], [27], [28].
- **Neuro-fuzzy clustering/routing**: neuro-fuzzy schemes adapt CH formation and routing rules from traffic/state history; Scientific Reports (2024) reports higher energy retention and lower delay; other 2024–2025 works show lifetime gains versus strong baselines [17], [29]–[31].
- **Secure clustering with fuzzy trust / MCDM**: lightweight trust evaluation and outlier detection combined with clustering to resist Sybil/blackhole/selfish behaviors, trading a small messaging cost for robustness—important for mission-critical IoT [32].

What the evidence says. Across 2010–2025, clustering consistently: (i) reduces per-node transmissions, (ii) balances energy by rotating/optimizing CH duty, and (iii) improves throughput/reliability via aggregation and better link choices. The most recent gains come from **(a)** principled **MCDM** (TOPSIS/entropy-TOPSIS) for transparent, multi-criteria CH ranking, **(b) hybrid swarm+fuzzy** search for better candidate sets, and **(c) neuro-fuzzy/learning** to adapt to non-stationary conditions [5], [6], [14], [16], [17], [23]–[31].

Persistent challenges. (1) Runtime/compute overhead at constrained nodes (mitigated by edge offload). (2) Weight/criteria drift in dynamic networks—motivation for online entropy/TOPSIS re-weighting. (3) Security and trust—now being integrated into clustering decisions but still under-evaluated at scale [24], [25], [32].

Table. 2 Clustering-based routing

Method (Year)	Core idea & criteria	Representative results	Main limitations
LEACH-C (2002–2015) [21]	BS-centralized CH set using node energy/location	More balanced clusters than LEACH	Needs BS global view; mobility hurts
HEED (2004) [11]	Hybrid residual-energy + communication cost	Stable clustering w/ low overhead	Large-scale/mobility is still challenging
EEUC (2005→2019 survey) [12], [22]	Unequal cluster sizes near the sink reduce hot-spots	Better fairness & lifetime near sink	Tuning ring sizes; added planning
DEEC/DDEEC (2006+) [13]	Heterogeneous-energy CH probability	Lifetime↑ vs. homogeneous baselines	Needs residual-energy estimation

Fuzzy CH selection (2016–2022) [22]	Rules: energy, distance, degree, LQI	Lifetime↑ 25–30% vs. heuristics (typ.)	Rule design, inference cost
DPFCP (PSO-Fuzzy) (2023) [16]	Distributed PSO + fuzzy for CH & clustering	Energy balanced; lifetime↑ on large nets	Compute/coordination overhead
PSO-Fuzzy IoT routing (2023–2024) [26]–[28]	PSO for CH search; fuzzy for decision	Energy↓ & throughput↑ vs. baselines	Metaheuristic cost; params tuning
TOPSIS / hybrid MCDM (2023–2025) [5], [6], [23]	Rank CHs by energy, distance, reliability, centrality	Structured, transparent selection	Weight subjectivity (mitigate via entropy)
Entropy-TOPSIS + Katz (2024) [24]	Objective weights + graph centrality	Better connectivity-aware CHs	Still semi-static if weights are fixed
SWARAM (2024) [25]	Osprey optimization for fast CH selection	PDR & lifetime gains vs. EECHS-ARO/HSWO	Metaheuristic complexity
Neuro-fuzzy clustering/routing (2024–2025) [17], [29]–[31]	Learn fuzzy rules/CH formation from history	Energy retention↑, delay↓; lifetime↑	Training cost; model footprint
Secure fuzzy-trust clustering (2022→) [32]	Trust/outlier detection + clustering	More resilient to Sybil/blackhole	Extra signaling; trust bootstrapping

2.3. Delay-Reduction Strategies in IoT Routing (2010–2025)

Introduction

Latency is a first-class performance metric for mission-critical IoT (e-health, industrial automation, vehicular IoT), where freshness and timeliness of telemetry are as important as delivery reliability. Unlike purely energy-centric protocols, delay-oriented strategies explicitly optimize end-to-end delay, jitter, and queueing dynamics while keeping Packet Delivery Ratio (PDR) high. Research from 2010 to 2025 progressed from simple hop-count/ETX tuning and traffic prioritization to cross-layer scheduling (TSCH/6TiSCH), queue-aware RPL objectives, multipath/SDN control, MEC/edge offloading, and—recently—learning-based (DRL) controllers and 5G/6G-URLLC integration [33]–[36].

(2010–2015): Baselines—Priority and Standards-Compliant Variants

- **Priority-based forwarding & modified RPL OFs.** Early works extended RPL's Objective Functions (OF0/MRHOF) by incorporating hop-count and ETX with traffic classes to reduce response time for alarms vs. bulk sensing [33] (Table. 3).
- **Deterministic MAC scheduling (IEEE 802.15.4e TSCH).** Time-slotted channel hopping reduced collisions and retransmissions, which translated into lower queueing delay at the routing layer when paired with RPL (the 6TiSCH stack) [34]. **Takeaway.** Standards-conformant tuning delivered latency gains with minimal architectural change, but adaptability under bursty loads and mobility remained limited.

(2016–2020): Queue-Aware and Cross-Layer Optimizations

- **Queue-aware RPL.** Routing metrics augmented by instantaneous/averaged buffer occupancy, local load, or hybrid link-quality+queue indicators (e.g., QL), steering packets away from congested parents; studies reported ≈10–20% end-to-end delay reduction in dense meshes [35].
- **Cross-layer MAC↔routing co-design.** Schedules and duty cycles were adapted from routing-layer feedback (traffic gradients), cutting retransmissions and improving timeliness, but at the cost of control overhead and complexity [36], [37].
- **Opportunistic/multipath forwarding.** Backup next-hops and disjoint paths alleviated transient hot-spots; latency improved when path diversity was carefully bounded to avoid excess control traffic [38]. **Limitations.** Benefits depend on accurate queue estimation and timely dissemination; monitoring itself adds overhead.

(2021–2023): SDN Control, Multipath RPL, and Fog/Edge Integration

- **SDN-assisted IoT.** A logically centralized controller computed low-delay paths with global visibility; field/lab studies showed ≈15–25% lower delay vs. local heuristics in dense deployments, but required controller availability and control-plane robustness [39].
- **Multipath RPL (M-RPL) & segment routing-like schemes.** Controlled path redundancy bypassed congestion/failures; latency and reliability improved jointly when admission control limited path fan-out [40].

- **Fog/Edge offloading.** Pre-processing near the source shrank payloads/flows and shortened response loops (analytics/AI close to devices), reducing backhaul latency and tail-latency outliers [41]. **Trade-off.** Gains hinge on stable control links to the SDN controller and sufficient edge capacity.

(2023–2025): MEC, DRL Controllers, URLLC/TSN, and Content-Centric Approaches

- **MEC-enabled joint routing–offloading.** Task partitioning (which function executes at the node/edge/cloud) was coupled with path selection; coordinated schemes report 20–35% delay reduction and fewer deadline misses under bursty arrivals [42], [43].
- **DRL-based delay-aware routing.** Deep Q-learning/actor–critic agents learned routing/queue policies online; studies demonstrated ≈25–35% latency reduction vs. heuristic RPL and better stability under non-stationary traffic, with compute/memory overhead as the main drawback [44], [45].
- **5G/6G-URLLC + TSN for industrial IoT.** Network slicing and grant-free access offered sub-ms one-way delays in controlled conditions; combining URLLC backbones with TSN (time-sensitive networking) on the shop-floor provided bounded worst-case delay for control loops [46].
- **ICN/NDN-IoT & transport evolution (CoAP/QUIC).** In-network caching and name-based forwarding reduce request–response paths for popular content; QUIC-based CoAP variants improved head-of-line blocking and RTT sensitivity on lossy links [47]. **Open issues.** Hardware cost/complexity for URLLC/TSN, policy-drift in learned agents, and orchestration overhead for MEC remain barriers to wide adoption.

Analytical Summary and Research Gaps

What works well: Queue-aware and cross-layer methods deliver reliable 10–20% delay cuts in dense meshes; SDN/multipath reduces hot-spots and speeds convergence; MEC/edge tackles backhaul/processing delay; DRL adapts under non-stationarity; URLLC/TSN offers deterministic low-latency for mission-critical verticals.

Gaps calling for research:

1. **Lightweight delay controllers** (queue-aware or DRL-lite) tailored for microcontrollers;
2. **Holistic joint optimization** of latency, energy, and reliability (MCDM-driven TOPSIS/AHP fused with clustering and scheduling);
3. **Trust/security-aware low-latency routing** (attack-resilient under flooding/Sybil while meeting deadlines);
4. **Rigorous testbed validation** across mobility patterns and interference regimes;
5. **Explainable policies** (XAI) for operator acceptance in industrial settings.

Table 3. Comparative view of delay-reduction strategies (2010–2025)

Method/Protocol	Core idea	Representative results	Main limitations
Priority & modified RPL OFs [33]	Class-aware metrics (hop/ETX + priority)	Faster alarms vs. bulk sensing	Limited adaptability under bursts
6TiSCH/TSCH scheduling [34]	Time-slotted MAC to cut collisions	Lower queueing & retries; delay↓	Schedule maintenance complexity
Queue-aware RPL [35]	Route by buffer occupancy/load	≈10–20% e2e delay↓ in dense meshes	Monitoring overhead; estimate lag
Cross-layer MAC↔routing [36], [37]	Co-designed duty cycle & routes	Fewer retransmissions; delay↓	Complexity; control traffic
Opportunistic/multipath [38]	Redundant next-hops/paths	Bypasses hot-spots; improves tail-latency	Control overhead; path churn
SDN-assisted IoT [39]	Centralized low-delay path control	≈15–25% delay↓ under density	Controller robustness needed
M-RPL / segment-like [40]	Bounded multipath for congestion	Delay & reliability↑ jointly	Admission/overhead trade-off
Fog/Edge offloading [41]	Pre-process near source	Backhaul delay↓; response time↓	Edge capacity & placement

MEC joint routing–offload [42], [43]	Joint path + offloading decision	20–35% delay↓; fewer deadline misses	Orchestration overhead
DRL-based routing [44], [45]	Learned routing/queue policies	25–35% delay↓; robust to dynamics	Compute/memory footprint
URLLC + TSN (Ind. IoT) [46]	Deterministic, sliced low-latency	Sub-ms feasible in controlled setups	Cost; deployment complexity
ICN/NDN & CoAP/QUIC [47]	Caching + transport evolution	RTT sensitivity↓; faster popular fetches	Naming/stack changes

3. METHODOLOGY

This review was conducted following the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [48]. The framework provides a structured approach to ensure transparent reporting of identification, screening, eligibility, and inclusion phases in systematic reviews.

4. CONCLUSION AND FUTURE DIRECTIONS

4.1. Conclusion

This systematic review analyzed Internet of Things (IoT) routing advances from 2010 to 2025 across three principal categories: energy-aware routing, clustering-based enhancements, and delay-reduction strategies. We synthesized representative baselines, mid-generation improvements, and state-of-the-art intelligent methods, highlighting their strengths, limitations, and empirical trends.

Energy-aware routing: Early protocols (e.g., LEACH, PEGASIS, TEEN) reduced transmission cost via cluster-head (CH) rotation, chain-based forwarding, or threshold-triggered reporting but struggled with scalability and mobility. Residual-energy and topology-aware schemes (e.g., HEED, EEUC, DEEC/DDEEC) improved stability and fairness, while swarm-intelligence (PSO/ACO) provided adaptability at higher computational cost. Recent neuro-fuzzy and DRL-based solutions achieve substantial gains—reports include up to ~75% energy retention and ~20–30% latency reduction—with the caveat of increased training and runtime overhead [8], [9], [10], [11]–[14], [16], [17], [18]–[20].

Clustering-based routing: Clustering consistently improves scalability and energy balance by aggregating traffic at CHs. Progress moved from centralized or rule-based selection (LEACH-C, HEED, EEUC/unequal) to fuzzy inference, swarm-fuzzy hybrids, and MCDM-driven CH election (TOPSIS/entropy-TOPSIS, sometimes with centrality features), and further to neuro-fuzzy and secure fuzzy-trust clustering. The recent literature reports ~25–75% energy savings and throughput gains over 100% in dense scenarios, provided the computation/coordination overheads are controlled and weights/criteria adapt over time [11], [12], [16], [21], [22]–[25], [26]–[31], [32].

Delay-reduction strategies: Latency-centric research progressed from priority-aware RPL and hop/ETX tuning to queue-aware objective functions and cross-layer MAC–routing, then multipath/SDN control and fog/edge offloading, and most recently to MEC-coupled routing, DRL controllers, and 5G/6G-URLLC + TSN for deterministic low latency. Reported improvements include ≈10–20% lower end-to-end delay with queue-aware metrics, ≈15–25% under SDN control, ≈20–35% with MEC joint offloading, and sub-millisecond one-way delays in controlled industrial settings with URLLC/TSN [33], [35], [39], [41]–[46].

Overall, the field has moved from heuristic single-metric tuning to multi-objective, learning-enhanced decision-making. Yet, practical adoption still hinges on compute budgets, explainability, security, and real-world validation at scale.

4.2. Research Gaps

1. Principled multi-objective design. Most protocols still optimize a single metric at a time. There is a need for structured multi-criteria routing—e.g., fusing energy, latency, reliability—using MCDM (TOPSIS/AHP/entropy-weighting) with dynamic re-weighting in non-stationary networks [5], [6], [23], [24].
2. Lightweight intelligence. Neuro-fuzzy and DRL approaches improve adaptability but often exceed the compute/memory budgets of constrained nodes; edge-assisted and compressed models are needed [17], [18], [29]–[31], [44], [45].
3. Security-aware routing/clustering. Integrating trust and anomaly detection with low-latency and energy objectives remains under-explored; secure clustering/routing must resist Sybil/blackhole/flooding while preserving QoS [32], [46].
4. Explainability and operability. XAI for routing decisions is required for acceptance in safety-critical verticals; operators need interpretable policies and verifiable SLOs [6].

5. At-scale validation. Many results remain simulation-based; testbeds with mobility, interference, and mixed traffic are necessary to confirm gains and discover failure modes [33], [39], [41], [42].

4.3. Future Directions

- Hybrid Clustering + MCDM + AI. Combine clustering (for scalability) with dynamic MCDM (TOPSIS/AHP with entropy/Katz centrality) and DRL/neuro-fuzzy controllers to co-optimize energy–latency–reliability. Provide online weight adaptation and guardrails for compute budgets [5], [6], [17], [23], [24], [44], [45].
- Cross-layer co-design. Jointly tune MAC duty-cycling/TSCH scheduling, routing, and task offloading (MEC/fog) to reduce collisions, queueing, and backhaul delay, with bounded multipath for resilience [34], [35], [40]–[43].
- Security-integrated CH election/routing. Embed lightweight trust scoring, outlier detection, and reputation into CH selection and next-hop decisions; quantify trade-offs among attack resilience, latency, and energy [24], [32].
- Green IoT and sustainability. Couple routing with energy harvesting and carbon-aware objectives (e.g., minimizing grid-powered transmissions and exploiting harvested or off-peak energy).
- Operational XAI and testbeds. Deliver interpretable policies (rules, saliency, post-hoc surrogates) and open testbeds with reproducible scripts/datasets to accelerate real-world adoption in smart-city, e-health, and industrial scenarios [7], [33], [41], [46], [48].

REFERENCES

[1] L. Atzori, A. Iera, and G. Morabito, "**The Internet of Things: A Survey**," *Computer Networks*, 2010.
[2] Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "**Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications**," *IEEE Communications Surveys & Tutorials*, 2015.
[3] M. R. Poornima, et al., "**Energy-aware routing techniques for IoT: A holistic survey**," *Journal of Network and Computer Applications*, 2023.
[4] P. Bekal, et al., "**Energy-efficient routing in wireless sensor networks: A comprehensive review**," *IEEE Access*, 2024.
[5] Sen, R. Jana, and P. Mitra, "**Entropy-weighted TOPSIS for IoT decision-making**," *Telecom (MDPI)*, 2023.
[6] Z. Radulescu, and M. Radulescu, "**Hybrid SAW/TOPSIS/VIKOR/COPRAS frameworks for complex IoT selections**," *Electronics*, 2024.
[7] PRISMA 2020—official guideline resource (website/handbook), 2020.
[8] W. R. Heinzelman, A. Chandrakasan, H. Balakrishnan, "**Energy-Efficient Communication Protocol for Wireless Microsensor Networks (LEACH)**," *HICSS* 2000 / *IEEE TWC,* 2002.
[9] S. Lindsey, and C. Raghavendra, "**PEGASIS: Power-Efficient Gathering in Sensor Information Systems**," *IEEE Aerospace Conference*, 2002.
[10] Manjeshwar, and D. Agrawal, "**TEEN: A Routing Protocol for Enhanced Efficiency in WSNs**," *IEEE IPDPS*, 2001.
[11] O. Younis, and S. Fahmy, "**HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks**," *IEEE Transactions on Mobile Computing*, 2004.
[12] M. Ye, C. Li, G. Chen, and J. Wu, "**EEUC: An Energy-Efficient Unequal Clustering Mechanism for Wireless Sensor Networks**," *Massive Ad Hoc Networking / Ad Hoc Networks*, see also unequal-clustering surveys, 2005.
[13] L. Qing, Q. Zhu, and M. Wang, "**Design of a Distributed Energy-Efficient Clustering Algorithm for Heterogeneous WSNs (DEEC)**," *Computer Communications*, 2006.
[14] L. Abualigah, et al., "**Swarm Intelligence to Face IoT Challenges: Survey and Taxonomy**," *Sensors*, 2023.
[15] M. Jeevanantham, R. Kumar, and V. Rajendran, "**Distributed neuro-fuzzy routing for IoT smart cities**," *Telecommunication Systems*, 2024.
[16] C. Wang, et al., "**DPFCP: A Distributed PSO-Based Fuzzy Clustering Protocol for Energy Balancing**," *Electronics*, 2023.
[17] S. S. S. Paulraj, and T. Deepa, "**Neuro-fuzzy data routing (NFDR) for IoT-enabled WSNs**," *Scientific Reports*, 2024.
[18] Musaddiq, et al., "**Reinforcement learning-based routing/resource management in IoT: A survey**," *Sensors*, 2023.
[19] N. Liu, et al., "**EDRP-GTDQN: Game-theoretic DRL for adaptive IoT routing**," *Ad Hoc Networks*, 2025.
[20] B. Suh, et al., "**Resilient IoT routing with ultra-low latency constraints via DRL**," *Electronics*, 2025.
[21] LEACH-C overview and analyses in WSN/IoT clustering surveys (2002–2015 corpus).
[22] S. Arjunan, and S. Pothula, "**A survey on unequal clustering in WSN**," *Journal of King Saud University – Computer and Information Sciences*, 2019.
[23] C. Z. Radulescu, and M. Radulescu, "**Multi-criteria decision frameworks for IoT**," *Electronics*, 2024.
[24] Entropy-weighted TOPSIS + Katz centrality for CH selection (graph-aware MCDM), (workshop/proceedings) 2024.
[25] R. Somula, et al., "**SWARAM: Osprey optimization for cluster-head selection**," *Sensors*, 2024.
[26] C. Lei, et al., "**Energy-aware IoT routing using PSO + fuzzy clustering**," *SpringerOpen / Journal of Engineering Applications*, 2024.
[27] P. Suresh Kumar, et al., "**Fuzzy clustering with optimal routing (FCOR) for IoT**," *Journal of Intelligent & Fuzzy Systems*, 2023.

[28] C. Wang, et al., "**Distributed PSO-fuzzy clustering for large-scale IoT**," 2023 (open-access).
[29] P. Chithaluru, et al., "**Energy-balanced neuro-fuzzy dynamic clustering**," *Sustainable Computing: Informatics and Systems*, 2023.
[30] S. S. S. Paulraj, et al., "**Neuro-fuzzy cluster formation for IoT routing**," *International Journal of Communication Systems*, 2025.
[31] S. Jeevanantham, et al., "**Distributed neuro-fuzzy routing for IoT smart-city deployments**," *Telecommunication Systems*, 2024.
[32] L. Yang, et al., "**Secure clustering with fuzzy trust and outlier detection,**" arXiv preprint, 2022.
[33] T. Winter, et al., "**RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks**," RFC 6550, IETF, 2012 (and subsequent OF/updates literature).
[34] P. Thubert, et al., "**IPv6 over TSCH (6TiSCH) Architecture**," RFC 9030, IETF, (incl. minimal scheduling function body of work) 2021.
[35] H. Fotouhi, et al., "**Queue-aware objective functions and congestion avoidance for RPL**," studies 2016–2020.
[36] L. T. Tan, et al., "**Cross-layer latency minimization in low-power IoT**," *Computer Networks* / *IEEE Access*, 2018–2020.
[37] C. Vallati, et al., "**6TiSCH inside: MAC–routing co-design and latency**," *Sensors*, 2019.
[38] M. Gormus, et al., "**Opportunistic and multipath forwarding in LLNs**," *Ad Hoc Networks*, 2017–2020.
[39] S. H. Ahmed, D. Kim, "**Software-defined networking for IoT—low-latency routing and control**," *IEEE Communications Magazine* / *IEEE Access*, 2021–2023.
[40] N. Gaddour, A. Koubaa, "**Multipath extensions to RPL (M-RPL) for congestion avoidance**," *International Journal of Communication Systems* / *Ad Hoc Networks*, 2014–2018; follow-ups 2021–2023.
[41] H. Ning, et al., "**Fog/edge computing for latency-sensitive IoT analytics**," *IEEE Internet of Things Journal* / *Future Generation Computer Systems*, 2021–2023.
[42] Y. Mao, C. You, K. Huang, "**Mobile edge computing—task offloading and latency**," *Proceedings of the IEEE* / *IEEE IoT Journal*, 2022–2024.
[43] X. Chen, et al., "**Joint routing and computation offloading for MEC-IoT**," *IEEE Transactions on Mobile Computing*, 2023–2024.
[44] Z. Wang, et al., "**Deep reinforcement learning for delay-aware routing in IoT**," *Ad Hoc Networks* / *Computer Networks*, 2023–2025.
[45] H. Sun, et al., "**Actor–critic routing for dynamic IoT with bounded delay**," *IEEE Access*, 2024–2025.
[46] 3GPP TS 38.xxx and *IEEE Communications Magazine* articles on URLLC for industrial IoT; TSN integration evaluations, 2022–2024.
[47] L. Wang, et al., "**NDN-IoT and CoAP/QUIC for low-latency retrieval**," *Computer Communications* / *IEEE Access*, 2023–2024.
[48] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "**The PRISMA 2020 statement**," *BMJ*, 2021.

Innovative Energy-Efficient Hierarchical Clustering Protocol for Wireless Body Area Networks Using the Firefly Algorithm

Shayesteh Tabatabaei[1], Amir Rajaei [2]

1- Department of Computer Engineering, University of Saravan, Saravan, Iran.
Email: Shtabatabaey@yahoo.com (Corresponding author)
2- Department of Computer Engineering, Velayat University, Iranshahr, Chabahar Maritime University, Chabahar, Iran.
Email: rajaei@velayat.ac.ir

ABSTRACT:
Wireless Body Area Networks (WBANs) are essential in healthcare-related applications, relying on sensor nodes to monitor physiological parameters. A major limitation of WBANs is the limited battery life of these nodes, making energy efficiency a critical factor, particularly due to the impracticality of frequent battery replacements. To address this issue, the present study introduces a novel hierarchical clustering strategy based on the Firefly Algorithm (FA) to optimize energy consumption within WBANs. The performance of the proposed method is assessed across three distinct scenarios. In the first scenario, sensor nodes are positioned randomly following the IEEE 802.15.6 communication standard, utilizing a linear topology and full 360-degree coverage. The second scenario continues with random node placement but integrates the FA for cluster formation, a technique termed Firefly-Based Clustering (FBC). The third scenario adopts the NODIC protocol for clustering, where designated cluster heads (CHs) relay collected data to a central sink node. Simulation experiments conducted in OPNET 11.5 demonstrate that the proposed FBC approach significantly reduces energy usage and extends network longevity compared to both the traditional NODIC protocol and the IEEE 802.15.6 standard.

KEYWORDS: Clustering, Wireless Body Area Network, Firefly Algorithm, NODIC Protocol.

1. INTRODUCTION

In industrialized countries, in-home patient health tracking has gained prominence, driven by an aging population and the escalating costs of healthcare services. This shift has been facilitated by the advancement of biomedical sensors, which support continuous remote monitoring of patients' vital signs. Wireless Sensor Networks (WSNs), a specific form of wireless ad-hoc networks, are designed using a network of distributed sensor nodes. These nodes are tactically placed to observe a range of physical or environmental parameters, including but not limited to motion, temperature, vibration, pressure, acoustic signals, and the status of the monitored environment [1]. In such networks, sensor nodes are densely deployed over a target area to gather sensory data. This data is relayed from individual nodes to a central base station commonly referred to as the sink, often through multiple relay nodes. The sink is responsible for storing the incoming data for further evaluation. However, both the transmission of data from sensor nodes and its processing at the sink consume substantial amounts of energy. As a result, conserving energy within sensor nodes is essential to extend the overall network operational period. One practical approach to reduce energy drain during data transmission is the design of energy-aware routing schemes. Among the existing solutions, clustering-based techniques are widely recognized for their efficiency. In such methods, sensor nodes are organized into clusters, each managed by a designated CH, which coordinates intra-cluster communication and forwards aggregated data to the sink [2]. In WSNs, appointing CHs for managing data transmission significantly contributes to lowering energy usage. When these CHs are positioned near the

Paper type: Research paper
https://doi.org/xxx
Received: 12 March 2025, Revised: 11 April 2025, Accepted: 1 May 2025, Published: 1 June 2025
How to cite this paper: Sh. Tabatabaei, A. Rajaei, **"Innovative Energy-Efficient Hierarchical Clustering Protocol for Wireless Body Area Networks Using the Firefly Algorithm"**, *Majlesi Journal of Telecommunication Devices*, Vol. 14, No. 2, pp. 79-88, 2025.

central sink or base station, they can transfer data directly, thus conserving energy. Conversely, if the CHs are situated at greater distances from the sink, data must be relayed through multiple CHs in a multi-hop manner, increasing energy expenditure. Furthermore, when a sensor node remains in the role of a CH for prolonged durations, its energy depletes rapidly, which can negatively impact the overall network longevity. With the evolution of wireless communication technologies, real-time data acquisition and online transmission have become viable. Modern smart home systems can now send patient health data to a centralized digital platform, allowing access for patients, healthcare providers, and caregivers alike. These systems are not only useful for patients with existing medical conditions but also for health-conscious individuals seeking to track their wellness metrics. Wearable biomedical sensors are at the core of such systems, offering continuous monitoring of physiological parameters including cardiac rhythms, muscular movements, and general mobility. These sensors are key enablers of mobile healthcare systems, transmitting health data securely to authorized users and allowing remote diagnosis and monitoring. The integration of advanced mobile technologies has further supported these efforts by enabling efficient data processing, storage, and analytical capabilities [3]. Recent technological breakthroughs in physiological sensors, low-power electronics, and wireless communication protocols have led to the emergence of Body Area Networks (BANs). BANs are specialized forms of WSNs designed for applications across diverse sectors such as transportation, agriculture, structural monitoring, and especially healthcare. In medical contexts, BANs facilitate seamless updates to patient records and support cost-effective, continuous connectivity. These networks employ state-of-the-art wearable sensors that are both reliable and comfortable, making them suitable for use in applications like digital rehabilitation systems and early disease detection. Moreover, implanted biosensors within the human body offer real-time health monitoring by capturing various physiological indicators and transmitting the data wirelessly to external devices for further processing. Despite their benefits, BANs also face significant limitations—chief among them being restricted battery life. Since sensor nodes rely on finite energy sources and are often difficult to recharge—particularly when implanted—enhancing energy efficiency is a key research focus. Clustering techniques serve as an effective solution in this context by reducing energy demand, especially during inter-cluster communication, where information is exchanged between nodes and their respective CHs [4]. Reducing the distance between sensor nodes and their corresponding CHs is a key strategy for minimizing energy usage in Body Area Networks (BANs) [5]. Cluster-based communication techniques aim to extend the operational lifespan of the network by lowering communication overhead, establishing optimal routing paths, and enhancing both intra-cluster and inter-cluster data exchange efficiency [6]. As a result, the development of energy-aware routing strategies becomes vital for sustaining network functionality over time. This study addresses hierarchical structures within BANs and introduces a novel clustering mechanism powered by intelligent optimization, specifically the Firefly Algorithm, to enhance energy efficiency among sensor nodes. To assess the performance of the proposed approach, simulations are carried out using OPNET version 11.5 [7]. The effectiveness of the method is benchmarked against two established models: the NODIC protocol [8] and the IEEE 802.15.6 standard, which is widely implemented in BAN environments. Key evaluation criteria include end-to-end transmission delay, energy consumption levels, signal-to-noise ratio (SNR), and overall data throughput.

2. RELATED WORKS

Due to the unique requirements of WBANs, sensor nodes used in such systems must be compact, affordable, and energy-efficient. However, this leads to inherent limitations, particularly regarding restricted power capacity. Numerous studies have sought to address these limitations by proposing strategies that focus on energy preservation. Among these, clustering has proven to be a highly efficient approach for organizing the network and managing communication among nodes. While dynamic clustering and the periodic reelection of CHs generally consume limited energy, the initial setup and coordination phases—such as forming clusters and selecting heads—can result in notable energy expenditure due to the volume of control messages exchanged. A major ongoing challenge in WBSNs is ensuring reliable and timely data delivery, especially for applications like elderly healthcare monitoring. Performance issues such as inconsistent data transmission, significant latency, and high power consumption further complicate this objective. To tackle these challenges, the study presented in [9] introduces an algorithm called Enhanced Reliability, Energy-Efficient, and Latency-aware (EREEAL), designed specifically to improve the accuracy and efficiency of data communication in WBSNs. EREEAL reduces packet loss, decreases latency, and improves overall communication reliability by leveraging Time-Division Multiple Access (TDMA), allowing sensors to transmit data in designated time intervals. Additionally, it limits the transmission of repetitive or non-essential data, leading to significant energy savings. Simulation findings confirm that the EREEAL algorithm enhances communication stability, lowers delay, and reduces power consumption. This contributes to more efficient remote patient monitoring by reducing interference among sensors and limiting data loss. By optimizing network performance and ensuring timely, accurate health data transmission, the method allows healthcare providers to perform continuous, remote diagnostics with greater reliability. In [10], a protocol known as Energy-Aware Routing (EAR) is introduced, specifically aimed at reducing power consumption in BANs by analyzing

the quality of wireless links. The EAR approach integrates two essential parameters—residual energy and link stability—into its decision-making framework for route selection. Each node evaluates neighboring links and remaining energy levels to determine the optimal forwarding path using a weighted formula. This formula considers factors such as initial battery capacity, signal reliability, and the node's current energy status. The objective of this strategy is to evenly distribute the communication load among nodes, enhance routing efficiency, and extend the network's operational life. Simulation-based validation was conducted to assess the protocol's performance across several criteria, including energy usage, packet delivery success, network delay, and throughput. The outcomes demonstrated that the EAR protocol significantly reduces power drain while maintaining effective and dependable data transmission within WBAN environments.

Furthermore, the research presented in [11] introduces a robust multi-path routing protocol designed to deliver stable and efficient communication in WBANs. This protocol emphasizes both service quality prediction and intelligent bandwidth utilization. It employs Time-Division Multiple Access (TDMA) for structured communication, enabling optimal bandwidth management. The protocol continuously evaluates service quality by sustaining reliable routes between source and destination nodes. When data transmission is initiated, it identifies unused time intervals (idle slots) and calculates available bandwidth based on the current connectivity between nodes. Additionally, it enhances route stability by monitoring mobility trends, collecting statistical data on adjacent nodes, and estimating link expiration durations. These predictive capabilities help preserve route integrity and support timely route repairs, minimizing interruptions in data flow. While the protocol excels in load balancing, data scheduling, and link resilience, it does require additional bandwidth for its routing operations. This can lead to elevated power demands on individual nodes, potentially shortening their battery lifespan despite improved communication performance. In [12], the authors propose an energy-optimized routing protocol tailored for WBANs, with the goal of enhancing power efficiency during data transmission. This protocol designates two nodes responsible for handling sensitive data and introduces a cost function based on two primary parameters: the shortest communication distance and the highest available energy. Since longer transmission distances inherently require more energy, minimizing distance is a key factor. Additionally, energy thresholds are established, and nodes that are depleted or close to depletion are given lower priority. The protocol adopts a two-tiered approach—routine data is transmitted via multi-hop routing to conserve energy, while critical information is sent directly through a single-hop path to ensure speed and reliability. The sender node's role involves gathering data and forwarding it to the sink node, dynamically choosing relay nodes based on their proximity to the sink and their residual energy levels. Signal strength or link quality is also a critical criterion when selecting the forwarding node. It's worth noting that a sensor with a high energy reserve might not always have a stable communication link, highlighting the need to consider both energy and signal integrity when establishing connections.

Separately, radiation therapy remains a key method in the treatment of cancer, targeting tumor cells with high-energy rays to either eliminate or restrict their proliferation. Precision in targeting is crucial, as even slight patient movements or respiratory activity can alter the tumor's position during therapy sessions. To counteract this, real-time tracking of tumor movement is essential to ensure accurate radiation delivery while sparing surrounding healthy tissue from exposure.

In [13], the researchers present a novel tumor-tracking method based on spatial sparsity principles. Their approach employs a radio frequency (RF) emitter embedded within the body to generate signals, which are then captured by a strategically placed array of sensors located beneath the patient. These signals are processed to accurately determine the tumor's location. Unlike conventional techniques that typically depend on magnetic transmitters and numerous sensors, this method achieves high precision with fewer sensing elements. Two testing scenarios were explored: one with clearly defined tissue structure and another with uncertain or poorly defined tissue boundaries. In both cases, the approach demonstrated strong accuracy, making it a promising alternative for real-time tumor localization, particularly in clinical settings where tissue delineation may be challenging. Energy efficiency is a pivotal concern in the design and operation of WSNs, as reducing power usage is crucial for extending the operational lifespan of individual sensor nodes. In response to this need, researchers in [14] propose a novel routing framework for WBASNs that integrates a clustering algorithm to minimize energy consumption. This method emphasizes a centralized clustering strategy to create a tree-like routing structure among sensor nodes, thereby shortening communication distances. By enforcing a uniform cluster layout, the approach ensures that nodes are evenly distributed. The clustering process accounts for both the spatial distance between nodes and their remaining energy levels, enabling the strategic selection of source nodes. Within each cluster, a multi-phase adaptive mechanism is employed to optimize intra-cluster communication by limiting data transmission ranges. This strategy balances energy use across the network and reduces data forwarding costs. Simulation outcomes confirm that this method leads to notable reductions in energy expenditure, thereby extending the active life of sensor nodes and improving overall network efficiency.

In [15], the authors explore an innovative technique that utilizes backup, or "reserve," nodes to limit the frequency of message exchanges in clustering-based protocols, thereby conserving energy. This reserve mechanism is specifically

designed to reduce the overhead incurred during cluster formation and the election of CHs. The proposed algorithm introduces a reserve phase during the initial setup of the network. By activating this phase early in the network's lifecycle, the approach significantly lowers the number of control messages needed during dynamic clustering operations. The algorithm's performance is benchmarked against the LEACH protocol, and results demonstrate that the inclusion of reserve nodes can markedly enhance energy conservation and extend network longevity. This strategy proves effective in managing communication load while maintaining network functionality with reduced power consumption. In paper [16], the authors introduce a novel exploratory routing algorithm, the QoS Multi-objective Hybrid Routing Algorithm (Q-MOHRA), specifically designed for heterogeneous sensor networks. Q-MOHRA evaluates multiple factors, including energy consumption, link quality, and route delay, when determining the optimal route for data transmission. The primary goal of this algorithm is to ensure service quality across the network by simultaneously addressing various objectives. By integrating considerations of energy efficiency, link quality, and route delay, Q-MOHRA enables well-informed decisions about the best path for data transmission, ultimately improving overall network performance while balancing energy consumption and ensuring reliable data transfer.

3. THE SUGGESTED METHOD

The limited energy resources in sensor nodes make it essential to optimize energy consumption in order to extend the lifespan of WSNs. The development of routing protocols based on intelligent methods can significantly contribute to enhancing network longevity. This paper proposes the use of the FA to design an efficient clustering algorithm for WSNs, with the primary goal of improving energy efficiency. In a different context, fireflies display a fascinating natural phenomenon through their luminous flashes, which are visible in tropical and temperate regions during the summer months. There are approximately 2,000 species of fireflies, each producing unique and short-lived flashes. The mechanism behind flash production, known as bioluminescence, remains an area of ongoing research and debate. Fireflies use these flashes for two main purposes: attracting mates and capturing prey. The patterns of the flashes, their rhythmic nature, flash rates, and the distances between signals are critical for firefly communication within species. Females respond to the flashes emitted by males, while some species, such as Photuris, mimic the flashes of other species to deceive and prey upon them. The intensity of firefly flashes follows an inverse-square law, where the intensity (denoted as I) diminishes as the distance (r) from the light source increases. This relationship is mathematically represented as $I \propto 1 / r^2$. Additionally, light absorption by the surrounding air further reduces the brightness as the distance increases. As a result, firefly flashes are typically visible only over short distances, generally a few hundred meters at night, which is sufficient for communication between fireflies. The FA, introduced by Yang in 2008, was inspired by the light-emitting behavior of fireflies [17]. To streamline the algorithm's definition, three key assumptions were made:

- All fireflies are of a single sex and are naturally attracted to one another for mating purposes, regardless of gender.
- The attractiveness of a firefly is determined by its luminous intensity. When two fireflies interact, the one with lower luminosity is drawn toward the one with higher luminosity. The attraction strength is directly proportional to their luminous intensities, which diminish as the distance between them increases. If the luminous intensities of both fireflies are equal, their movement becomes random.
- The luminous intensity of a firefly is representative of the value of a target function.
- Two key factors must be considered in the FA: variations in luminous intensity and the formulation of attractiveness or deceptiveness. To streamline the algorithm, the absorption of fireflies can be based on their luminosity, which is determined by the target function.

For maximization problems, the luminosity I of a firefly at a specific position x can be considered as $I(x) \alpha f(x)$, where f(x) represents the value of the target function. Although the attractiveness β is relative, it must be observed by other fireflies for their judgment. Therefore, this value may vary based on the distance r_{ij} between firefly i and firefly j. As the attractiveness of a firefly is influenced by the intensity of the flashes perceived by neighboring fireflies, the attractiveness β can be defined using Equation 1.

$$\beta = \beta_0 e^{-\gamma r^2} \quad (1)$$

The attractiveness parameter β_0 represents the attractiveness value at a distance of r=0, indicating the initial attractiveness between fireflies.

The distance between firefly i and firefly j, given their respective positions x_i and x_j, can be calculated using the Cartesian distance formula, as shown in Equation 2.

$$r_{ij} = x_i - x_j = \sqrt{\sum_{k=1}^{d}\left(x_{i,k} - x_{j,k}\right)^2} \quad (2)$$

That $x_{i,k}$ is the K_{th} component of the coordinate distance of x_i of the i_{th} firefly. In two-dimensional problems, the value of the distance is calculated according to Equation 3.

$$r_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2} \quad (3)$$

(3)

The movement of firefly i towards a more attractive (brighter) firefly j can be defined using Equation 4:

$$x_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}\left(x_j - x_i\right) + \alpha\, \grave{o}_i \quad (4)$$

Where β_0 is the attractiveness, α is the random generator parameter, and $\grave{o}_i$ is the random vector of the numbers shown by a Gaussian distribution or a uniform distribution.

For most implementations, we can consider β_0 =1 and $\alpha \in$ [0,1]. In this formula, the parameter γ indicates the attractiveness changes, and its value is used in determining the convergence velocity and how the FA algorithm behaves. In theory, it is $\gamma \in [0, \infty)$.

In this section, we will explain the proposed method for improving energy consumption in the body area sensor network.

Two important aspects are considered in the FA: the clustering process and the routing mechanism. The proposed method introduces a clustering algorithm based on the FA for body area sensor networks. It utilizes factors such as residual energy, inter-cluster distance, and distance to the sink node to determine the CH. In this approach, both sensor nodes and the sink node are treated as fireflies. A virtual backbone is then formed using the CHs to simplify data routing, with the sink node, having the highest luminosity, acting as the root of this backbone.

The algorithm consists of two primary phases: clustering and routing. During the clustering phase, fireflies are grouped into separate clusters. Each firefly independently adjusts its timing before competing to become the CH. The scheduling of a sensor node i, represented as t(i), is determined using Equation 5.

$$t\left(i\right) = \frac{E_m\left(i\right) - E_r\left(i\right)}{E_m\left(i\right)} \times T_{CH} \quad (5)$$

Where T_{CH} is the maximum time allotted for selecting the CH. $E_m\left(i\right)$ and $E_r\left(i\right)$ are the maximum initial energy and residual energy of i firefly, respectively. Also, the luminous intensity of I is considered according to Equation 6.

$$I \propto \left(\frac{1}{r_{iS}^2}\right) + E_r\left(i\right) \quad (6)$$

The algorithm uses Equation 6 to select the CH based on higher residual energy and a shorter distance to the sink node. A firefly with these characteristics is considered more attractive than others. After the scheduling process is completed, node i designates itself as the CH and broadcasts an introduction message within its transmission range R. The introduction message includes the node's ID, residual energy $E_r(i)$, and spatial information P(i). Upon receiving this message, firefly j withdraws from being a CH candidate, cancels its scheduling process, and functions as a normal

node until the next round. Additionally, firefly j maintains a list, referred to as the Neighbor CHs Collection ($N_{Ch}(j)$), which contains the fireflies that have been identified as CHs. In the subsequent rounds, firefly j determines its cluster membership using the information stored in $N_{Ch}(j)$. To form clusters, each normal node decides its membership by attempting to join one of the CHs listed in $N_{Ch}(j)$. Firefly j evaluates the attractiveness of the CHs in the $N_{Ch}(j)$ list using Equation 8.

$$\beta_{K,j=}E_r N_{Ch}(j)+\left(1/\left(r_{k,j}{}^2\right)\right) \quad (7)$$

$$E_r N_{Ch}(j)=\frac{\sum_{k=1}^{m}E_r(v_k)}{m} \quad (8)$$

Where $\beta_{K,j}$ is the attractiveness of the CH k from the firefly j's point of view, $E_r N_{Ch}(j)$ is the average residual energy of the CHs in N_{Ch} (j) list of the j-th firefly, which is calculated according to Equation 8, $r_{k,j}$ is the distance of the jth firefly from the kth CH in the $N_{Ch}(j)$ list.

Regarding Equation 7, out of the CH fireflies in the $N_{ch}(j)$ list, the firefly j chooses a firefly as the CH, which is less distant from that CH and also has the most residual energy. Then, the firefly j joins the nearest CH, the remaining energy of which is greater than or equal to $E_r N_{ch}(j)$. Accordingly, clusters are formed.

In the data routing phase to the sink, a directed virtual backbone (DVB) rooted in the sink is created for all CHs. In the beginning, the sink node sends a message requesting the route to all the CHs in its 2R range. This message contains data such as node ID, level (L), and spatial information. It is considered that the sink node level is zero. Once the CH u receives a message, the node raises its level a little bit and chooses the sink node as its father PN (u) = sink. To explain, the level of all the CHs within the 3R sink range is 1. Similarly, the CH u sends a message requesting the route again to all the CHs within its range. The message contains information such as ID, L(u), $E_r(u)$, and spatial information P(u). If a CH v receives the message, and if its level is less than or equal to the node u level, it ignores that message. Otherwise, it increases the value of the level by one greater than the level of u and places it as a PN. In the same way, these steps will continue, and all the CHs will broadcast the message requesting the completion of the DVB creation process. A CH may have multiple PNs inside the DVB, so it will have multiple routes to the sink.

4. SIMULATION OF THE PROPOSED METHOD

4.1 Simulation environment

The effectiveness of the proposed approach was assessed using version 11.5 of the OPNET simulation tool. The primary aim of this simulation was to benchmark the performance of the proposed algorithm against the NODIC protocol. OPNET operates based on a hierarchical architecture comprising three integral layers: Network, Node, and Process. This structure enables visual modeling of WSN configurations and offers extensive customization of simulation settings. The platform facilitates in-depth performance analysis and comparison between different routing techniques. A comprehensive overview of the simulation parameters is outlined in Table 1.

Table. 1 Simulation Parameters

Parameter	Amount
The Way of Scattering Nodes in the Environment	Random
The Size of the Simulation Environment	10m × 10m× 10m
Type of Sending	CBR
Packet Size	1024 byte
Packet Rate	250kbps
Simulation Type	100 seconds
Mac Layer	IEEE802.15.6

The Initial Amount of Energy	200 joules
The Number of Sinks	1
The Number of Nodes	60
The Range of Radio Communication	1 meters
Packet Inter Arrival Time	Constant

The routing approach introduced in this work is tailored for healthcare monitoring systems. It operates under the assumption that all sensor nodes are within each other's communication range and are aware of their neighboring node positions. Central to the network is the sink node, which has superior processing and communication capabilities compared to ordinary sensor nodes and is primarily responsible for aggregating the collected data. As shown in Fig. 1, the architecture of the WBAN is presented in both its structural and operational forms according to the proposed method. Within this configuration, the sink node is positioned internally within the body, whereas the gateway node is placed in proximity to the patient or at another suitable location. Due to the impact of human movement, which can disrupt data collection, the sink node is designed to reliably gather information despite potential mobility-induced challenges. To assess the effectiveness of the proposed strategy, simulations were conducted on a network comprising 60 sensor nodes under three separate conditions. The first configuration distributes nodes randomly based on the IEEE 802.15.6 standard, known for offering 360-degree coverage, one-dimensional data support, low-power operation, cost efficiency, and real-time communication capability. The second configuration applies the FBC technique introduced in this study, where clusters are dynamically formed based on energy and proximity metrics. The third scenario employs the NODIC protocol, wherein data is relayed to the sink node via CH nodes. Importantly, the same network setup is maintained across all scenarios to ensure consistency and validity in comparative evaluation.

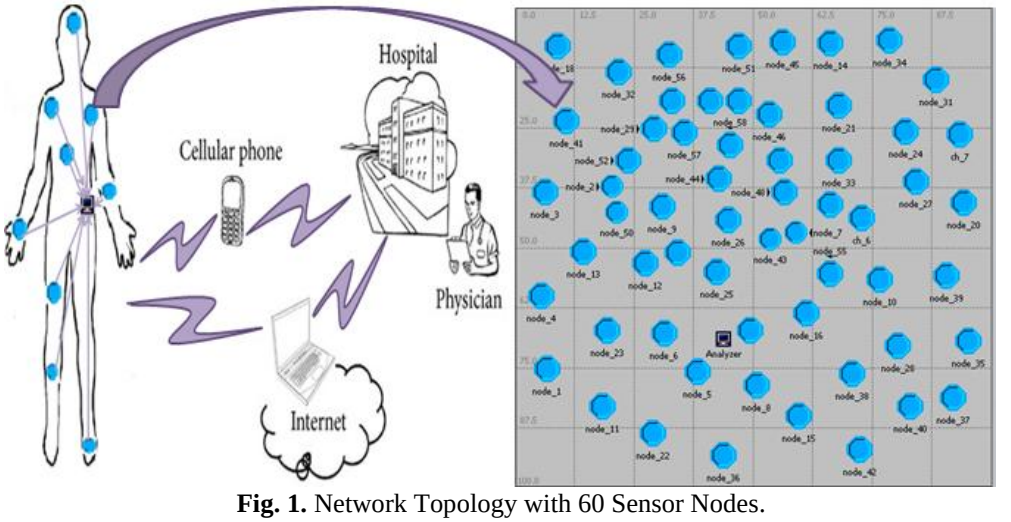
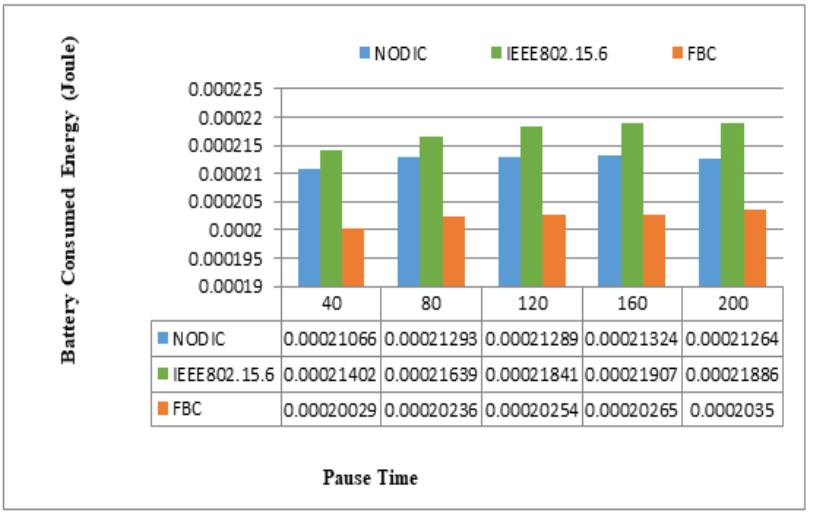


Fig. 1. Network Topology with 60 Sensor Nodes.

4.2 Simulation Results

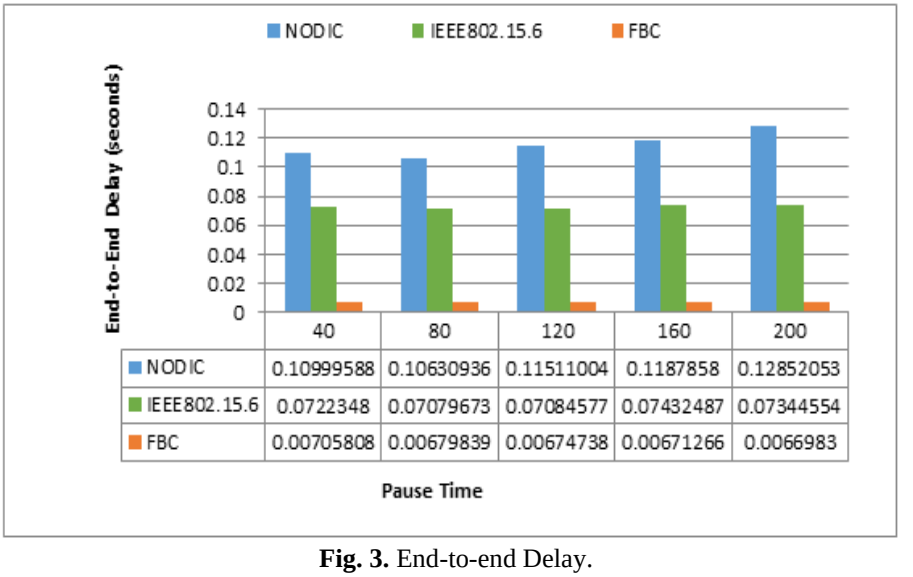
Fig. 2 illustrates a comparison of average energy consumption among three approaches: the proposed algorithm, IEEE 802.15.6, and the NODIC protocol. The y-axis denotes the amount of energy consumed, while the x-axis corresponds to simulation time. Energy consumption here refers to the total energy spent by nodes for transmitting and receiving data, as well as their anticipated energy usage. Among the evaluated methods, the IEEE 802.15.6 protocol demonstrates the highest power consumption. This is primarily because its sub-nodes transmit data to the sink node without considering their residual energy, leading to inefficient use of resources. The NODIC protocol, on the other hand, selects CHs based on node location and the number of neighboring nodes with adequate energy reserves. However, in scenarios where suitable nodes are not found, it does not perform re-clustering. This can result in the selection of CHs with insufficient energy or few neighboring nodes in future rounds, accelerating energy depletion and destabilizing the network structure. In contrast, the proposed method leverages a FBC strategy that chooses CHs with higher remaining energy and closer proximity to the sink. Member nodes are grouped based on their distance to the selected CH, which significantly reduces the energy required for communication. This strategy not only conserves energy but also maintains effective data transfer across the network.



	40	80	120	160	200
NODIC	0.00021066	0.00021293	0.00021289	0.00021324	0.00021264
IEEE802.15.6	0.00021402	0.00021639	0.00021841	0.00021907	0.00021886
FBC	0.00020029	0.00020236	0.00020254	0.00020265	0.0002035

Fig. 2. Average Network Energy Consumption.

Fig. 3 compares end-to-end delays observed in three scenarios: the proposed method, the IEEE 802.15.6 protocol, and the NODIC protocol. The y-axis represents the time delay from source to destination, while the x-axis shows the simulation time. End-to-end delay is defined as the total time required for a data packet to travel from the source node to the sink. In the IEEE 802.15.6 protocol, increased delays are noted due to nodes initiating transmissions with insufficient energy, often leading to incomplete data delivery and communication disruptions. Likewise, in the NODIC approach, premature energy depletion of CH nodes hinders their ability to forward the collected data effectively, causing additional delay. On the other hand, the proposed algorithm achieves reduced end-to-end latency by assigning CH roles to nodes with higher residual energy and organizing member nodes based on their closeness to the CH. This strategy ensures smoother and more efficient data transmission, thereby lowering delay times in comparison to the other two protocols.

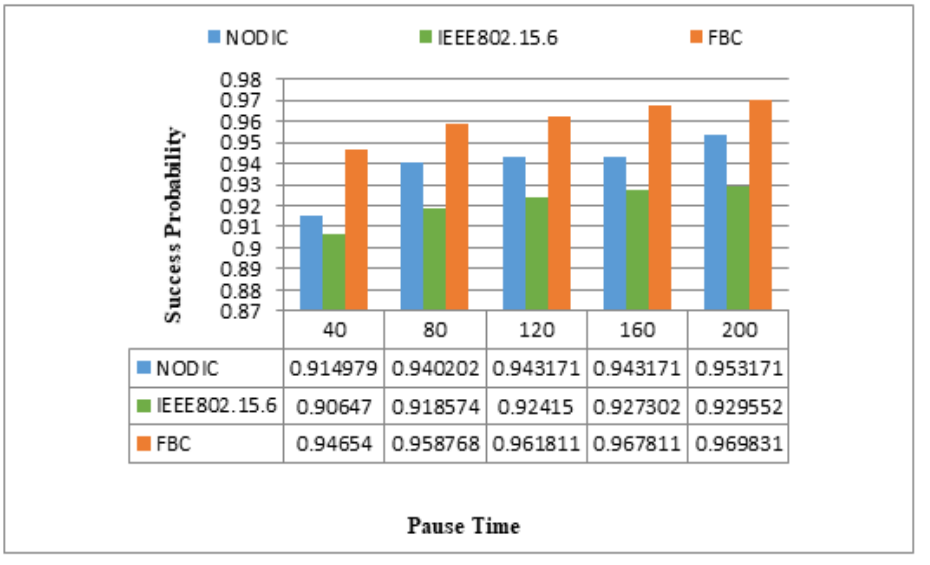


	40	80	120	160	200
NODIC	0.10999588	0.10630936	0.11511004	0.1187858	0.12852053
IEEE802.15.6	0.0722348	0.07079673	0.07084577	0.07432487	0.07344554
FBC	0.00705808	0.00679839	0.00674738	0.00671266	0.0066983

Fig. 3. End-to-end Delay.

Fig. 4 illustrates the likelihood of successful data delivery to the sink node under three different protocol conditions. The x-axis denotes the simulation time, while the y-axis reflects the calculated success rate of data transmission. This metric is obtained by dividing the amount of data correctly received at the sink by the total data volume transmitted at that moment, expressed in megabits per second. Analysis reveals that the IEEE 802.15.6 protocol demonstrates a lower packet delivery success rate compared to both the proposed and NODIC protocols. The lower performance of IEEE 802.15.6 is largely attributed to network congestion and the possible deactivation of nodes during operation. Conversely,

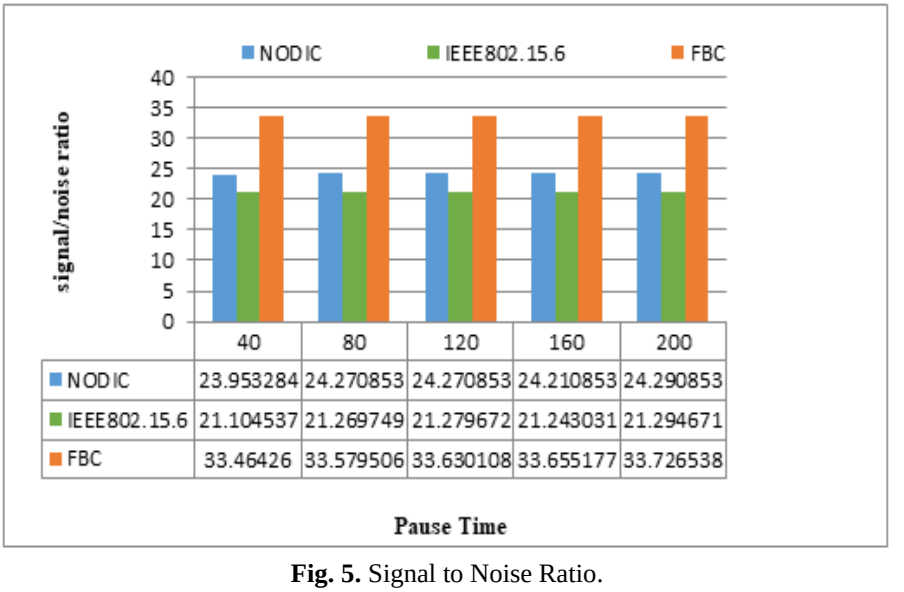
the proposed algorithm prioritizes the selection of sensor nodes with greater residual energy for cluster formation. This ensures that the routing paths remain stable and functional throughout the communication process, thereby reducing the risk of premature node failure and maintaining data flow integrity. As a result, the number of packets that successfully reach the sink node is notably increased. The method's emphasis on stable, energy-efficient routing contributes directly to the improved reliability of data delivery.



	40	80	120	160	200
NODIC	0.914979	0.940202	0.943171	0.943171	0.953171
IEEE802.15.6	0.90647	0.918574	0.92415	0.927302	0.929552
FBC	0.94654	0.958768	0.961811	0.967811	0.969831

Fig. 4. The Probability of Success in Sending Information to the Sink Node

Fig. 5 illustrates the comparison of signal-to-noise ratio (SNR) performance among three approaches: the newly developed algorithm, the IEEE 802.15.6 protocol, and the NODIC protocol. The x-axis indicates simulation time, while the y-axis reflects the SNR values. SNR is a key metric that quantifies the ratio of the intended signal's strength to the level of background interference—where greater values signify better signal clarity. The chart shows that the IEEE 802.15.6 protocol consistently produces lower SNR values than both the proposed and NODIC methods. This reduction in SNR can be attributed to the protocol's tendency to rely on less reliable transmission paths, which may introduce bit-level errors and data corruption. Moreover, network congestion and instability associated with the IEEE 802.15.6 protocol exacerbate these issues by increasing noise and reducing the integrity of the transmitted signal. In contrast, the proposed method and NODIC protocol offer enhanced route stability and more reliable data delivery, which in turn lead to higher and more consistent SNR measurements.



	40	80	120	160	200
NODIC	23.953284	24.270853	24.270853	24.210853	24.290853
IEEE802.15.6	21.104537	21.269749	21.279672	21.243031	21.294671
FBC	33.46426	33.579506	33.630108	33.655177	33.726538

Fig. 5. Signal to Noise Ratio.

5. CONCLUSION

This research addresses the problem of excessive energy usage in body area sensor networks, which are widely used in diverse applications. To mitigate this issue, an innovative clustering technique based on the FA is introduced. The effectiveness of this technique was assessed through simulation experiments conducted using the OPNET simulator. Comparative evaluations were performed against the IEEE 802.15.6 standard and the NODIC protocol. Key performance indicators, including power consumption, end-to-end transmission delay, signal-to-noise ratio, and the likelihood of successful data delivery to the sink, were analyzed. The findings indicate that the proposed method achieved superior performance and improved packet delivery accuracy compared to both benchmark protocols. These enhancements are largely due to the algorithm's ability to select stable communication paths that prioritize nodes with higher residual energy. Additionally, the approach integrates both global and local optimization strategies, enabling it to efficiently escape local optima while maintaining a fast convergence rate. Overall, the FA demonstrated robust clustering performance by effectively utilizing both exploratory and exploitative search mechanisms.

REFERENCES

[1] S. S. S. Farahani, **"Congestion control approaches applied to wireless sensor networks: A survey,"** *Journal of Electrical and Computer Engineering Innovations*, Vol. 6, No. 2, pp. 125-144, 2018.

[2] M. Ghaderi, V. Tabataba Vakili, and M. Sheikhan, **"STCS-GAF: Spatio-temporal compressive sensing in wireless sensor networks-A GAF-based approach, "** *Journal of Electrical and Computer Engineering Innovations (JECEI)*, Vol. 6, No. 2, pp. 159-172, 2018.

[3] A. T. Barth, M. A. Hanson, H. C. Powell Jr, D. Unluer, S. G. Wilson, and J. Lach, **"Body-coupled communication for body sensor networks, "** *in Proceedings of the ICST 3rd international conference on Body area networks*, pp. 1-4, 2008.

[4] E. I. Oyman and C. Ersoy, **"Multiple sink network design problem in large scale wireless sensor networks, "** *in 2004 IEEE international conference on communications (IEEE Cat. No. 04CH37577)*, Vol. 6, pp. 3663-3667, 2004.

[5] M. Hammoudeh and R. Newman, **"Adaptive routing in wireless sensor networks: QoS optimisation for enhanced application performance, "** *Information Fusion*, Vol. 22, pp. 3-15, 2015.

[6] M. Hammoudeh and R. Newman, **"Adaptive routing in wireless sensor networks: QoS optimisation for enhanced application performance, "** *Information Fusion*, Vol. 22, pp. 3-15, 2015.

[7] L. OPNET, **"Specialized Model: http://www. opnet. com, "** ed: LTE.

[8] İ. Abasıkeleş-Turgut and O. G. Hafif, **"NODIC: a novel distributed clustering routing protocol in WSNs by using a time-sharing approach for CH election, "** *Wireless Networks*, Vol. 22, pp. 1023-1034, 2016.

[9] K. Dhakal, A. Alsadoon, P. Prasad, R. S. Ali, L. Pham, and A. Elchouemi, **"A novel solution for a wireless body sensor network: Telehealth elderly people monitoring, "** *Egyptian Informatics Journal*, Vol. 21, no. 2, pp. 91-103, 2020.

[10] M. Ghafouri vaighan and M. A. Jabraeil Jamali, **"A multipath QoS multicast routing protocol based on link stability and** route **reliability in mobile ad-hoc networks, "** *Journal of Ambient Intelligence and Humanized Computing*, Vol. 10, No. 1, pp. 107-123, 2019.

[11] R. A. Khan et al**., "An energy efficient routing protocol for wireless body area sensor networks, "** *Wireless Personal Communications*, Vol. 99, pp. 1443-1454, 2018.

[12] M. Pourhomayoun, Z. Jin, and M. Fowler, **"Accurate tumor localization and tracking in radiation therapy using wireless body sensor networks, "** *Computers in Biology and Medicine*, Vol. 50, pp. 41-48, 2014.

[13] J. Y. Chang and P. H. Ju, **"An energy-saving routing architecture with a uniform clustering algorithm for wireless body sensor networks, "** *Future Generation Computer Systems*, Vol. 35, pp. 128-140, 2014.

[14] A. Zahedi, M. Arghavani, F. Parandin, and A. Arghavani**, "Energy efficient reservation-based cluster head selection in WSNs, "** *Wireless Personal Communications*, Vol. 100, pp. 667-679, 2018.

[15] N. Kulkarni, N. R. Prasad, and R. Prasad, **"Q-MOHRA: QoS assured multi-objective hybrid routing algorithm for heterogeneous WSN, "** *Wireless Personal Communications*, Vol. 100, pp. 255-266, 2018.

[16] X. S. Yang, **"Firefly algorithm, stochastic test functions and design optimisation, "** *International journal of bio-inspired computation*, Vol. 2, No. 2, pp. 78-84, 2010.

A Hybrid Approach for Intrusion Detection in the Internet of Things Using Harris Hawks Optimization and Deep Learning Algorithms

Reza Kohan[1], Hamid Barati[2], Ali Barati[3]

1- Institute of Artificial Intelligence and Social and Advanced Technologies, Dez.C., Islamic Azad University, Dezful, Iran.
Email: rezakohan.rk061@gmail.com
2- Institute of Artificial Intelligence and Social and Advanced Technologies, Dez.C., Islamic Azad University, Dezful, Iran.
Email: Hamid.barati@iau.ac.ir (Corresponding author)
3- Institute of Artificial Intelligence and Social and Advanced Technologies, Dez.C., Islamic Azad University, Dezful, Iran.
Email: alibarati@iau.ac.ir

ABSTRACT:
Intrusion detection in Internet of Things (IoT)-based smart cities is essential due to the increasing volume and complexity of cyberattacks. Traditional detection systems face two major challenges: achieving high accuracy and minimizing false alarms, particularly in large-scale and heterogeneous IoT networks. This paper proposes a novel hybrid intrusion detection system that combines Harris Hawks Optimization (HHO) for feature selection with a multi-layer neural network enhanced by learning automata for adaptive classification. The HHO algorithm efficiently reduces input dimensionality by selecting the most relevant features, while the learning automata optimize the network's weights dynamically, improving training stability and robustness. The proposed system is evaluated using the KDDCup99 dataset under both binary and multiclass scenarios. Experimental results show an average accuracy of 96.53%, a true positive rate (TPR) of 94.91%, and a false positive rate (FPR) of 2.80%. Compared to recent baseline models, the proposed method demonstrates superior performance in accuracy and false alarm reduction, confirming its suitability for real-time intrusion detection in dynamic IoT-based environments.

KEYWORDS: Intrusion Detection, Internet of Things, Harris Hawks Optimization, Neural Network, Learning Automata.

3. INTRODUCTION

With the rapid advancement of smart city technologies and the widespread deployment of Internet of Things (IoT) devices, urban environments are becoming increasingly interconnected, automated, and data-driven. These systems rely on distributed sensor networks, cloud-based platforms, and intelligent decision-making to improve urban services such as traffic control, energy management, public safety, and environmental monitoring [1]. However, the growing complexity and openness of IoT-based smart city infrastructures have also made them highly susceptible to cyberattacks, posing significant threats to data confidentiality, system availability, and operational integrity [2].

Intrusion Detection Systems (IDS) play a crucial role in safeguarding smart city networks by monitoring network traffic and detecting anomalous or malicious behavior [3]. Traditional IDS approaches, including signature-based and rule-based models, often fall short in detecting novel or evolving attacks and struggle to adapt to the dynamic and large-scale nature of IoT networks [4]. Moreover, the high volume of data and the heterogeneity of devices introduce significant challenges in scalability, accuracy, and false alarm reduction [5].

To address these challenges, machine learning (ML) and metaheuristic optimization algorithms have gained attention for their ability to model complex attack patterns and improve detection performance [6]. In particular, feature selection

Paper type: Research paper
https://doi.org/xxx
Received: 18 January 2025, Revised: 15 February 2025, Accepted: 3 May 2025, Published: 1 June 2025
How to cite this paper: R. Kohan, H. Barati, A. Barati, **"A Hybrid Approach for Intrusion Detection in the Internet of Things Using Harris Hawks Optimization and Deep Learning Algorithms"**, *Majlesi Journal of Telecommunication Devices*, Vol. 14, No. 2, pp. 89-104, 2025.

has emerged as a vital preprocessing step to reduce redundant information, improve classifier efficiency, and mitigate the impact of noisy or irrelevant data [7]. Similarly, the quality of the classifier, including its training stability and adaptability, plays a key role in determining the effectiveness of the IDS [8].

We propose a novel hybrid intrusion detection method that integrates Harris Hawks Optimization (HHO) for optimal feature selection and a multi-layer neural network optimized using learning automata for adaptive classification. The HHO algorithm, inspired by the cooperative hunting behavior of Harris hawks, efficiently explores the feature space to identify the most informative attributes. Meanwhile, the learning automata dynamically adjust the weights of the neural network based on training feedback, enhancing learning stability and classification accuracy.

The contributions of this work are summarized as follows:

- We design a feature selection mechanism based on Harris Hawks Optimization to reduce computational complexity and improve detection performance.
- We introduce a learning automata-driven neural network training approach to enhance adaptability and avoid overfitting.
- We validate the proposed system using the KDDcup99 benchmark dataset and demonstrate superior performance in terms of accuracy, false positive rate, and robustness across different attack types.
- We compare our model with two recent methods and show its effectiveness in handling class imbalance and rare attack detection.

The remainder of the paper is organized as follows: Section 2 reviews related work on intrusion detection in IoT environments. Section 3 presents the problem formulation and background concepts. Section 4 details the proposed methodology. Section 5 discusses the experimental results and comparative analysis. Finally, Section 6 concludes the paper and outlines directions for future work.

4. RELATED WORK

In recent years, intrusion detection in the Internet of Things (IoT) has attracted growing attention due to the proliferation of connected devices and the increasing sophistication of cyber threats. A variety of intelligent and hybrid models have been proposed to address the unique constraints of IoT environments, including resource limitations, data heterogeneity, and real-time processing requirements. Khan et al. [9] developed a deep neural network (DNN)-based IDS tailored for MQTT-based IoT environments. Their model leveraged three levels of feature abstraction—packet-flow, uni-flow, and bi-flow—achieving over 99% accuracy in uni-flow and bi-flow detection. However, classification accuracy declined for packet-flow due to class imbalance.

Zhao et al. [10] proposed a lightweight DNN with PCA-based feature reduction for low-resource IoT devices. Their architecture integrated inverted residuals and channel-wise fusion mechanisms to optimize performance. While the model maintained high accuracy and efficiency, its performance varied across feature types, particularly for packet-level data.

Wahab [11] introduced an online deep learning approach for intrusion detection, which addressed concept drift via Hedge Backpropagation. This two-phase method—detecting changes and adapting the model—enabled real-time learning and outlier detection, demonstrating strong resistance to evolving attack patterns.

Vishwakarma and Kesswani [12] presented the DIDS system based on DNN, designed for real-time detection using Netflow-based data. Their approach combined PCA for dimensionality reduction, dropout layers to avoid overfitting, and Hedge Backpropagation for adaptability. The model achieved high classification accuracy in both binary and multiclass tasks.

Chatterjee and Hanawal [13] developed a hybrid ensemble-based IDS called PHEC, combining KNN and Random Forest to improve detection accuracy while reducing false positives. They extended this model to a federated learning setting to preserve data privacy across distributed IoT devices. A noise-tolerant version of PHEC was also introduced to address label noise using a weighted convex loss. Evaluations on four benchmark datasets showed high TPR and low FPR in both clean and noisy scenarios, with federated performance closely matching the centralized setup.

Ngo et al. [14] designed the HH-NIDS framework using lightweight neural networks implemented on heterogeneous hardware platforms, including MAX78000EVKIT and FPGA. Their system achieved over 99% accuracy on UNSW-NB15 and IoT-23 datasets while significantly reducing energy consumption and inference time.

Awajan [15] proposed a dynamic IDS using deep neural networks equipped with adaptive feature extraction modules. Their model achieved a 93.21% average detection rate across five common attack types. Despite high performance, the system's complexity posed challenges for real-time deployment.

Wang et al. [16] introduced BT-TPF, a knowledge-distillation-based framework using a large Vision Transformer (teacher) and a compact Poolformer (student) for lightweight IDS. Their model achieved over 99% accuracy on CIC-IDS2017 and TON_IoT datasets, though training complexity and reliance on a powerful teacher model remained concerns.

Qaddos et al. [17] proposed a similar lightweight detection system by combining Siamese networks and knowledge distillation, also achieving over 99% accuracy. The model was optimized for limited-resource environments but required precise calibration and rich training data to maintain robustness.

Wang et al. [18] presented FeCo, a federated contrastive learning framework designed to preserve data privacy while improving detection performance. FeCo effectively handled non-IID data and achieved up to 8% accuracy improvement over state-of-the-art models, though computational complexity in large-scale deployments remained a limitation.

Lin et al. [19] proposed E-GRACL, a GNN-based IDS that modeled traffic correlations via graph structures. By enhancing the GraphSAGE model with attention and gating mechanisms, and integrating contrastive learning, E-GRACL demonstrated high accuracy in both binary and multiclass settings, but at the cost of computational overhead and complex model tuning.

Table 1 provides a comparative overview of recent IDS approaches in IoT contexts, summarizing their core techniques, advantages, and known limitations.

Table 1. Comparative summary of recent intrusion detection approaches in IoT environments

Ref	Year	Method / Model	Dataset(s) Used	Key Techniques	Strengths	Limitations
[9]	2021	DNN for MQTT IDS	MQTT-IoT-IDS2020	Multi-level flow (Uni/Bi/Packet), Adam optimizer	High accuracy, flow-specific feature modeling	Poor performance on Packet-flow due to class imbalance
[10]	2021	Lightweight DNN	Real-world datasets	PCA + DNN + NID loss	Compact model for low-resource devices	Lower precision on complex attack types
[11]	2022	Online DL + Hedge Backprop	Simulated stream data	Concept drift handling, PCA, outlier detection	Fast adaptation, reduced training-test gap	Complex update mechanism
[12]	2022	DIDS – Real-time DNN	NetFlow-based IoT traffic	PCA + Dropout + Hedge Backprop	Real-time detection, adaptive learning	May require high processing capacity
[13]	2022	PHEC + Federated Learning	NSL-KDD, DS2OS, Gas Pipeline, Water Tank	KNN + RF ensemble, Fed-stacking, weighted loss	Privacy-preserving, noise-tolerant	Higher computation, edge-device constraints
[14]	2023	HH-NIDS (Heterogeneous HW)	UNSW-NB15, IoT-23	Lightweight NN + FPGA & microcontroller	Energy-efficient, real-time inference	Needs specialized hardware
[15]	2023	Adaptive DNN	Custom dataset	Dynamic feature extraction, attack reduction phase	Modular, flexible model	High model complexity, slow inference
[16]	2024	BT-TPF (Knowledge Distill.)	CIC-IDS2017, TON_IoT	Vision Transformer + Poolformer	Model compression, high accuracy	Training dependent on teacher model
[17]	2024	Enhanced BT-TPF	CIC-IDS2017, TON_IoT	Siamese Net + Knowledge distillation	Efficient for constrained IoT	Sensitivity to data variation
[18]	2025	FeCo (Federated + Contrastive)	NSL-KDD, BaIoT	Fed. learning + Feature filtering + Contrastive learning	Privacy-preserving, handles non-IID	Slower convergence, complex coordination

[19]	2025	E-GRACL (Graph-based IDS)	Three benchmark sets	GNN (GraphSAGE+) + Global attention + Contrastive learning	Captures topology, high detection in binary/multiclass	High computational cost, tuning required

5. PROBLEM FORMULATION AND BACKGROUND CONCEPTS

3.1. Problem Definition

Intrusion detection in IoT-enabled smart city environments presents a challenging task due to the high volume of data, the heterogeneous nature of connected devices, and the presence of both known and unknown cyber threats. The goal is to design a detection system capable of identifying a wide variety of attacks—including volumetric, probing, and stealthy intrusions—while maintaining high accuracy and low false alarm rates [20].

Formally, let $D = \{x_1, x_2, \dots, x_n\}$ be a dataset of network traffic records, where each instance $x_i \in \mathbb{R}^m$ is characterized by m features. The task is to learn a function $f: \mathbb{R}^m \to C$, where C= {normal, DoS, Probe, R2L, U2R}, that maps each input instance to its correct class label. Due to high dimensionality and data imbalance, the learning function must be trained on a reduced and optimized feature subset $F' \subseteq F$, where $|F'| \ll |F|$, without sacrificing classification performance.

This problem requires a two-step solution:

1. Optimal Feature Selection: Identify a compact subset of relevant features to improve classification accuracy and reduce computational cost.
2. Robust Classification: Train an adaptable and generalizable model capable of handling diverse and imbalanced data efficiently.

3.2. Background Concepts

3.2.1 Feature Selection in Intrusion Detection

Feature selection plays a critical role in reducing noise and irrelevant information in high-dimensional data. It aims to select a subset of the most discriminative features that contribute significantly to the learning process. This not only improves the detection accuracy but also speeds up the training phase and prevents over fitting [21]. In intrusion detection, feature selection is particularly important because many features in datasets like KDDcup99 [22] are redundant or weakly correlated with attack behavior.

3.2.2. Harris Hawks Optimization (HHO)

Harris Hawks Optimization (HHO) [23] is a recent nature-inspired metaheuristic algorithm that simulates the cooperative hunting strategy of Harris hawks in nature. It employs both exploration (global search) and exploitation (local search) mechanisms, adapting its strategy based on the prey's escaping energy and the current fitness of candidate solutions. In the context of feature selection, each hawk represents a binary vector corresponding to a potential subset of features, and the fitness function evaluates classification performance using this subset. HHO has been shown to converge rapidly and explore diverse regions of the solution space, making it suitable for dynamic IDS environments.

3.2.3. Neural Networks for Classification

Artificial Neural Networks (ANNs) [24] are widely used in classification tasks due to their strong learning capabilities and non-linear mapping functions. In this work, a multi-layer feedforward neural network is employed as the core classifier. It is composed of an input layer, one or more hidden layers, and an output layer, and is trained using supervised learning with backpropagation. However, traditional training may suffer from local minima and slow convergence, especially when using large or imbalanced datasets.

3.2.4. Learning Automata for Adaptive Training

To overcome the limitations of standard backpropagation, Learning Automata (LA) [25] are integrated into the training process. A learning automaton is a probabilistic decision-making mechanism that adjusts its actions based on feedback from the environment. In this context, the automaton evaluates training progress by monitoring changes in classification error. It rewards configurations that reduce error and penalizes those that degrade performance. This adaptive process helps in avoiding local minima and improves generalization in neural networks, particularly under non-stationary or noisy training conditions.

6. PROPOSED METHOD

This section describes the proposed intrusion detection method for smart city-based Internet of Things (IoT) environments. The approach combines advanced feature selection through Harris Hawks Optimization (HHO) with an enhanced multi-layer neural network (MLNN) supported by learning automata for adaptive weight tuning. The method follows a structured pipeline including data preprocessing, feature selection, classification, and training optimization. A complete overview of the proposed architecture is illustrated in Fig. 1.

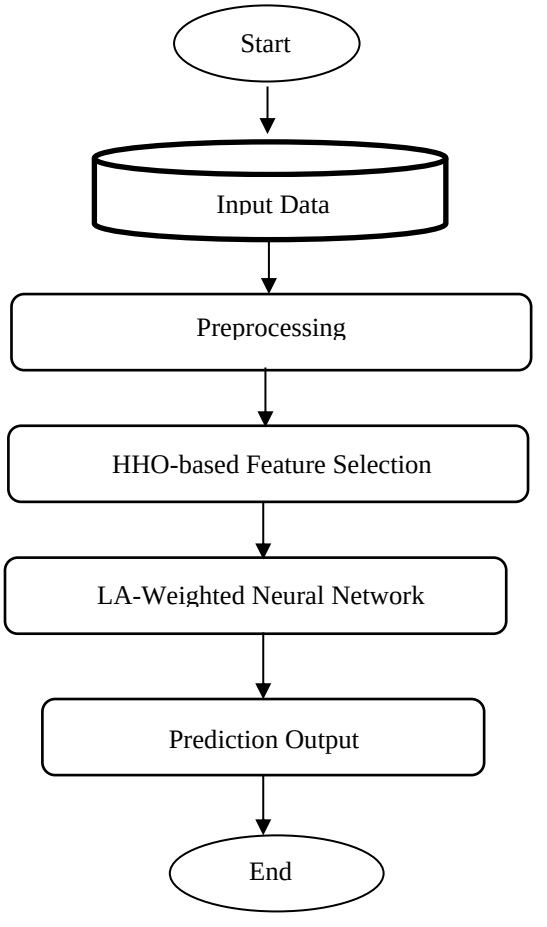


Fig. 1. Architecture of the proposed hybrid intrusion detection system.

4.1. Overview of the Architecture

The overall architecture of the proposed system includes the following major phases:

1.Input Dataset: The system is trained and evaluated using the KDDcup99 dataset.
2.Preprocessing: Includes normalization, categorical encoding, and dimensional transformation.
3.Feature Selection: Utilizes Harris Hawks Optimization (HHO) to select optimal subsets of features.
4.Classification: A supervised multi-layer neural network enhanced by learning automata.
5.Intrusion Detection: The final model classifies traffic as normal or malicious (DoS, Probe, R2L, U2R).

The goal is to improve accuracy, reduce computational complexity, and minimize false alarms in detecting network intrusions within smart city infrastructures.

6.1. Dataset Description and Preprocessing

The KDDcup99 dataset, derived from real network traffic simulations, contains 41 features per connection and labels for different attack types. This dataset is widely used for intrusion detection research, but is known to contain noise and class imbalance.

4.2.1. Label Consolidation

For simplicity and effectiveness, records are grouped into five categories:

- Normal
- Denial of Service (DoS)
- Probe
- Remote to Local (R2L)
- User to Root (U2R)

To mitigate severe class imbalance inherent in the original KDDCup99 dataset, oversampling techniques were applied to minority classes (R2L and U2R), leading to improved representativeness in test subsets.

4.2.2. Categorical Feature Encoding

Three categorical features (protocol_type, service, flag) are converted to numerical form using one-hot encoding:

- protocol_type: 3 values (TCP, UDP, ICMP)
- service: 70 distinct values
- flag: 11 values

Resulting in 122 features after transformation.

4.2.3. Normalization

Features with highly skewed distributions (e.g., src_bytes, dst_bytes, duration) are normalized to the range [0,1] using min-max scaling (Eq. 1):

$$x_i = \frac{x_i - Min}{Max - Min} \quad (1)$$

x_i is the i$th component of the feature vector. **Min** and **Max** refer to the minimum and maximum values of the ith column in the dataset, respectively. Normalization improves neural network performance by preventing dominance of large-scale features.

6.2. Feature Selection using Harris Hawks Optimization (HHO)

To reduce dimensionality and enhance classifier performance, Harris Hawks Optimization (HHO) is employed. HHO simulates the hunting behavior of Harris hawks with a balance between exploration and exploitation phases.

4.3.1. Initialization

Each hawk represents a binary vector indicating whether a feature is selected (1) or not (0). Initial parameters are:

- Population size: 20 hawks
- Maximum iterations: 100
- Feature encoding: 41-dimensional binary vectors (post one-hot expansion to 122 dimensions)

The Harris Hawks Optimization (HHO) algorithm is a continuous algorithm and generates continuous solutions. However, the proposed method for feature selection requires discrete results. Therefore, the solution values generated by HHO must be converted into binary values. To achieve this, the results are rounded to 0 or 1. An example of the initial population is shown in Table. 2.

Table 2. An example of the initial population

	Feature 1	Feature 2	Feature 3	Feature 4	.	.	.	Feature 39	Feature 40	Feature 41
Solution 1	1	0	0	1	.	.	.	1	1	1
Solution 2	0	0	1	1	.	.	.	0	1	0
Solution 3	1	0	1	0	.	.	.	0	0	0
	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .
Solution n-1	1	1	0	0	.	.	.	1	1	0
Solution n	1	1	0	0	.	.	.	1	1	0

4.3.2. Fitness Function

At this stage, the evaluation of candidate solutions is performed. After generating the initial population, the performance of each solution is assessed using a fitness function within the Harris Hawks Optimization (HHO) algorithm. The fitness of each hawk represents the quality of the feature subset it has selected. The primary objective is to identify optimal or near-optimal parameters that yield the most accurate solution.

The fitness function in the proposed method aims to minimize the classification error while maximizing the true positive rate and selecting the most informative feature subset. To achieve this, each candidate feature subset is evaluated based on three criteria: true positive rate (TPR), error rate, and the number of selected features. The fitness function used in the proposed method is defined by Eq. (2).

$$Fitness = R_{tp} + (1 - R_E) + (1 - \frac{N_F}{N}) \quad (2)$$

R_{tp} represents the true positive rate, calculated using Eq. (3); R_E denotes the error rate, computed based on Eq. (4); and N_F is the number of selected features, while N is the total number of features.

According to Equation (2), the fitness value increases when the true positive rate is higher, the error rate is lower, and the selected feature subset is smaller. This ensures that the proposed method not only improves classification performance but also promotes dimensionality reduction by selecting the most relevant features.

This formulation ensures the selected feature subset achieves high classification accuracy with minimal dimensionality.

The true positive rate (TPR) is calculated by dividing the number of true positives (TP) by the sum of true positives (TP) and false negatives (FN). This metric reflects the model's ability to correctly identify positive instances. Eq. (3) presents the formula used to compute the true positive rate.

$$R_{tp} = \frac{TP}{TP+FN} \quad (3)$$

TP denotes the number of true positive instances, and FN represents the number of false negatives.

The error rate refers to the proportion of instances that are incorrectly predicted by the classifier out of all samples. It reflects the overall misclassification and is calculated using Eq. (4).

$$R_E = \frac{FP + FN}{TP + FN + TN + FP} \quad (4)$$

Where TP denotes the number of true positives, FP represents the number of false positives, TN is the number of true negatives, and FN refers to the number of false negatives. These values form the foundation of the confusion matrix and are essential for evaluating the performance of the classification model.

4.3.3. Exploration and Exploitation Phases

- Exploration: Hawks perform random search via Lévy flight and stochastic positioning.
- Exploitation: Hawks adjust their positions based on the best solution (prey).

The transition between phases depends on the prey's escaping energy E, updated each iteration (Eq. 5):

$$E = 2E_0(1 - \frac{t}{T}) \quad (5)$$

E represents the escaping energy of the prey, T denotes the maximum number of iterations, and E_0 is the initial energy of the prey. The value of E_0 can vary randomly within the range (–1, +1) in each iteration. When E_0 decreases from 0 to –1, it implies that the prey (rabbit) is becoming physically weaker. Conversely, when E_0 increases from 0 to +1, it indicates that the prey is gaining strength.

The exploration phase in the Harris Hawks Optimization (HHO) algorithm involves the hawks moving randomly within the search space to discover better positions. This process enables the hawks to explore diverse regions and prevents stagnation in a specific area. The movement of hawks in this phase is random and guided by techniques such as Levy Flight, a stochastic movement pattern characterized by a combination of short and long jumps. This strategy allows the hawks to explore distant regions through long jumps and examine local areas through short jumps.

In the HHO algorithm, the Harris hawks represent candidate solutions, and the best candidate at each iteration is

considered the target prey or the near-optimal solution. Initially, hawks perch randomly in the environment, waiting for opportunities. Depending on a probability parameter q, two different strategies for locating the prey are applied:

- If q<0.5, the hawks perch based on the positions of other hawks and the rabbit, simulating coordinated waiting behavior.
- If q≥0.5, the hawks perch randomly on tall trees (i.e., random positions within the population's range).

This behavior is mathematically described by Eq. (6):

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 X(t) & q \geq 0.5 \\ \left(X_{rabbit}(t) - X_m(t)\right) - r_3\left(LB + r_4(UB - LB)\right) & q < 0.5 \end{cases} \quad (6)$$

Where $X(t+1)$ is the position vector of a hawk in iteration t+1, X(t) is its current position, $X_{rand}(t)$ is a randomly selected hawk's position, $X_{rabbit}(t)$ is the position of the rabbit (best solution), $X_m(t)$ is the mean position of the current population, r_1, r_2, r_3, r_4 and q are random numbers in the range (0,1), updated at each iteration, LB and UB denote the lower and upper bounds of the search space.

In the first rule (q≥0.5), new solutions are generated based on a random position influenced by both the previous position and the location of other hawks. In the second rule (q<0.5), the update rule incorporates a random location within the search boundaries, along with the difference between the current best position and the population mean. The variables r_3 and r_4 act as scaling factors to enhance the stochastic behavior of the hawks within the defined bounds.

4.3.4. Stopping Criteria

HHO terminates when either:

- The best fitness value does not improve significantly (≤ 0.001), or
- The algorithm reaches the maximum number of iterations.

The best feature subset obtained is passed to the classifier. On average, HHO reduced the number of input features from 122 to approximately 38, representing a 69% reduction in dimensionality without compromising classification performance.

6.3. Intrusion Classification using Neural Network with Learning Automata

The selected features are used to train a supervised multi-layer feedforward neural network. To enhance convergence, training is supported by a Learning Automaton (LA) which dynamically adjusts weight updates.

4.4.1. Neural Network Architecture

The classifier consists of:

- Input Layer: Corresponding to the number of selected features
- Hidden Layer: Nonlinear transformation using sigmoid activation
- Output Layer: Binary output (normal vs. attack) or multiclass if needed

Weight initialization is random, and the model is trained using backpropagation with error feedback. The training process is primarily based on standard backpropagation. However, to improve convergence and generalization, learning automata are integrated into the training loop, dynamically modifying weight updates based on performance feedback.

4.4.2. Learning Automata Integration

Learning automata improve training efficiency and prevent local minima:

- Monitors the error change after each epoch.
- If error decreases and remains below a threshold → reward current weights.
- If error increases or plateaus → penalize and replace weights.

Automata-based regulation enables:

- Faster convergence
- Prevention of overfitting
- Discarding ineffective weight configurations

4.4.3. Weight Update Mechanism

To optimize the neural network's weight assignment, a Learning Automaton (LA) is employed. Learning automata are decision-making systems modeled as abstract machines that operate in stochastic environments. These systems are capable of updating their decisions based on inputs received from the environment and corresponding probability values,

thereby improving overall system performance.

Each decision or action taken by the system is evaluated by the environment, which provides feedback to the learning automaton. Based on this feedback, the automaton updates its strategy and selects the most suitable action for the next step. A learning automaton can be represented as a triplet E={α, β, c}, where its components are defined as follows:

- Parameter α: This represents the set of inputs to the learning automaton, denoted as $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$. It corresponds to a set of actions or functions that may control a group of standard neurons, typically with a sigmoid activation function.
- Parameter β\beta: Represented as $\beta = \{\beta_1, \beta_2, \dots, \beta_n\}$, this parameter defines the set of outputs. It is a vector that describes the probability distribution over the actions selected by the Learning Control Unit (LCU).
- Parameter c: This feedback parameter, defined as c={0, 1}, represents the environment's response to the automaton's current action. A value of 0 indicates a penalty, while a value of 1 represents a reward.

Through iterative interaction with the environment, the learning automaton adjusts the probabilities of its actions to favor those that result in rewards, enabling it to effectively guide the neural network's learning process and improve performance.

a) Initial Data

In the proposed method, it is assumed that the neural network has N inputs. The neurons in the input layer act as a mapping function that transforms the N-dimensional input space into the required domains. Each neuron in the hidden layer represents a cluster of points related to the same group. The neurons in the output layer correspond to the number of classes.

During training, the network's weights and biases are adjusted so that the trained network responds to different inputs according to a specific rule.

It is assumed that there are K data samples, each having mm dimensions, where each dimension represents a feature in the proposed method. The dataset is categorized into two classes: intrusion and normal.

b) Weighting of Neural Network Layers

In each of these N functions, the network is initially trained with random weights, which are assumed to be identical for all functions at the start. The procedure is as follows: training data of the i^{th} class is fed into the first layer of the i^{th} function. The input data is denoted as x_i. These inputs are multiplied by their corresponding weights, which are initially assigned random values. The activation of each neuron in the hidden layer is then calculated according to Eq. (7), assuming v_{ij} as the weights and v_{oj} as the bias of the hidden layer.

$$Z_{in_j} = v_{oj} + \sum_{i=1}^{n} x_i v_{ij} \quad (7)$$

The activation function in Eq. (8) is used to compute the output of the hidden (intermediate) neuron.

$$Z_j = f\left(Z_{in_j}\right) \quad (8)$$
$$F1(x) = \frac{1}{1+\exp(-x)}$$

The values obtained in the previous step are multiplied by the weights of the next layer, and the network output is then computed. Similar to the previous step, the activation function defined in Eq. (9) is used to obtain the output.

$$Y_{in_k} = w_{oj} + \sum_{j=1}^{p} z_i w_{jk} \quad (9)$$
$$Y_k = f\left(Y_{in_k}\right)$$

c) Evaluation of Assigned Weights

The error backpropagation for each output unit $Y_k = 1,2, \dots, m$, is computed as the difference between the network output and the actual output, defined by Eq. (10):

$$\delta_k = (t_k - y_k) f'(Y_{in_k}) \quad (10)$$

At this stage, a learning automaton is employed to check whether the propagation error falls below a predefined

threshold. If the error is less than the threshold, this is considered a favorable condition for the network. Subsequently, it is checked whether the current error is lower than the previous error. If so, it indicates that the network is progressing toward reducing the propagation error, which is desirable. As a partial reward, the network is allowed to continue its training. However, if the propagation error exceeds the previous epoch's error, the initially selected random weights are penalized. If this increase persists over several iterations, the penalization involves replacing the current weights with new random weights.

On the other hand, if the initial condition is not met — meaning the propagation error remains above the threshold and does not reduce below it over several predefined iterations — the corresponding weights are discarded and replaced with new weights. Finally, if the error falls below the lower threshold, the algorithm terminates.

This learning automaton mechanism offers several advantages: it prevents the learning error from escalating uncontrollably, eliminates inappropriate weights, and if certain weights produce a learning error exceeding the threshold, those weights are removed. These actions result in faster network training and, more importantly, yield accurate and appropriate weights that enhance the overall performance.

For each hidden unit (neuron) $z_j, j = 1,2, \dots, p$, the sum of the input deltas is computed as shown in Eq. (11):

$$\delta_{in_j} = \sum_{k=1}^{m} \delta_k \, w_{jk} \tag{11}$$

Using the delta values derived from the activation function, the formulas for updating the input layer weights connected to the hidden layer and the biases of the input layer are obtained as Eq. (12):

$$\Delta \, v_{ij} = \alpha v_{ij} = \alpha \delta_j x_j \tag{12}$$
$$\Delta \, v_{oj} = \alpha \delta_j$$

At this stage, based on the previously obtained information, the weights are updated. For all output units, weights and biases are updated according to Eq. (13), and for all hidden units, according to Equation (14):

$$w_{jk}(new) = w_{jk}(old) + \Delta w_{jk} \tag{13}$$

$$v_{ij}(new) = v_{ij}(old) + \Delta v_{ij} \tag{14}$$

These steps are repeated for both classes in the proposed method. Through iteration of this process, the network is trained, producing a set of weight matrices corresponding to each class. For network testing and deployment, this procedure is repeated for each network.

7. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

7.1. Objective and Evaluation Criteria

The main objective of the evaluation is to assess the effectiveness and efficiency of the proposed hybrid intrusion detection system—based on Harris Hawks Optimization (HHO) for feature selection and neural network classification enhanced with learning automata—in detecting various types of cyber-attacks within IoT networks in smart cities.

The performance is assessed using standard evaluation metrics:

- Accuracy (overall correct classifications)
- True Positive Rate (TPR) (detection rate of actual attacks)
- False Positive Rate (FPR) (incorrect alarms on normal traffic)
- True Negative Rate (TNR) and False Negative Rate (FNR)

All metrics are calculated under both binary and multiclass classification scenarios using the KDDcup99 dataset, which includes five traffic types: Normal, DoS, Probe, R2L, U2R.

7.2. Experimental Setup

The proposed method was implemented using MATLAB 2017 and evaluated under simulated network conditions. All experiments were carried out on a system equipped with an Intel Core i7-10700 CPU running at 2.90 GHz and 16 GB of RAM. The software environment included Python 3.9, TensorFlow 2.x, and Scikit-learn libraries, operating on Ubuntu 22.04 LTS. Model performance was assessed using a 10-fold cross-validation approach to ensure robustness and reliability of the results. The simulation parameters are presented in Table 3.

Table 3. Simulation parameters

Parameter	Value
Simulation time	1500 seconds
Network size	100 × 100 m²
Number of nodes	100
HHO population	20 hawks
Maximum HHO iterations	100
Input features (after encoding)	122 features

7.3. Dataset Description

We use a preprocessed subset of the KDDCup99 dataset containing 494,021 labeled records. Each record consists of 41 features, categorized into five classes. The distribution of these classes along with the number of records per class is presented in Table 4.

Table 4. Class distribution and number of labeled records in the preprocessed KDDCup99 dataset used for model evaluation

Class Type	Description	Percentage (Train)	Percentage (Test)
Normal	Benign traffic	19.69%	19.48%
DoS	Denial of Service	79.24%	73.90%
Probe	Network reconnaissance	0.83%	1.34%
R2L	Remote access to local user	0.23%	5.21%
U2R	User privilege escalation	0.01%	0.07%

7.4. Performance Analysis

To comprehensively assess the performance of the proposed intrusion detection system, a series of evaluation charts was generated, each corresponding to a specific performance metric.

Fig. 2 illustrates the True Positive Rate (TPR) achieved by the proposed method for different categories of network attacks. The True Positive Rate (TPR) presents the system's ability to correctly detect various types of attacks. The highest TPR was observed for DoS attacks, reaching 99.21%, which is expected due to the high frequency and distinguishable patterns of such attacks. Probe attacks also achieved a strong TPR of 95.44%, reflecting the model's effectiveness in identifying scanning-based threats. More challenging attacks, such as R2L and U2R, which are typically underrepresented in the dataset and more difficult to detect, achieved TPRs of 93.86% and 91.13%, respectively. These results indicate that the proposed method performs well not only on frequent attacks but also on rare and subtle ones, thanks to the robust feature selection and adaptive classification components.

Fig. 3 illustrates the False Positive Rate (FPR) achieved by the proposed method for different categories of network attacks. The FPR illustrates the proportion of normal traffic mistakenly classified as malicious. The system demonstrated a low average FPR of 2.80%, with the lowest FPR recorded for DoS at 1.88%. These results are critical, as a high false alarm rate can burden security personnel and compromise the usability of a detection system. The low FPR across all classes highlights the model's capability to distinguish legitimate traffic from actual threats with high precision, ensuring operational efficiency in real-world deployments.

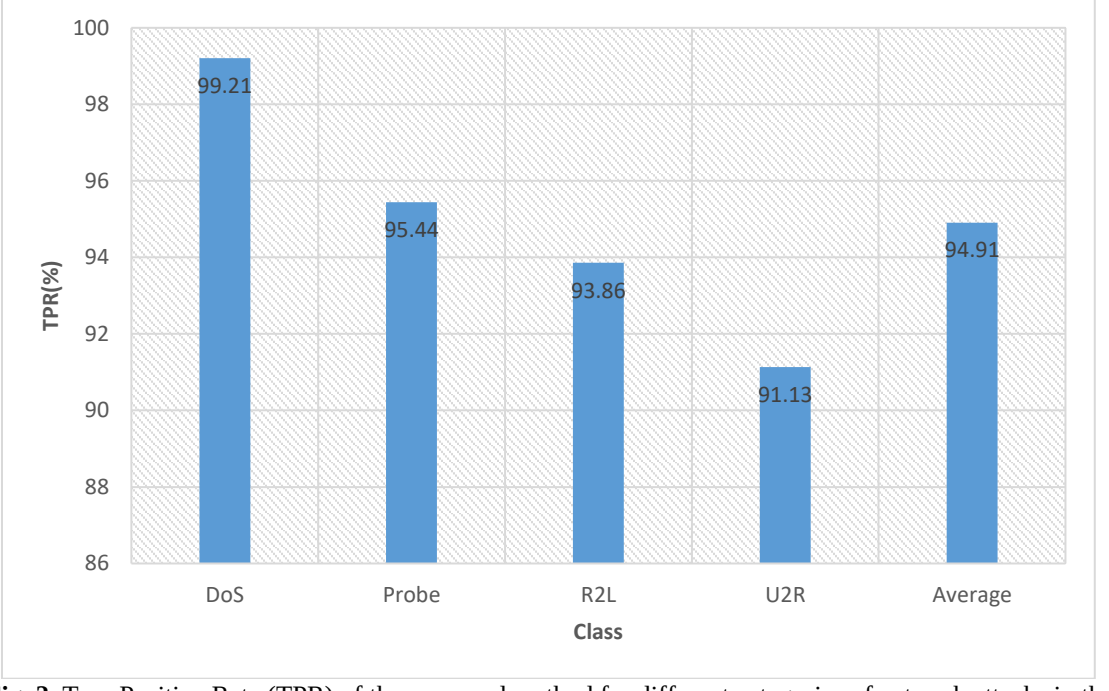


Fig. 2. True Positive Rate (TPR) of the proposed method for different categories of network attacks in the KDDCup99 dataset.

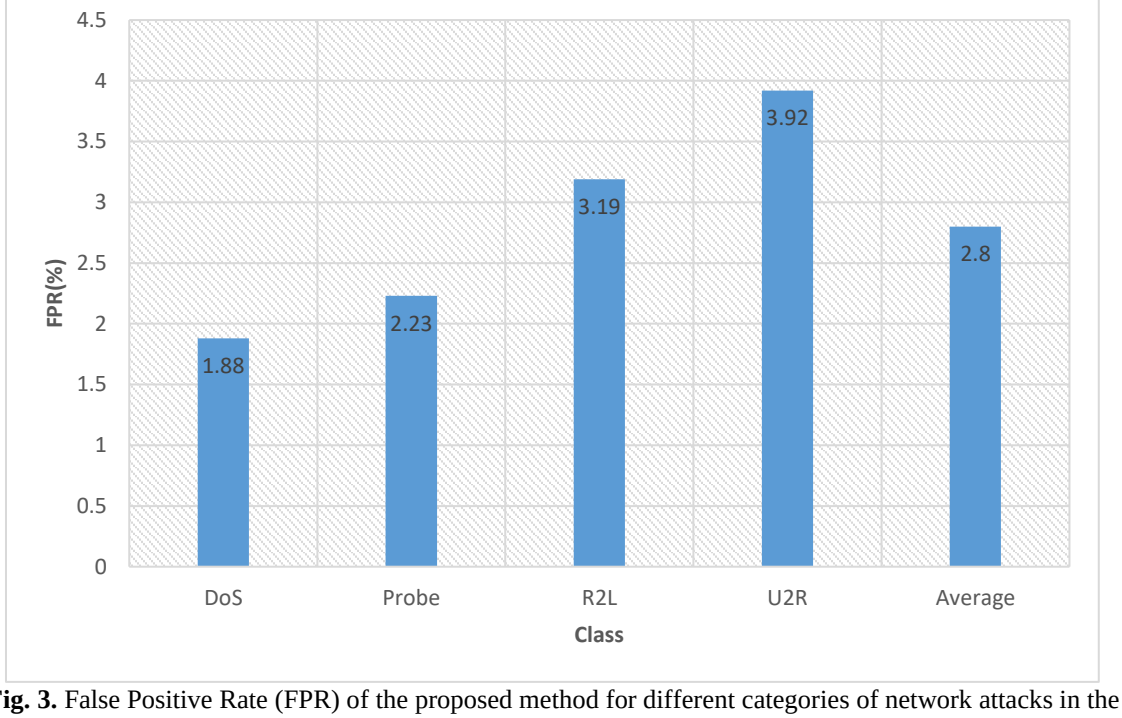


Fig. 3. False Positive Rate (FPR) of the proposed method for different categories of network attacks in the KDDCup99 dataset.

Fig. 4 illustrates the True Negative Rate (TNR) achieved by the proposed method for different categories of network attacks. TNR further confirms this capability by showing how accurately the system identifies normal, non-malicious traffic. The overall TNR reached 97.20%, with most classes (including Probe and R2L) exceeding 96%. Such high TNR values suggest that the model effectively minimizes false alarms and maintains reliability in classifying benign behavior—an essential requirement for smart city infrastructures where service continuity is critical.

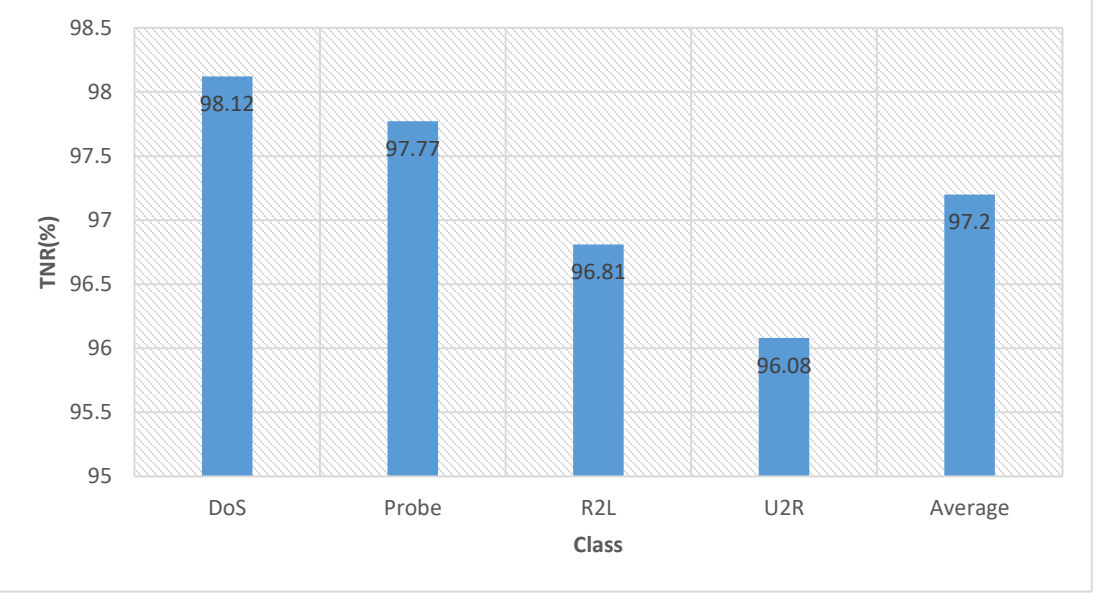


Fig. 4. True Negative Rate (TNR) of the proposed method for different categories of network attacks in the KDDCup99 dataset.

Fig. 5 illustrates the False Negative Rate (FNR) achieved by the proposed method for different categories of network attacks. On the other hand, the FNR quantifies the percentage of attacks that were incorrectly labeled as normal. The average FNR across all classes was 5.09%, with the lowest for DoS (0.79%) and the highest for U2R (8.87%). These values remain within acceptable thresholds for intelligent detection systems, especially given the difficulty of detecting rare attack types. A low FNR implies that the system is unlikely to miss actual threats, further enhancing its robustness and trustworthiness. The higher FNR observed for U2R (8.87%) is consistent with its rarity in the dataset. Despite this, the model outperforms many baseline approaches by achieving reliable detection of such low-frequency attack types.

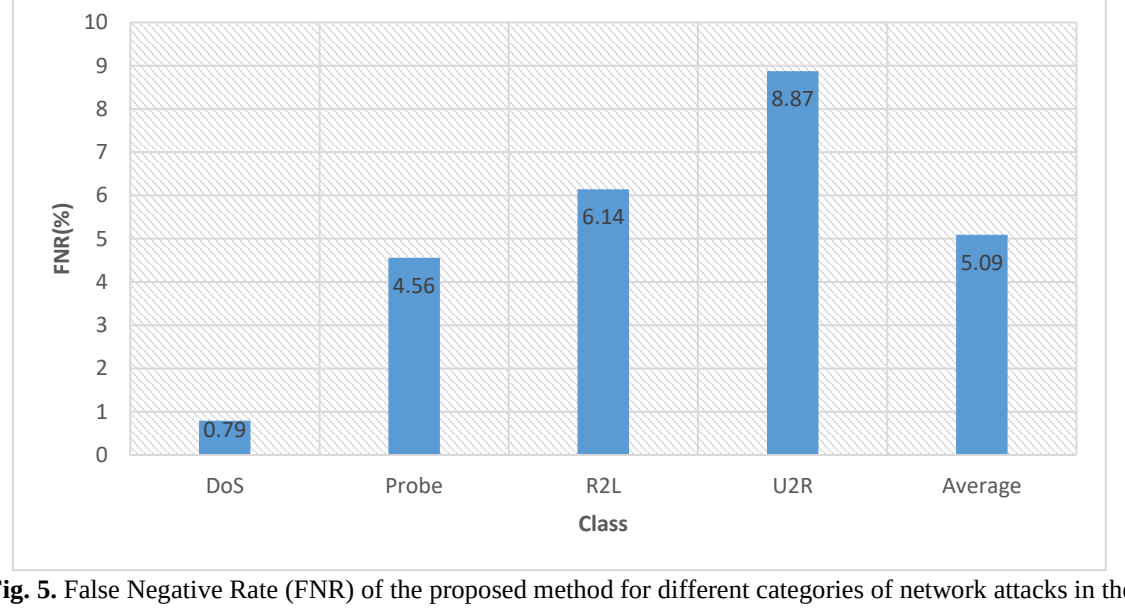


Fig. 5. False Negative Rate (FNR) of the proposed method for different categories of network attacks in the KDDCup99 dataset.

Fig. 6 illustrates the accuracy of the proposed method across different attack categories in the KDDCup99 dataset. The Accuracy shows the overall correctness of the system in classifying both attack and normal instances. The model achieved its highest accuracy for DoS attacks (98.67%) and its lowest for U2R (94.03%), with an average accuracy of 96.53% across all classes. This consistently high performance, even on difficult-to-detect classes, demonstrates the

effectiveness of integrating Harris Hawks Optimization for feature selection with learning automata-based neural network training, resulting in a well-generalized and stable classifier.

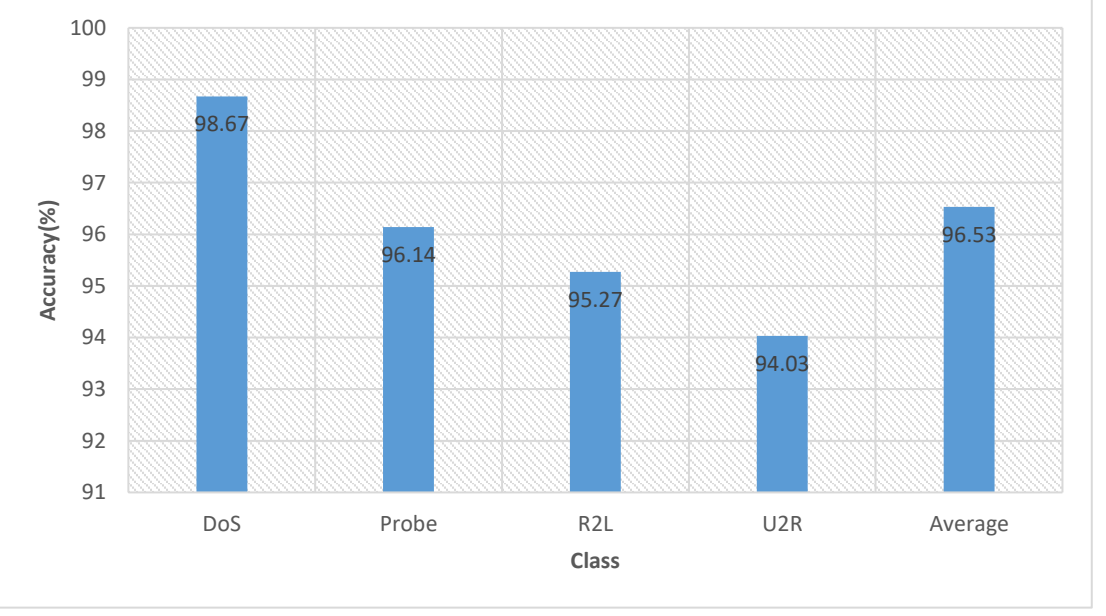


Fig. 6. Accuracy of the proposed method in detecting different categories of attacks using the KDDCup99 dataset.

Fig. 7 presents a comparative analysis of the proposed method against two recent approaches. In addition to individual class performance, a comparative chart was generated to benchmark the proposed model against two recent methods [13], [14]. The proposed method achieved an overall accuracy of 96.53%, compared to 94.32% and 95.18% for the respective reference methods. This performance gain of +2.2% and +1.35%, respectively, underscores the superiority of the proposed hybrid approach in both detection accuracy and system reliability. The improvement is mainly attributed to the dynamic and adaptive nature of the combined HHO and learning automata strategies, which enhance both training efficiency and classification accuracy.

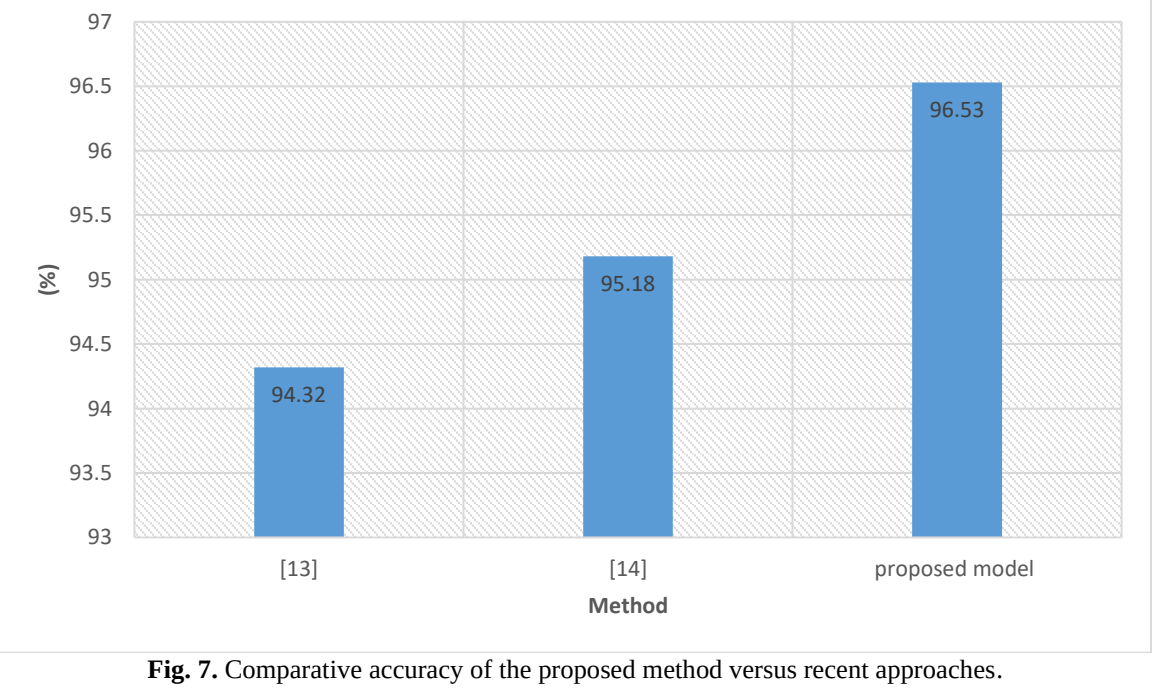


Fig. 7. Comparative accuracy of the proposed method versus recent approaches.

8. DISCUSSION

The proposed HHO + LA-NN model shows notable improvements in precision and recall across all classes, particularly for underrepresented attack types. The use of HHO reduces input dimensionality while preserving critical patterns, and Learning Automata improves the adaptiveness of the classifier. These combined mechanisms contribute to higher generalization, faster convergence, and lower false positive rates, making the system highly suitable for real-world IoT.

One limitation of our approach is the computational overhead introduced by the Harris Hawk Optimization (HHO) during feature selection, particularly for large-scale IoT datasets. Additionally, the Learning Automata mechanism requires fine-tuning of reward and penalty parameters, which may not generalize optimally across all IoT scenarios. Finally, our current evaluation is based on the KDDCup99 dataset, which, while widely used, may not fully capture the complexities of modern IoT traffic. Future work will address these limitations by experimenting with more realistic datasets and investigating online feature selection mechanisms.

For future work, we plan to extend our hybrid IDS in several directions. First, we aim to validate the system using more recent and realistic IoT datasets such as TON_IoT, CIC-IDS2017, and IoTID20, to better assess its applicability to real-world deployments. Second, we intend to develop an online variant of the model capable of real-time learning and adaptation to evolving attack patterns. Third, the proposed LA-NN architecture will be evaluated on embedded IoT hardware platforms (e.g., Raspberry Pi, Nvidia Jetson Nano) to analyze its performance under resource-constrained conditions. Finally, integration with edge and fog computing infrastructures will be explored to enable distributed and low-latency intrusion detection for smart city applications.

REFERENCES

[1] H. Allioui, and Y. Mourdi, "**Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey**," *Sensors*, Vol. 23, No. 19, pp. 8015, 2023.

[2] R., Ahmad, and I., Alsmadi, "**Machine learning approaches to IoT security: A systematic literature review,**" *Internet of Things*, Vol. 14, pp. 100365, 2021.

[3] W. H., Hassan, "**Current research on Internet of Things (IoT) security: A survey**," *Computer networks*, Vol. 148, pp. 283-294, 2019.

[4] K., He, D. D., Kim, M. R., and Asghar, "**Adversarial machine learning for network intrusion detection systems: A comprehensive survey**," *IEEE Communications Surveys & Tutorials*, Vol. 25, No. 1, pp. 538-566, 2023.

[5] N., Islam, F., Farhin, I., Sultana, M. S., Kaiser, M. S., Rahman, M. Mahmud, and G. H., Cho, "**Towards machine learning based intrusion detection in IoT networks**," *Comput. Mater*, Contin, Vol. 69, No. 2, pp. 1801-1821, 2021.

[6] M. A., Alsoufi, S., Razak, M. M., Siraj, I., Nafea, F. A., Ghaleb, F., Saeed, and M., Nasser, "**Anomaly-based intrusion detection systems in iot using deep learning: A systematic literature review**," *Applied sciences*, Vol. 11, No. 18, pp. 8383, 2021.

[7] A. N., Alsheavi, A., Hawbani, X., Wang, W., Othman, L., Zhao, Z., Liu, and M. A., Al-qaness, "**IoT Authentication Protocols: Classification, Trend and Opportunities**," *IEEE Transactions on Sustainable Computing*, 2024.

[8] A. A., Laghari, K., Wu, R. A., Laghari, M., Ali, and A. A., Khan, "**A review and state of art of Internet of Things (IoT)**," *Archives of Computational Methods in Engineering*, pp. 1-19, 2021.

[9] M. A., Khan, M. A., Khan, S. U., Jan, J., Ahmad, S. S., Jamal, A. A., Shah, and W. J., Buchanan, "A deep learning-based intrusion detection system for MQTT enabled IoT," *Sensors*, Vol. 21, No. 21, pp. 7016, 2021.

[10] R., Zhao, G., Gui, Z., Xue, J., Yin, T., Ohtsuki, B., Adebisi, and H., Gacanin, "**A novel intrusion detection method based on lightweight neural network for internet of things**," *IEEE Internet of Things Journal*, Vol. 9, No. 12, pp. 9960-9972, 2021.

[11] O. A., Wahab, "**Intrusion detection in the iot under data and concept drifts: Online deep learning approach**," *IEEE Internet of Things Journal*, Vol. 9, No. 20, pp. 19706-19716, 2022.

[12] M., Vishwakarma, and N., Kesswani, "**DIDS: A Deep Neural Network based real-time Intrusion detection system for IoT**," *Decision Analytics Journal*, Vol. 5, pp. 100142, 2022.

[13] S., Chatterjee, and M. K., Hanawal, "**Federated learning for intrusion detection in IoT security: a hybrid ensemble approach**," *International Journal of Internet of Things and Cyber-Assurance*, Vol. 2, No. 1, pp. 62-86, 2022.

[14] D. M., Ngo, D., Lightbody, A., Temko, C., Pham-Quoc, N. T., Tran, C. C., Murphy, and E., Popovici, E., "**HH-NIDS: Heterogeneous Hardware-Based Network Intrusion Detection Framework for IoT Security**," *Future Internet*, Vol. 15, No. 1, pp. 9, 2023.

[15] A., Awajan, "**A novel deep learning-based intrusion detection system for IOT networks**," *Computers*, Vol. 12, No. 2, pp. 34, 2023.

[16] Z., Wang, J., Li, S., Yang, X., Luo, D., Li, and S., Mahmoodi, "A lightweight IoT intrusion detection model based on improved BERT-of-Theseus," Expert Systems with Applications, Vol. 238, pp. 122045, 2024.

[17] A., Qaddos, M. U., Yaseen, A. S., Al-Shamayleh, M., Imran, A., Akhunzada, and S. Z., Alharthi, "**A novel intrusion detection framework for optimizing IoT security**," *Scientific Reports*, Vol. 14, No. 1, pp. 21789, 2024.

[18] N., Wang, S., Shi, Y., Chen, W., Lou, and Y. T., Hou, "**FeCo: Boosting intrusion detection capability in IoT networks via contrastive learning**," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[19] L., Lin, Q., Zhong, J., Qiu, and Z., Liang, "**E-GRACL: an IoT Intrusion Detection System Based on Graph Neural Networks**," *The Journal of Supercomputing*, Vol. 81, No. 1, pp. 42, 2025.
[20] M. A., Rahman, A. T., Asyhari, L. S., Leong, G. B., Satrya, M. H., Tao, and M. F., Zolkipli, "**Scalable machine learning-based intrusion detection system for IoT-enabled smart cities**," *Sustainable Cities and Society*, Vol. 61, pp. 102324, 2020.
[21] A., Alazab, M., Hobbs, J., Abawajy, and M., Alazab, "**Using feature selection for intrusion detection system**," *In 2012 international symposium on communications and information technologies (ISCIT*), IEEE, pp. 296-301, 2012.
[22] M., Tavallaee, E., Bagheri, W., Lu, and A. A., Ghorbani, "**A detailed analysis of the KDD CUP 99 data set**," *In 2009 IEEE symposium on computational intelligence for security and defense applications*, IEEE, pp. 1-6, 2009.
[23] A. A., Heidari, S., Mirjalili, H., Faris, I., Aljarah, M., Mafarja, and H., Chen, "**Harris hawks' optimization: Algorithm and applications**," *Future generation computer systems*, Vol. 97, pp. 849-872, 2019.
[24] K. T., Yang, "Artificial Neural Networks (ANNs): a New Paradigm for Thermal Science and Engineering", 2008.
[25] Thathachar, M. A., and P. S., Sastry, "Varieties of learning automata: an overview," *IEEE Transactions on Systems, Man, and Cybernetics*, Part B (Cybernetics), Vol. 32, No. 6, pp. 711-722, 2002.

Critical Review on Deep Learning-Based Breast Cancer Detection and Segmentation: Challenges, Gaps, and Future Directions

Hassan Mahichi[1], Vahid Ghods[2], Mohammad Karim Sohrabi[3], Arash Sabbaghi[4]
Department of Electrical and Computer Engineering, Se.C., Islamic Azad University, Semnan, Iran.
Email: V.ghods@iau.ac.ir (Corresponding author)

ABSTRACT:
Breast cancer remains a leading cause of cancer-related mortality among women worldwide, where early and accurate detection is vital for effective intervention and prognosis. Deep learning has emerged as a cornerstone in the automated analysis of breast cancer imaging, offering substantial improvements in tumor detection, segmentation, and classification across modalities such as mammography, ultrasound, and MRI. Despite notable progress, clinical integration remains constrained by challenges including limited dataset availability, suboptimal generalization, lack of interpretability, high computational complexity, and insufficient multi-task learning optimization. This critical review synthesizes findings from 70 peer-reviewed studies published between 2018 and 2025, encompassing convolutional neural networks, U-Net derivatives, Vision Transformers, instance segmentation models, and hybrid frameworks. Comparative evaluation highlights architectural strengths, modality-specific adaptations, and diagnostic performance metrics. Emphasis is placed on the comparative analysis of single-task versus multi-task frameworks, the integration of handcrafted features, transfer learning, and optimization strategies to improve model generalizability and robustness. Key limitations are identified in areas such as cross-domain robustness, real-time applicability, interpretability, and standardized benchmarking. Emerging solutions are examined, including self-supervised and semi-supervised learning strategies, lightweight and explainable architectures, adaptive loss balancing for MTL, cross-modal fusion techniques, and unified end-to-end pipelines.

KEYWORDS: Breast cancer detection, Segmentation, Medical Image Analysis, Deep learning, Convolutional neural network.

1. INTRODUCTION

Breast cancer remains the most commonly diagnosed cancer and the leading cause of cancer-related death among women worldwide, accounting for a substantial burden on global healthcare systems [1]. Early detection and accurate localization of breast tumors play a crucial role in improving prognosis, enabling timely intervention, and guiding therapeutic strategies. Medical imaging modalities—such as mammography, ultrasound (US), and magnetic resonance imaging (MRI)—are integral to breast cancer screening and diagnosis[2], [3]. However, the manual interpretation of these images by radiologists is subject to inter-observer variability, fatigue, and diagnostic error, especially in complex cases or dense breast tissues [4], [5]. In recent years, deep learning (DL) has emerged as a transformative technology in medical image analysis, particularly for tasks such as tumor detection, segmentation, classification, and localization. Architectures such as CNNs, transformers, and multi-task learning (MTL) frameworks have achieved remarkable success in various computer vision domains and are increasingly applied to breast imaging due to their ability to learn hierarchical features from large datasets [6]-[8]. Despite significant advancements, several critical challenges remain. These include the scarcity of annotated datasets, poor cross-dataset generalization, lack of model interpretability,

Paper type: Research paper
https://doi.org/xxx
Received: 25 February 2025, Revised: 5 April 2025, Accepted: 2 May 2025, Published: 1 June 2025
How to cite this paper: H. Mahichi, V. Ghods, M. K. Sohrabi, A. Sabbaghi, **"Critical Review on Deep Learning-Based Breast Cancer Detection and Segmentation: Challenges, Gaps, and Future Directions"**, *Majlesi Journal of Telecommunication Devices*, Vol. 14, No. 2, pp. 105-129, 2025.

and computational demands that limit real-time clinical deployment. Furthermore, many models are designed in isolation for either detection or segmentation, lacking the integration required for practical, end-to-end diagnosis workflows [6]-[9]. Moreover, while transformer-based models and hybrid architectures improve contextual understanding, their real-world applicability is often constrained by their complexity and high resource requirements [10], [11].

While several reviews have surveyed deep learning techniques in medical imaging, few have delivered a critical, task-specific, and end-to-end analysis tailored for breast cancer detection and segmentation across modalities. This critical review aims to systematically evaluate the current landscape of deep learning-based breast cancer detection and segmentation models. The primary objectives are to:

- Analyze recent methods across different imaging modalities and model types;
- Identify technical gaps, methodological inconsistencies, and clinical challenges;
- Highlight the strengths and weaknesses of existing detection, segmentation, and unified models;
- And propose future research directions that address outstanding issues and move toward more robust, interpretable, and deployable solutions.

Through an in-depth synthesis of over 70 peer-reviewed studies from 2018 to 2025, this review provides a comprehensive knowledge base for researchers, clinicians, and developers aiming to enhance AI-driven breast cancer diagnosis systems. By framing the discussion around methodological innovation, clinical utility, and scalability, this work contributes a forward-looking perspective on how deep learning can evolve into clinically reliable tools for breast cancer care.

2. CONCEPTUAL AND THEORETICAL FRAMEWORK

Deep learning-based breast cancer detection and segmentation are rooted in the intersection of medical imaging, artificial intelligence (AI), and pattern recognition. The theoretical framework for this research domain encompasses several key components: the nature of breast imaging modalities, the principles of deep neural networks, the role of encoder-decoder architectures for segmentation, attention mechanisms, transformers, and the emerging trend of multi-task learning. This section provides a detailed overview of these concepts and explains how they form the scientific foundation of modern breast cancer analysis systems.

2.1. Imaging Modalities in Breast Cancer Diagnosis

Breast cancer diagnosis relies on various imaging techniques, each providing distinct structural and textural information. The three most widely used modalities in clinical and research settings include [4]:

Mammography: This is the gold standard for breast cancer screening. Mammograms are grayscale 2D X-ray images that reveal calcifications, asymmetries, and masses. Their high resolution is beneficial, but the modality often suffers from reduced sensitivity in dense breast tissues [12].

Ultrasound (US): Ultrasound imaging is non-invasive, inexpensive, and safe. It is frequently used as an adjunct to mammography, particularly for younger women or patients with dense breast tissues. However, ultrasound images are often noisy, operator-dependent, and may exhibit lower spatial resolution, making automated analysis more challenging [13].

Magnetic Resonance Imaging (MRI): Breast MRI offers high sensitivity and 3D anatomical detail. It is especially useful for high-risk patients and for evaluating tumor extent. Despite its benefits, it is expensive, time-consuming, and requires expert interpretation [14]

Each modality introduces unique challenges in image interpretation and thus influences the architecture and preprocessing strategies of deep learning models (Fig. 1).

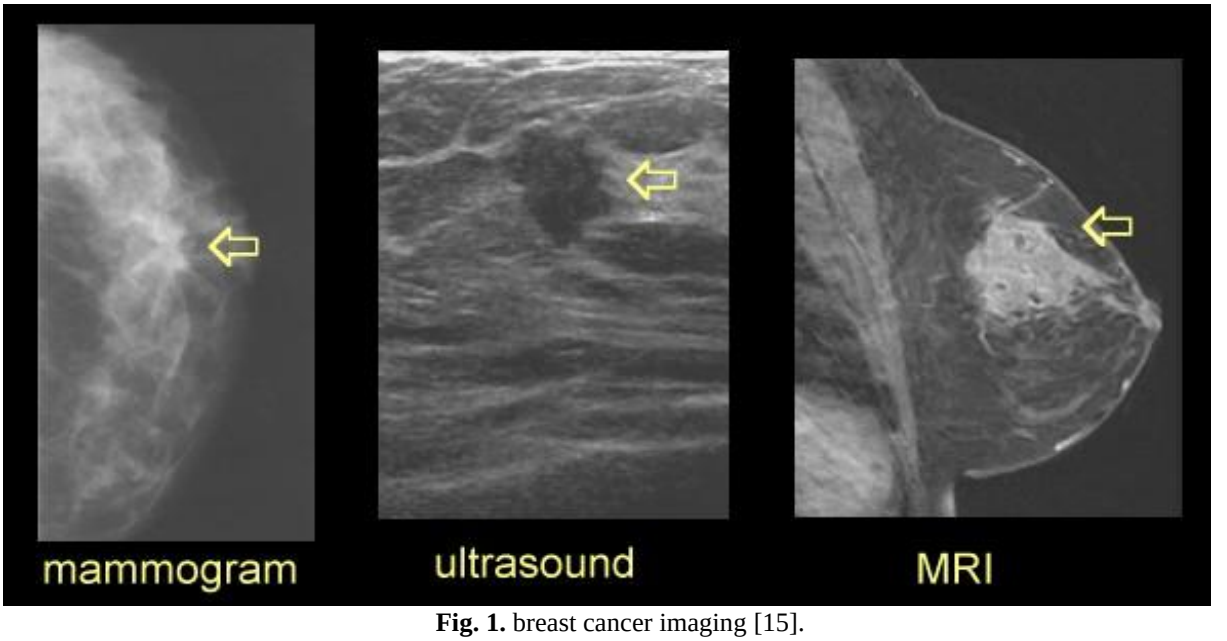


Fig. 1. breast cancer imaging [15].

2.2. Fundamentals of Deep Learning in Medical Imaging

Deep learning is a subset of machine learning that automatically learns hierarchical representations from data. It has revolutionized computer vision tasks and shown exceptional promise in the medical imaging domain [16]. Convolutional Neural Networks (CNNs) are the most widely used architecture in breast cancer diagnosis tasks due to their ability to capture spatial features, detect edges, and model patterns in images with varying levels of abstraction [17]. CNNs consist of layers such as convolution, pooling, and fully connected layers [18], [19].

Deeper CNNs like ResNet (with residual skip connections) [19], DenseNet [20] (which encourages feature reuse), and InceptionNet [21] (multi-scale feature extraction) have improved image classification performance by addressing issues like vanishing gradients and overfitting [22]. These architectures are commonly used for tumor detection and classification tasks and have been further fine-tuned using transfer learning approaches in medical contexts [23].

However, standard CNNs are typically limited to image-level predictions. For tasks such as delineating tumors pixel by pixel, specialized architectures like encoder-decoder networks are required.

2.3. Encoder-Decoder Architectures for Tumor Segmentation

Segmentation, which involves delineating tumor boundaries at the pixel level, is a critical task in medical imaging. U-Net [24], introduced by Ronneberger et al. [25], is the foundational architecture for biomedical segmentation (Fig. 2). It follows an encoder-decoder scheme, where the encoder progressively downsamples the input to extract features, and the decoder upsamples these features to the original resolution, enabling fine-grained segmentation [25]. Key enhancements to the U-Net architecture (Fig. 2) have been proposed to address limitations such as class imbalance, low contrast, and small lesion size:

- **U-Net**$^{++}$ introduces dense skip pathways to bridge semantic gaps between encoder and decoder features [26].
- **Attention U-Net** incorporates attention gates that allow the model to focus on relevant tumor regions while suppressing irrelevant background [27].
- **ResUNet**, **V-Net**, and **DenseUNet** adapt popular classification backbones into segmentation frameworks to enhance feature richness [28].

These models have shown superior performance on various breast imaging datasets; however, their generalization across datasets and imaging conditions remains a challenge.

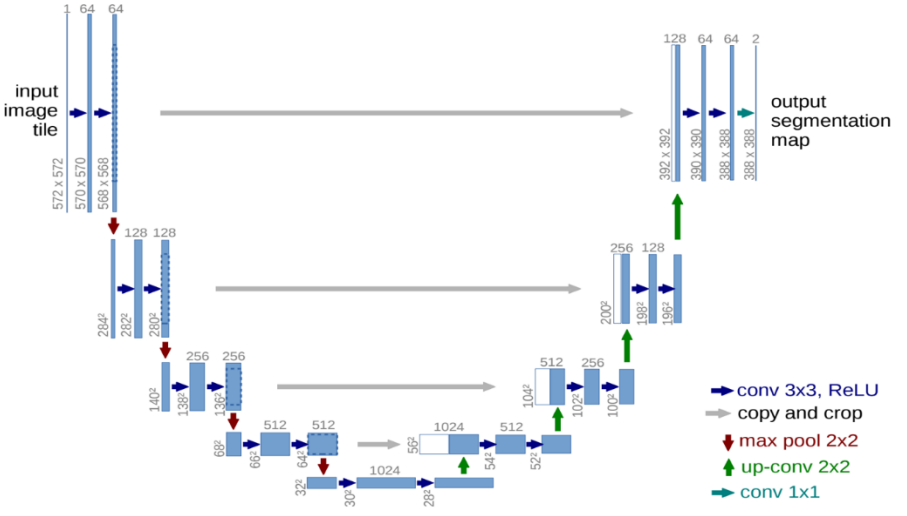


Fig. 2. U-net architecture [29].

2.4. Role of Attention Mechanisms

Inspired by human visual attention, attention mechanisms enable neural networks to focus selectively on important regions of an image[30]. In breast cancer segmentation, attention modules help improve model robustness by suppressing irrelevant background and emphasizing tumor boundaries. The most common types include channel attention, spatial attention, and self-attention mechanisms [30]. Attention U-Net (Fig. 3) and its variants improve segmentation quality, particularly when tumors are small or embedded in complex tissue structures. Moreover, coordinate attention and non-local attention blocks have also been applied to improve feature representation in noisy and low-contrast ultrasound and mammogram images [14].

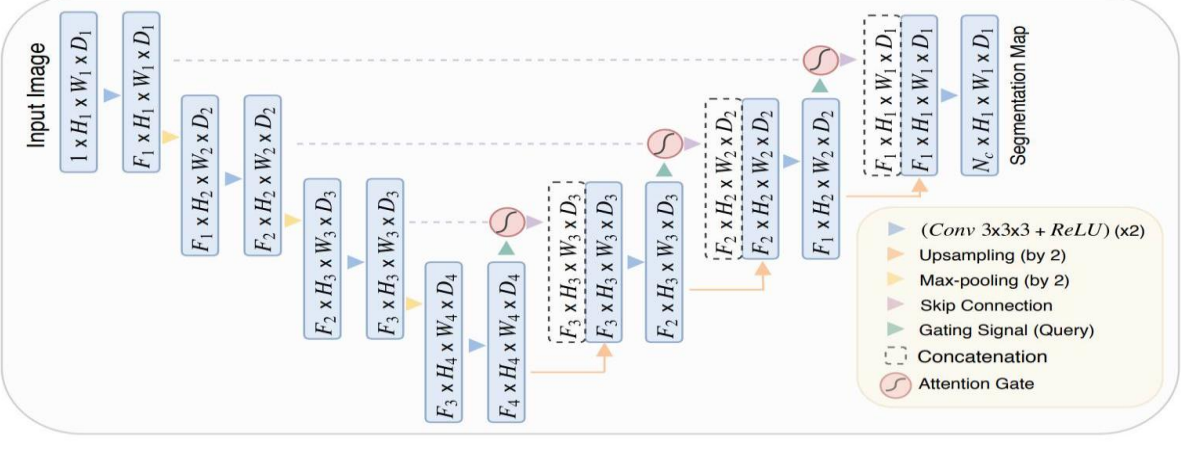


Fig. 3. Attention U-Net [14].

2.5. Transformers in Medical Imaging

While CNNs are effective for local feature extraction, they often struggle with modeling long-range dependencies. Vision Transformers (ViTs) [31], introduced in 2020, overcome this limitation by treating images as sequences of patches and applying self-attention across the entire image (Fig. 4). Although initially developed for natural images, transformers have been successfully adapted for medical tasks such as tumor segmentation and detection [31].

TransUNet, one of the first hybrid models, combines a transformer-based encoder with a U-Net decoder, achieving state-of-the-art results in several biomedical segmentation benchmarks [32]. Transformers provide better global context modeling, which is especially useful in breast cancer detection where the tumor may be diffuse, poorly defined, or appear in multiple regions [31], [32].

However, transformer-based models typically require large datasets and significant computational resources. This poses limitations in medical applications where annotated datasets are often limited. To address this, recent studies have explored lightweight transformer variants and semi-supervised training techniques to make these models more practical [33], (Fig. 4).

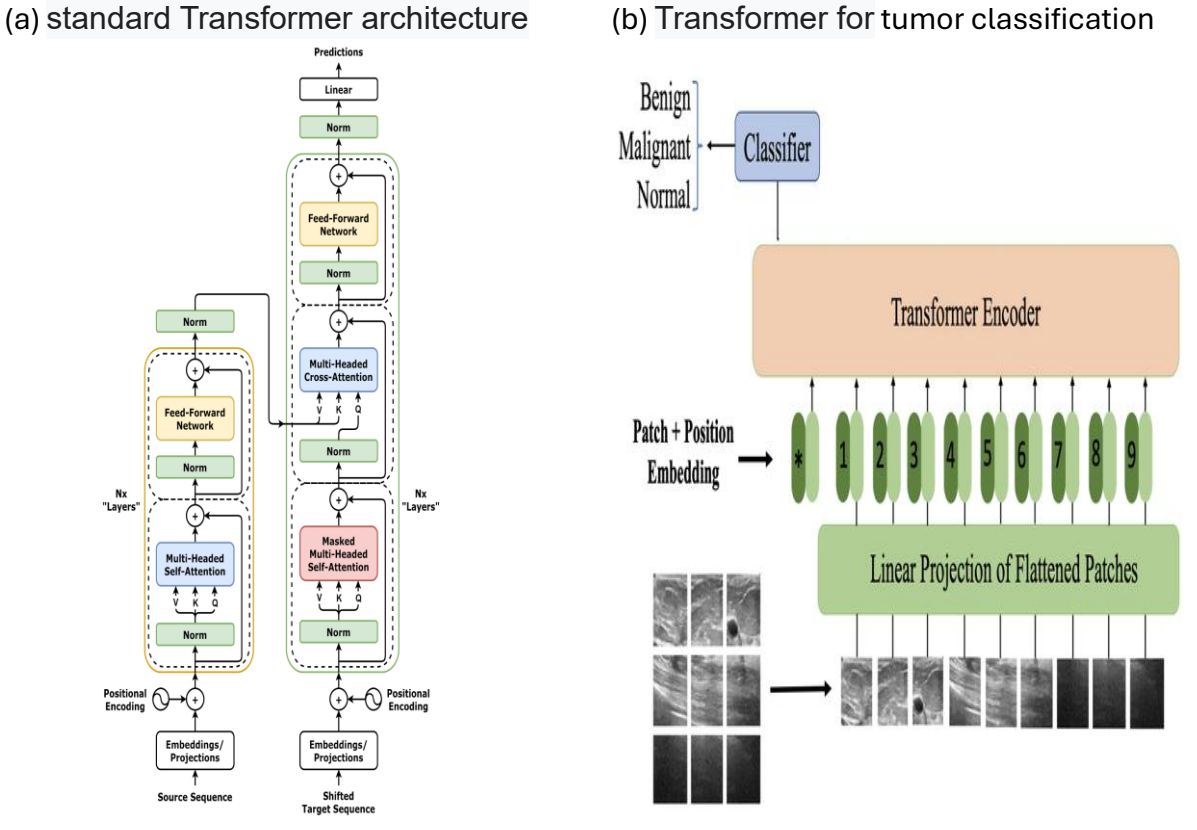


Fig. 4. Transformers [31]: (a): simple standard Transformer architecture, and (b): example of using transformer encoder for breast tumor classification.

2.6. Multi-Task Learning Paradigm

Traditional models perform either detection or segmentation as independent tasks. However, breast cancer diagnosis in clinical practice often involves both tasks simultaneously. Multi-task learning (MTL) addresses this by training a single model to optimize multiple objectives (e.g., classification and segmentation) concurrently, encouraging shared feature learning and improving model efficiency [8]. Shared encoder-decoder structures are commonly used in MTL, with task-specific heads for segmentation and classification. While MTL (Fig. 5) can improve performance and reduce training time, balancing the learning of each task is a challenge. Improper loss weighting or architectural imbalance can lead to dominance of one task over the other [34].

Recent works have introduced dynamic loss balancing, task-specific attention modules, and cross-task consistency learning to enhance MTL performance in breast cancer imaging [17]. These frameworks align well with real-world clinical workflows and represent an important trend in future model development [31], [34].

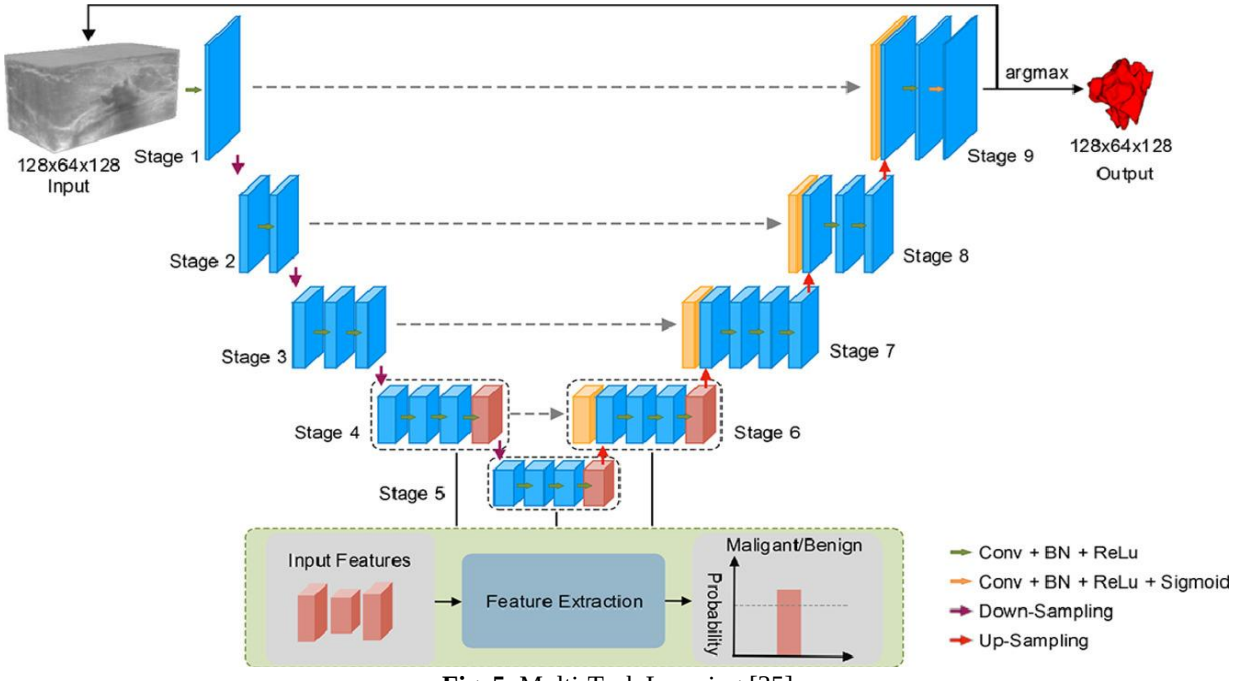


Fig. 5. Multi-Task Learning [35].

3. LITERATURE SELECTION APPROACH

To ensure the integrity, relevance, and scientific rigor of this critical review, a systematic and reproducible literature selection strategy was adopted. The focus was to identify high-quality, peer-reviewed research works that employed deep learning approaches for breast cancer detection, classification, and segmentation across different imaging modalities, including mammography, ultrasound, and MRI. The literature selection methodology adhered to widely accepted guidelines such as PRISMA and was guided by a multi-phase screening process encompassing database querying, inclusion/exclusion filtering, and quality assessment.

Initially, a comprehensive search was conducted across five major academic databases—IEEE Xplore, Scopus, PubMed, Web of Science, and ScienceDirect—owing to their extensive coverage of medical imaging, computer vision, and AI research. The search queries were designed using a combination of targeted keywords and Boolean operators. Typical queries included phrases such as: (“breast cancer” OR “mammography” OR “ultrasound” OR “MRI”) AND (“deep learning” OR “CNN” OR “transformer” OR “multi-task learning” OR “semantic segmentation”) AND (“detection” OR “classification” OR “segmentation” OR “localization”). To enhance the specificity of the search, filters were applied to restrict the publication period from January 2015 to July 2025, including only peer-reviewed articles, and exclude non-English publications and preprints.

The literature selection process proceeded in multiple stages. In the first stage, all retrieved articles were screened based on titles and abstracts. Articles that did not clearly indicate the use of deep learning for breast cancer analysis or that focused exclusively on traditional machine learning, radiomics without DL components, or non-imaging modalities were excluded. In the second stage, the remaining papers were subjected to full-text review to determine methodological depth, dataset transparency, and the presence of quantitative performance metrics. Only those papers which provided sufficient implementation detail, comparative evaluation with state-of-the-art methods, and performance results on benchmark datasets (e.g., INbreast, DDSM, BUSI, CBIS-DDSM)[36] were retained.

To maintain the scientific quality of this review, an internal quality appraisal framework was used. This framework evaluated each article against criteria such as clarity in problem formulation, innovation in methodology, use of public or multi-institutional datasets, inclusion of ablation studies, reproducibility of results, and relevance to breast imaging. Articles were rated on a 5-point scale for each criterion, and only those achieving a minimum threshold across all dimensions were included. This step ensured that the review captures robust, and impactful studies rather than merely reiterating superficial trends.

In total, 70 primary studies were selected for detailed analysis. These works covered a diverse range of deep learning techniques, from standard CNNs to advanced architectures such as Vision Transformers [4], hybrid attention mechanisms [12], U-Net variants [25], and generative adversarial networks (GANs) used for data augmentation or unsupervised segmentation [37]. The selected studies also addressed different challenges such as multi-task learning [8],

[34]-[35], class imbalance [5], [7], [38]-[40], model generalization across domains [37], [41]-[42], and interpretability [4], [20], [31], [43]-[46], reflecting the breadth and complexity of the current research landscape. Special attention was paid to recent innovations such as lightweight DL models for edge deployment [47], multimodal learning frameworks integrating clinical metadata [48], [49], and large-scale self-supervised pretraining [50].

Furthermore, priority was given to studies published in high-impact journals such as IEEE Transactions on Medical Imaging, Medical Image Analysis, Nature Scientific Reports, Computer Methods and Programs in Biomedicine, and *Pattern Recognition*. Inclusion of recent systematic reviews and meta-analyses published between 2020 and 2025 helped validate the selection strategy and ensured thematic consistency.

4. CRITICAL ANALYSIS OF DETECTION METHODS

Over the past decade, CNNs have become the dominant framework for breast cancer detection due to their ability to extract high-level semantic features from medical images[5], [13], [17], [51]-[55]. Despite considerable improvements in detection accuracy, these methods exhibit diverse performance characteristics depending on architecture, dataset, optimization technique, and modality.

Traditional CNN Approaches: Early CNN-based models for breast cancer detection, such as those by Mikhailov et al.[56], focused on relatively shallow architectures applied to histopathological images, achieving moderate accuracy (~85%) but struggling with generalization across magnification levels and datasets [56]. Later studies incorporated deeper networks and transfer learning. For instance, Gonçalves et al. utilized ResNet50, DenseNet201, and VGG16 with transfer learning, significantly improving classification accuracy to 91.67% on the DMR-IR dataset [57]. However, these models often relied on pre-trained weights from natural image datasets, introducing domain mismatch issues that limit their performance in breast imaging.

Optimization-Based CNN Enhancements: To address the limitations of conventional architectures, optimization techniques such as evolutionary algorithms have been integrated. The BreastCNet model [57], for instance, introduced a dual-optimizer approach using the Grey Wolf Optimizer (GWO) [58] to fine-tune neuron counts in dense layers and the Parrot Optimizer (PO) [59] to adjust the learning rate dynamically. These strategies mitigated overfitting, reduced convergence time, and improved generalization, yielding 98.10% validation accuracy and an AUC of 0.995 on the BUSI dataset. The integration of multi-task learning for simultaneous classification and bounding box regression further distinguished BreastCNet from traditional single-task CNNs, enabling efficient, real-time diagnostic inference [5].

Transfer Learning and Feature Fusion: Numerous studies have leveraged pre-trained models such as ResNet, EfficientNet, and DenseNet to transfer learned features from large-scale datasets like ImageNet to the breast imaging domain [60], [61], [62]. While EfficientNet-based models have shown good trade-offs between accuracy and parameter count, they are typically optimized for classification only and require architectural adaptation for localization. For example, Petrini et al.'s EfficientNet-based CAD system achieved 0.9344 AUC on CBIS-DDSM, yet lacked integrated lesion localization capabilities [9]. Similarly, YOLO-based detectors (YOLOv4, YOLOv5) demonstrated real-time performance but required additional classification branches for dual-task capability, adding complexity and memory overhead [63], [9].

Multi-Task Learning (MTL) for Joint Detection and Localization: MTL-based models have gained attention for their ability to share feature representations across related tasks. Ding et al. introduced a ResNet-GAP architecture that jointly performed classification and localization in ultrasound images, though with limited optimization and modest accuracy (88.6%)[5], [64]. In contrast, the BreastCNet architecture offered significant innovation by integrating MTL with hybrid optimization, allowing efficient learning of both tasks in a unified framework [57]. Compared to previous models, BreastCNet's dual-branch output achieved higher F1-scores (0.98) and Intersection over Union (IoU) scores (0.96), indicating superior localization precision [5].

Lightweight Models and Clinical Feasibility: While heavy models like YOLOv4 (~60M parameters) offer state-of-the-art detection accuracy, their computational cost poses barriers for clinical deployment. Lightweight architectures such as MobileNetV2 [41], SqueezeNet [42], and EfficientNet-B0 [21] offer fewer parameters (~3–5M) and reduced inference latency, but require additional modules for localization and lack specialization for medical image characteristics. BreastCNet, with ~1.12M parameters and ~1.58 GFLOPs, achieves a practical balance between performance and efficiency, making it suitable for edge deployment in clinical workflows [57].

Generalization and Cross-Modality Validation: Many CNN models suffer from limited generalization due to overfitting on specific datasets[65]. BreastCNet's training on BUSI and validation on DDSM and INbreast datasets confirmed its robust cross-modality performance. It achieved 97.50% accuracy on DDSM and 96.20% on INbreast, outperforming other models that showed performance degradation when exposed to unseen data [57], [23], [35].

Interpretability in Detection Models: Interpretability is essential for clinical adoption. Grad-CAM was used in BreastCNet to visualize the decision-making process, improving trust and transparency [57]. However, the most

detection models in the literature do not incorporate sufficient interpretability tools, reducing their acceptance in clinical settings[5], [57].

In conclusion, current detection methods in breast cancer imaging show a clear progression from basic CNN architectures to sophisticated multi-task, optimization-enhanced, and lightweight models. While high accuracy has been achieved in many studies, key limitations persist in terms of computational cost, lack of real-time integration, generalization across imaging modalities, and explainability.

5. CRITICAL ANALYSIS OF SEGMENTATION METHODS

Segmentation of breast tumors from medical images plays a vital role in identifying the extent and location of the malignancy, aiding in treatment decisions. With advances in deep learning techniques, CNN based methods have gained attention for their ability to achieve an accurate and automated segmentation [66]. Early and accurate segmentation of breast tumors is essential for precise diagnosis, treatment planning, and monitoring. CNN methods have emerged as powerful tools for medical image analyses, including breast cancer segmentation [67]. Segmentation methods, such as U-Net[1], YOLO[2], Mask R-CNN, and contour-based methods, have been extensively studied and developed for various applications in computer vision and medical imaging [68]-[73]. However, instance segmentation goes beyond semantic segmentation by not only assigning class labels, but also distinguishing between individual instances of objects within a given class [74]. It involves identifying and segmenting each distinct object instance separately, providing precise localization and differentiation of objects. It is typically built upon object detection algorithms, and extends them to include pixel-level segmentation masks for each detected object instance [74].

Other segmentation methods, such as graph-cut algorithms, region-based approaches, and deep learning-based architectures, have also been explored and developed for various segmentation tasks [72], [73]. These methods aim to improve segmentation accuracy, efficiency, and generalization capabilities. However, the main problem is that a highly accurate and efficient instance segmentation of breast cancer tumors in images should be reached, a challenging task due to varying tumor shapes, sizes, and appearances. Methods range from basic CNN architectures to more advanced models, such as U-Net, Mask R-CNN, and DenseNet for medical image segmentation tasks. While these methods are likely to be promising, they often struggle with complex variations in tumor presentation and suffer from lower generalization across diverse datasets[12]. Currently, models like Mask R-CNN and U-Net are considered among the best for medical image segmentation tasks due to their ability to handle detailed pixel-wise segmentation and detection. However, even the most advanced models have limitations in terms of generalizability, especially when applied to different datasets like BUSI, DDSM, and INbreast. Breast cancer segmentation methods like UNet, VGG16, and similar CNN-based models face multiple limitations[12], [75]. These include poor generalization across different imaging datasets due to overfitting, difficulty in accurately detecting small, irregular or diffuse lesions, and high sensitivity to noise and artifacts in medical images. Additionally, the computational cost and resource requirements of these models make real-time clinical application challenging, particularly in resource-limited settings. The models also lack interpretability, limiting clinical trust, and fail to effectively capture multiscale information, which is critical for segmenting tumors of varying shapes and sizes. These methods also lack full automation for clinical workflows and are unable to perform precise instance segmentation, which could otherwise distinguish between multiple, adjacent tumor instances [12].

Despite considerable advancements, current segmentation methods face key limitations, particularly in generalization, precision, and clinical applicability.

5.1. Traditional CNN-Based Segmentation: From U-Net to Its Variants

The seminal U-Net architecture by Ronneberger et al. [25] revolutionized biomedical segmentation through its encoder-decoder framework with skip connections, enabling pixel-level precision with relatively few annotated samples. However, baseline U-Net often underperforms in segmenting small or low-contrast lesions, a common issue in mammography and ultrasound imaging. To address this, U-Net++ introduced nested dense skip pathways, reducing the semantic gap between encoder and decoder features [76]. Meanwhile, Attention U-Net incorporated attention gates to emphasize tumor-relevant features [77].

Despite these innovations, several studies have highlighted U-Net's limitations. Its rigid architecture struggles with significant shape variability and inter-patient heterogeneity. For example, experiments on the BUSI and DDSM datasets revealed sensitivity drops when segmenting irregular or low-contrast tumors, particularly in ultrasound and dense breast tissues[78], [79]. Moreover, performance often degrades when models trained on one dataset are applied to another, due to domain shift and overfitting [12], [80]-[83].

[1] U-shaped encoder-decoder network
[2] You only look once

5.2. Instance Segmentation: Mask R-CNN and Beyond

Instance segmentation, unlike semantic segmentation, aims to identify each object instance separately. Mask R-CNN has emerged as a leading architecture for this task by extending Faster R-CNN with an additional segmentation head, allowing simultaneous object detection and pixel-level mask generation [13], [55], [84]-[85]. In breast imaging, Mask R-CNN has shown superior results on datasets like INbreast and CBIS-DDSM, especially for well-defined lesions [86]-[87]. However, its reliance on bounding box proposals makes it less effective for diffuse or small-scale tumors. Additionally, the model's computational complexity (~140M parameters) hinders real-time clinical application, particularly in low-resource settings [13].

To reduce computational burden, lightweight adaptations like Lite-Mask R-CNN and EfficientDet+Mask modules have been proposed. While these models maintain reasonable segmentation performance (IoU ~0.82 on BUSI), they often sacrifice fine-grained accuracy and are prone to false positives in noisy modalities like ultrasound [53].

5.3. Transformer-Based and Hybrid Architectures

Transformers have recently been incorporated into segmentation pipelines to model long-range dependencies. TransUNet, Swin-UNet, and MedT combine CNN backbones with transformer encoders, yielding better contextual awareness and improved Dice coefficients, particularly in MRI segmentation [88]-[89]. However, their reliance on large-scale datasets and high GPU memory limits widespread adoption in clinical practice.

Hybrid models like CNN-ViT and Res-TransUNet have been proposed to balance spatial feature extraction and global context modeling [90]. For instance, Res-TransUNet achieved a 3–5% gain in Dice and Jaccard scores over pure CNN models on the BUSI dataset, especially for multi-region tumors [90] .However, these models require careful tuning of transformer depth and attention granularity to avoid overfitting in small datasets.

5.4. Graph-Based and Edge-Aware Segmentation

To overcome boundary ambiguity, several studies have explored graph-cut methods and edge-aware segmentation. Deep contour-aware networks (e.g., DCAN) and GCN-integrated U-Nets leverage boundary supervision to improve precision around lesion edges[91]-[92]. This is especially beneficial in ultrasound, where tumors often lack clear demarcation. However, integrating these modules increases architectural complexity and training time.

Recent efforts using graph neural networks (GNNs) in post-processing stages have enhanced connectivity preservation and reduced artifacts [93]. For example, GNN-refined U-Nets improved boundary recall by 6% on BUSI but required additional annotations for edge maps, limiting scalability [93].

5.5. Generalizability, Robustness, and Cross-Modality Performance

A recurring challenge in breast cancer segmentation is the generalization of models across imaging modalities and datasets. Many studies report high Dice coefficients (>0.90) on training datasets but observe significant performance drops on external validation sets due to domain shifts in image quality, resolution, and annotation styles [31,33,80]. Few-shot learning, domain adaptation (e.g., CycleGAN-based style transfer), and data augmentation using generative adversarial networks (GANs) have shown promise in mitigating this issue [37], [41], [94].

However, GAN-based augmentation, while beneficial, can introduce unrealistic samples that skew model learning. Unsupervised domain adaptation techniques remain underexplored in breast cancer imaging compared to other fields like retinal or brain imaging [36].

5.6. Computational Constraints and Clinical Integration

Many of the top-performing segmentation models—such as Transformer-based or hybrid networks—demand significant memory, long inference times, and are not optimized for real-time application. Lightweight segmentation architectures like Mobile-UNet and Squeeze-UNet have addressed this by reducing parameter counts while maintaining reasonable accuracy (Dice ~0.85) [95]. Nonetheless, a trade-off persists between efficiency and performance.

Moreover, clinical adoption is hindered by the lack of model interpretability. Few segmentation frameworks integrate explainable AI (XAI) tools like Grad-CAM or saliency maps, which are vital for clinician trust[4], [47], [65], [88]. Embedding such mechanisms in future models is essential for transparency in segmentation decisions.

Despite the considerable progress in deep learning-based segmentation methods for breast cancer, several unresolved issues remain. One major challenge is the poor generalization of models across different imaging modalities and institutions, often caused by dataset bias and overfitting. Additionally, many models struggle with accurately detecting small, irregular, or overlapping tumors, particularly in low-resolution or noisy images such as those produced by ultrasound. The high computational complexity of advanced architectures further limits their feasibility for real-time deployment on edge devices, especially in resource-constrained clinical environments. Moreover, the lack of

interpretability and clinical explainability in most segmentation models diminishes their acceptance and trust among radiologists. Lastly, there is limited integration of multi-modal information, such as patient metadata and radiomics features, which could otherwise enhance diagnostic precision and model robustness.

6. CRITICAL ANALYSIS OF HYBRID AND MULTI-TASK LEARNING MODELS

Hybrid models and multi-task learning (MTL) frameworks have recently emerged as promising paradigms in breast cancer analysis due to their capacity to jointly address related objectives—such as classification, detection, and segmentation—within a single, integrated architecture. These frameworks aim to enhance computational efficiency, promote feature sharing, and improve predictive consistency across tasks. However, several fundamental challenges still hinder their full potential in clinical applications.

6.1. Efficiency vs. Specialization Trade-off

MTL models are increasingly adopted in breast cancer analysis due to their ability to jointly handle tasks such as tumor detection and segmentation within a shared computational framework. A central advantage of these models lies in their computational efficiency: by employing a shared encoder with dedicated decoders for each task, MTL architectures significantly reduce model size, training time, and inference latency. This shared representation facilitates the transfer of common features, which can enhance generalization, particularly when data resources are limited [96].

However, this shared structure also introduces a fundamental trade-off between efficiency and task-specific specialization. While shared layers capture general features beneficial to multiple tasks, they may struggle to model the fine-grained distinctions needed for each individual objective. This can result in task interference, where gradients from competing tasks update shared parameters in conflicting directions, ultimately hindering convergence or degrading performance on one or more tasks [97].

To address these challenges, recent studies have explored adaptive sharing strategies, such as soft-parameter sharing or modular encoders with attention-based routing, which aim to maintain efficiency while allowing task-specific specialization [96], [98]. Nevertheless, determining the optimal balance between shared and task-specific components remains a significant design challenge in MTL architectures, particularly in complex medical imaging scenarios where tasks demand differing levels of semantic abstraction.

6.2. Architectural Complexity and Optimization Challenges

Architectural design plays a central role in the effectiveness of hybrid models. Simple hard parameter sharing (i.e., full encoder sharing) often fails to capture task-specific nuances. More sophisticated designs, such as partially shared encoders, gated multi-branch networks, and task-aware attention mechanisms, aim to separate shared and private representations more effectively. For example, Task-Aware Attention Networks (TAAN) have shown notable improvements by dynamically routing features to task-specific decoders based on learned attention maps [2], [52].

Nonetheless, determining the optimal depth, width, and connection schema for shared versus private layers remains an open research problem. There is no one-size-fits-all solution, especially for medical imaging tasks that differ in spatial precision and semantic interpretation.

6.3. Joint Loss Function Formulation

The joint optimization of multiple tasks relies heavily on well-designed loss functions. Most hybrid models combine individual losses such as Dice loss (for segmentation), focal loss (for detection), and cross-entropy loss (for classification). A fixed weighting of these losses can cause training imbalance, where one task dominates the learning process. Adaptive approaches, like Uncertainty Weighting [99], Gradient Normalization (GradNorm) [75], [100] and Adaptive Loss Balancing (ALB), offer dynamic weighting schemes to stabilize training and improve convergence across all tasks.

However, these strategies are computationally intensive and may still be sensitive to hyperparameter tuning, making real-world implementation more complex.

6.4. Evaluation Metrics and Benchmarking Inconsistencies

Evaluating hybrid models is challenging due to the diversity of metrics involved. Typically, segmentation is assessed using Dice coefficient and IoU [101], while detection/classification uses AUC [102] accuracy, sensitivity, and specificity. Reporting these metrics independently makes it difficult to compare models holistically. Furthermore, a model that excels in one task might underperform in another, which is not acceptable in clinical workflows.

Some researchers have proposed composite evaluation metrics or task-weighted scoring systems to provide a balanced perspective [35], [96]-[97]. However, such methods are not yet standardized, leading to inconsistencies across published studies.

6.5. Data and Annotation Bottlenecks

High-quality labeled datasets are essential for effective multi-task learning. Unfortunately, datasets that provide aligned annotations for all tasks (e.g., segmentation masks along with class/detection labels) are rare in the medical imaging domain[12], [103]. Many publicly available datasets offer either classification labels or segmentation masks, but not both.

To mitigate this, hybrid training schemes have been proposed, including semi-supervised learning, weak supervision, and pseudo-label generation. For instance, self-supervised pre-training followed by fine-tuning on limited labeled data has shown promise [50]. Nevertheless, the lack of comprehensive, multi-task annotated datasets remains a significant bottleneck.

6.6 Generalization and Clinical Translation Challenges

Despite their academic success, hybrid models often face performance degradation when applied to real-world clinical data due to domain shifts—variations in imaging modality, scanner settings, and patient demographics. Models trained on public datasets like BUSI or CBIS-DDSM may fail to generalize to images from other institutions or populations [12], [82], [104].

Transfer learning and domain adaptation techniques have been employed to mitigate this issue, but the robustness and interpretability of these models are still under scrutiny [39]. Clinical translation also requires transparency and reproducibility—qualities that many current MTL models lack due to their complex architecture and black-box nature.

hybrid and multi-task learning frameworks present a compelling solution for unified breast cancer diagnosis, offering the potential to perform classification, detection, and segmentation in a single model. While these models show significant promise in reducing computational redundancy and improving diagnostic coherence, several challenges persist. Future research must address task interference, architectural optimization, adaptive loss balancing, and data scarcity. Moreover, cross-domain generalization and clinical interpretability should be prioritized to ensure successful real-world deployment of these advanced AI systems.

7. COMPARATIVE ANALYSIS AND SUMMARY

This section presents a critical comparative analysis of recent deep learning approaches for breast cancer detection, segmentation, and unified end-to-end diagnosis frameworks. By systematically evaluating performance metrics, architectural design, computational demands, and clinical relevance, we identify the strengths, limitations, and gaps across state-of-the-art models. The methods are categorized into three core areas: (1) tumor detection, (2) tumor segmentation, and (3) integrated models capable of simultaneous detection and segmentation. Within each category, related models are grouped and contrasted based on shared methodologies or design goals, enabling a more cohesive understanding of trends and challenges in the field. This structured comparison facilitates the identification of promising directions for future research and practical deployment in clinical settings.

7.1. Breast cancer detection

The landscape of breast cancer detection using deep learning is rich with diverse approaches, ranging from classical CNNs to transformer-based and optimization-enhanced architectures. Traditional models, like those used by Das & Rana [100] and Kumar et al. [103], employ deep CNNs or ResNet variants but primarily focus on classification without localization. This limits their clinical utility, as they fail to offer tumor-specific visual guidance. While they achieve moderate to high accuracy (~88–97%), these methods generally lack interpretability and task diversity (e.g., no segmentation or bounding box regression).

In contrast, models like those by Petrini et al. [56] and Sait & Nagaraj [104] explore transfer learning with advanced networks like EfficientNet, demonstrating improved accuracy (~85–99%) on large datasets. However, they still fall short on localization and lack multi-task capabilities, often requiring high computational resources without offering real-time or clinical insight.

Optimization-enhanced methods such as BreastCNet [30] mark a substantial advancement. By integrating hybrid optimizers (GWO and PO) and a multi-task learning (MTL) structure, BreastCNet not only achieves superior performance (AUC: 0.995, IOU: 0.96) but also addresses critical limitations such as task separation, overfitting, and interpretability. This is in contrast to YOLO-based detectors [57], which offer real-time performance but suffer from high false positives and architectural complexity.

Transformer-based models like BUViTNet [55] show promising global context awareness but are limited by their data and compute requirements, as well as their lack of bounding box outputs. Meanwhile, deformable-attention approaches (e.g., ACSNet [43]) offer innovative gating mechanisms yet still lack full integration of classification and localization. Table 1 provides a comprehensive comparison of several methods in recent years (Table 1).

Table 1. Comparison of breast cancer detection models

Author/Model	Year	Technique	Dataset	Accuracy / AUC / F1 / IOU	Gaps/Shortcomings
Das & Rana [105]	2021	ResNet variants	BUSI	Accuracy: 88.89%	Accuracy decreased with deeper models; no localization; lacks optimization and interpretability.
Ding et al. [64]	2022	ResNet-GAP + elastography	BUSI	Accuracy: 88.6%	High complexity; lacks efficient optimization; no cross-dataset generalization shown.
Petrini et al. [9]	2022	EfficientNet + transfer learning	CBIS-DDSM	AUC: 0.9344, Acc: 85.13%	No localization; no MTL; weak generalization; lacks hybrid optimization.
Huynh et al. [106]	2023	YOLOX + EfficientNet	Private	Accuracy: 92%	No mention of optimization or interpretability; only classification; lacks cross-dataset validation.
Ayana et al. [62](BUViTNet)	2022	Vision Transformer	BUSI	AUC: 0.968, Kappa: 0.959	No localization; lacks optimization; limited to ultrasound only.
Sahu et al. [107] (HADLCM)	2023	Hybrid deep-layer cascade	BUSI	Accuracy: 95.0%	No bounding box regression; lacks optimizer usage; evaluated on one dataset.
Kumar et al.[108]	2022	CNN (3 conv layers)	DDSM, CBIS-DDSM	Accuracy: 97.2%	Only classification; no MTL or localization; lacks interpretability framework.
Sait & Nagaraj [109]	2024	EfficientNetB7 + LightGBM	CMMD	Accuracy: 99.9%	Very high complexity; not tested on ultrasound; no localization or MTL.
Prinzi et al.[63]	2023	YOLOv5 + transfer learning	Proprietary	Accuracy: 95.3%, mAP: 0.621	High false positives; lacks classification branch; no dual-task support.
Yu H et al. (ACSNet)[50]	2024	Deformable attention + gated CNN	BUSI	Accuracy: 94.44%	No bounding box output; no optimization; interpretability limited.
Mishra et al.[110] (MultiRUSNet)	2024	UNet-ResNet hybrid	BUSI	Accuracy: 95.2%, Dice: 0.741	No localization; no optimizer tuning; lacks cross-modality testing.
Mahichi et al.[5] BreastCNet	2025	CNN + GWO + PO (MTL)	BUSI, DDSM, INbreast	Accuracy: 98.10%, AUC: 0.995, F1: 0.98, IOU: 0.96	Balanced performance, MTL, hybrid optimization, interpretable with Grad-CAM.

7.2. Breast cancer segmentation

Breast tumor segmentation methods (Table 2) primarily rely on U-Net variants, instance segmentation models like Mask R-CNN, and more recently, transformer hybrids. Classical U-Net and its extensions (e.g., U-Net++, Attention U-Net) [71]-[72] are favored for their pixel-wise segmentation capabilities. However, they struggle with small or irregular lesions, particularly in ultrasound images where contrast is low. Studies such as those by Vakanski et al. [101] and Zhao & Dai [103] demonstrate strong performance (DSC: ~90%) but often lack robustness across datasets.

Multi-scale and attention-based networks like MDF-Net [104] and CV-VAE [106] introduce advanced feature fusion and spatial disentanglement, achieving better generalization and Dice scores (up to 93.70%). However, they are often

limited by training complexity and computational demands, making real-time deployment difficult. Similarly, hybrid and ensemble models like those by Bobowicz et al. [105] and Karunanayake et al. [82] offer high performance but require complex setups that hinder scalability and interpretability.

Instance segmentation models (e.g., Mask R-CNN) provide finer object-level delineation, yet their reliance on bounding box proposals often limits performance on diffused or small tumors. They also introduce high computational overhead (~140M parameters), making them impractical for edge deployment.

Table 2. Comparison of Breast tumor segmentation methods

Author	Year	Dataset	Method	Results	Gaps/Shortcomings
Vogl et al. [111]	2019	Multi-modal data (34 Patients)	Random Forest Classifier with mpPET/MRI	Segmentation Dice Coeff: 0.665	small dataset size, feature predictiveness, less modalitation between classification and segmentation
Vakanski et al. [112]	2020	510 breast ultrasound images	U-Net with attention blocks incorporating visual saliency	DSC of 90.5%	Lower accuracy compared to other advanced models, Reliance on visual saliency may not fully capture variability in medical images, Limited dataset scope, Need for further validation on diverse datasets, Potential overfitting, Scalability concerns
Irfan et al. [113]	2021	Breast ultrasonic image	Dilated Semantic Segmentation Network (Di-CNN) + DenseNet201 + 24-layer CNN + SVM	Accuracy: 98.9%, Mean-IoU: 52.89%, Mean Accuracy: 79.61%, Weighted-IoU: 73.83%, Mean-BF Score: 0.18218	Potential overfitting due to high complexity, Time-consuming training with 500 epochs, Limited explanation on performance with varied dataset sizes,
Zhao and Dai [114]	2022	Ultrasound images	U-Net + Residual Block + Attention Mechanism	Dice Index: 0.921	Evaluation metrics limited to Dice Index, IoU, and HD Index, Dataset details not clearly specified
Qi W et al. [115]	2023	Dataset A: Breast Ultrasound Lesions Dataset Dataset B: Baheya Hospital Ultrasound Dataset	Multi-scale Dynamic Fusion Network (MDF-Net)	Dataset A: DSC: 83.63%, IoU: 75.83%, Sensitivity: 85.07%, Specificity: 99.42% Dataset B: DSC: 78.20%, IoU: 70.22%, Sensitivity: 79.44%, Specificity: 98.22%	- Sensitivity slightly lower on Dataset B - No explicit reporting on overall accuracy, AUC, F1 Score, and mAP, 500 epoch cause overfitting.

Zhang S et al. [77]	2023	External datasets for BUS imaging	Dual-branch model (classification & segmentation)	AUC: 0.991 (classification); DSC: 0.898 (segmentation)	Slightly lower segmentation precision compared to specialized models; Increased complexity due to dual-branch architecture; Synchronization challenges between classification and segmentation tasks; Complex loss function design needed.
Karunanayake N et al. [88]	2024	Three distinct ultrasound datasets	AI-based hybrid model combining deep learning and multi-agent artificial life	Dice coefficients: 0.96 (easy), 0.91 (medium), 0.90 (hard); Relative Hausdorff distance: H3 = 0.26 (easy), H3 = 0.82 (medium), H3 = 0.84 (hard)	Reliance on initial DL segmentation quality; Lack of direct comparison to instance segmentation methods, Computational Complexity
Bobowicz M et al. [116]	2024	BUS B, OASBUD, BUSI, UCC BUS	PraNet, CaraNet, FCBFormer (ensemble classifiers)	IoU: 0.81, 0.80, 0.73; Dice: 0.89, 0.87, 0.82	Lower segmentation accuracy compared to ISRFE-DO; Requires complex ensemble setup for classification, Computational complexity, scalability issues, complexity in implementation.
Ma Yet al. [117]	2024	CBIS-DDSM, INbreast	Cross-View Variational Autoencoder (CV-VAE) with spatial hidden factor disentanglement, FPN-based classifier, and U-Net-like decoder	DSC: 92.46%, 93.70%. F1-score: 0.864	Requires large labeled datasets and high computational power, limiting use in small clinics.

7.3. End to end breast cancer detection and segmentation Simultaneously

End-to-end models that simultaneously detect and segment tumors represent a promising, holistic approach for breast cancer diagnosis. Most such models attempt to unify classification and segmentation, although often at the expense of performance trade-offs in one domain.

For example, EDCNN by Islam et al. [14] combines MobileNet and Xception for improved detection, yet it falls short in segmentation accuracy (IoU: 0.77). Models like DDA-AttResUNet [96] and DAU-Net [107] emphasize segmentation using attention-enhanced U-Net variants, achieving Dice scores above 92%, but do not support classification or localization, reducing their diagnostic completeness.

UCapsNet [108] and CoAtUNet [109] incorporate capsule networks and transformer modules, respectively, offering strong classification and segmentation performance (~99% accuracy and Dice ~95%). However, their high computational cost, lack of real-time capability, and evaluation on limited datasets weaken their generalizability and clinical viability. Table 3 summarizes several methods for breast cancer classification, localization and tumor segmentation, (Table 3).

Table 3. Comparison of breast cancer detection and segmentation Simultaneously methods

Author (Year)	Dataset	Model Architecture	Performance Metrics	Strengths	Gaps/Shortcomings
Islam M. et al. [24](2024)	BUSI, UDIAT	EDCNN (Ensemble Deep CNN - MobileNet + Xception)	Accuracy: 87.82%, AUC: 0.91, F1-score: 86.00%, IoU: 0.77	Combined MobileNet and Xception for enhanced detection, interpretability with Grad-CAM	Limited segmentation effectiveness, lower accuracy than BreaVisioNet, no classification component
Hekal A. et al.[101] (2024)	BUSI	DDA-AttResUNet (Dual Decoder Attention ResUNet)	Dice: 92.92%, IoU: 87.39%, Accuracy: 98.82%, Sensitivity: 92.16%, Precision: 93.90%	Focuses on segmentation accuracy, Dual Decoder Attention enhances tumor region emphasis	Restricted to segmentation without classification, image resizing reduces fine tumor detail, limited generalizability
Carriero A. et al. [118](2024)	BUSI	Unet3+ (with FCN-32s, Unet, SegNet, DeepLabV3+, PSPNet)	Accuracy: 82.53%, IoU: 52.57%, Weighted IoU: 89.14%, F1-score: N/A	Unet3+ showed high accuracy and was compared with various models	Struggles with small/irregular tumors, poor generalization (training vs. eval IoU gap), dataset size is insufficient
Pramanik et al. [119](2024)	BUSI, UDIAT	DAU-Net (Dual Attention U-Net with PCBAM & SWA)	Dice: 74.23% (BUSI), 78.58% (UDIAT)	Enhanced feature extraction with dual attention mechanisms, hybrid loss function	Lacks tumor classification, limited to segmentation, no real-time processing capabilities, dataset-specific evaluation
Madhu et al.[120] (2024)	BUSI	UCapsNet (Hybrid U-Net + Capsule Network)	Classification Accuracy: 99.22%, Segmentation Accuracy: 99.07%, Dice Score: 95.14%	Strong segmentation, capsule network improves classification accuracy	High computational cost, not optimized for real-time deployment, limited dataset evaluation (only BUSI dataset)
Zaidkilani et al. [11] (2025)	BUSI, UDIAT	CoAtUNet (Convolutional Transformer U-Net)	(Results not specified yet)	Strong transformer-based architecture, attention mechanisms to enhance feature extraction	Potential computational complexity, may require further optimization for clinical deployment
Schutte et al. [121] (2024)	Private Dataset	CNN-based Model (Multi-stage CNN)	Accuracy: 86.4%; Sensitivity: 90.2%; Specificity: 85.1%	Uses multi-stage processing for improved sensitivity	Dataset limited to a private source, restricting generalizability

8. IDENTIFIED RESEARCH GAPS AND CHALLENGES

Despite the rapid advancement of deep learning in breast cancer detection and segmentation, numerous unresolved challenges continue to hinder real-world clinical integration. These challenges span across data availability, model generalization, interpretability, efficiency, and evaluation consistency. Based on critical analyses throughout this review, the following research gaps have been identified.

8.1. Data Limitations and Annotation Scarcity

The development of effective deep learning models is fundamentally limited by the scarcity of large, annotated, and diverse breast cancer imaging datasets. Most existing datasets (e.g., BUSI, INbreast, CBIS-DDSM) either lack segmentation masks, classification labels, or are heavily imbalanced in terms of lesion types and sizes [5], [36]-[37] . Particularly, datasets with aligned annotations for multi-task learning (e.g., joint classification, detection, and segmentation) are rare, hindering unified model training [5], [8], [12]. Moreover, manual annotation of breast lesions requires expert radiologists, making the process costly and time-consuming [103]. Although weakly supervised, semi-supervised, and self-supervised methods have been introduced to reduce dependency on labels [51], their use in breast imaging remains under-investigated and not yet mature for deployment.

8.2. Generalization and Domain Shift Challenges

A significant number of models perform well on internal validation sets but struggle to generalize across external datasets or modalities. Domain shift caused by differences in imaging protocols, scanner types, or patient demographics often leads to performance degradation [31], [41], [82], [122]. Techniques such as domain adaptation using CycleGANs [37], few-shot learning, and cross-modality training[42] have been explored, but robust cross-institutional generalization remains limited, especially for segmentation tasks [80], [83].

8.3. Limited Interpretability and Clinical Trust

Deep learning models are often viewed as "black-box" systems, which reduces trust and hinders clinical acceptance. Many high-performing models lack mechanisms to explain their decision-making process, making them unsuitable for clinical workflows where transparency is essential [4], [20], [43]. While methods like Grad-CAM, saliency maps, or attention visualization have been adopted in some frameworks (e.g., BreastCNet [5]), few models embed interpretability natively within their architectures, particularly in segmentation pipelines [57], [65], [88].

8.4. Trade-off Between Model Complexity and Real-Time Feasibility

Transformer-based models, capsule networks, and ensemble frameworks often offer superior accuracy at the cost of massive computational overhead, making them unsuitable for real-time diagnosis or use in low-resource clinical environments[62], [88], [90]. On the other hand, lightweight models such as MobileNet [95]and EfficientNet-B0 [9] offer reduced complexity but typically suffer from performance trade-offs in segmentation accuracy or lesion localization. Striking an optimal balance between model performance and deployment feasibility remains a key challenge for breast imaging systems [95].

8.5. Multi-Task Learning Optimization and Interference

Multi-task learning (MTL) frameworks enable simultaneous learning of classification, detection, and segmentation, promoting efficiency and feature sharing [34], [35], [50], [123]. However, MTL introduces task interference, where one task (e.g., classification) may dominate training and suppress others (e.g., segmentation), especially under fixed loss weights [8], [99]. Although adaptive loss weighting strategies like GradNorm [75], Uncertainty Weighting [99], and ALB have shown promise, MTL optimization is still highly sensitive to architecture design and hyperparameter tuning, limiting stability and reproducibility [98], [100].

8.6. Inconsistent Evaluation Protocols and Benchmarking

Current studies adopt diverse datasets, metrics, and evaluation criteria, making direct comparisons difficult and impeding progress [24], [101]-[102]. Some report only segmentation metrics (Dice, IoU), others emphasize classification (accuracy, AUC), while very few attempt a comprehensive multi-task evaluation. The lack of standardized benchmarks or leaderboards for breast cancer imaging (as seen in fields like natural image segmentation) hampers reproducibility and model comparison, leading to fragmented progress[8], [35].

8.7. Limited Integration of Multimodal and Clinical Metadata

Most models rely solely on imaging data, overlooking valuable clinical context such as patient history, genetic markers, and radiomics features [48]-[49]. This represents a missed opportunity for improving diagnostic accuracy and robustness. While some recent models integrate structured metadata with imaging via multimodal fusion techniques, this remains rare, technically complex, and poorly standardized. Effective fusion of non-image data with imaging pipelines demands further exploration to enhance diagnostic precision and personalization [50].

Table 4 summarizes the major research gaps and challenges identified in recent deep learning-based breast cancer imaging studies, including issues related to data scarcity, model generalization, interpretability, computational efficiency, multi-task learning optimization, evaluation inconsistency, and multimodal integration.

Table 4. Gaps and challenges identified in recent deep learning-based breast cancer imaging studies

Research Gap / Challenge	Description
Limited Annotated Datasets	Public datasets (e.g., BUSI, CBIS-DDSM, INbreast) often lack multi-task annotations (classification + detection + segmentation). Manual labeling is time-consuming and expert-dependent. Multi-task datasets are rare, limiting unified model development.
Poor Generalization and Domain Shift	Models trained on specific datasets often fail on unseen domains due to variations in image quality, acquisition devices, and patient demographics. Generalization across modalities is not guaranteed.

Research Gap / Challenge	Description
Lack of Interpretability and Clinical Trust	Most models are black-box systems. Few integrate interpretable tools like Grad-CAM or saliency maps natively. Clinicians require transparent decision support.
High Computational Complexity	Transformer-based, capsule, and ensemble models achieve high accuracy but are resource-intensive. Lightweight models reduce complexity but often compromise segmentation accuracy or localization ability.
Multi-Task Learning (MTL) Interference	MTL frameworks often face imbalanced task learning where one objective dominates. Optimizing joint loss functions is difficult. Adaptive balancing methods (e.g., GradNorm) are promising but computationally demanding.
Inconsistent Evaluation Protocols	Studies use varied datasets, metrics (e.g., AUC, Dice, IoU), and lack statistical validation. Absence of unified benchmarks prevents direct model comparisons.
Limited Multimodal Integration	Most models use imaging data alone. Clinical metadata (e.g., age, family history, radiomics) is underutilized. Integration techniques are complex and poorly standardized.

9. FUTURE RESEARCH DIRECTIONS

Building on the identified challenges and gaps outlined in the preceding sections, future research in deep learning-based breast cancer diagnosis must prioritize several critical areas to ensure greater clinical applicability, robustness, and interpretability of AI systems. Despite the rapid evolution of deep neural network architectures and their promising diagnostic capabilities, translating these systems into real-world healthcare environments requires overcoming persistent limitations related to data scarcity, generalization, transparency, computational efficiency, and workflow integration.

While our review highlights significant progress in the field, a critical evaluation reveals a dichotomy in the development of DL models for breast cancer detection and segmentation. On one hand, complex, state-of-the-art models like YOLOv4 demonstrate impressive accuracy but often come with a substantial computational cost and latency, making them less suitable for real-time clinical applications. On the other hand, lightweight models such as BreastCNet offer faster inference times but may sacrifice a degree of diagnostic precision. This trade-off between performance and efficiency presents a key challenge for clinical translation. Furthermore, the limited interpretability of many black-box models, despite their high-performance metrics, remains a significant barrier to their adoption by clinicians who require confidence and transparency in AI-driven diagnoses. This highlights a critical need for future research to not only pursue higher accuracy but to do so in a manner that is both computationally efficient and inherently interpretable.

A foundational priority is addressing the heavy reliance on large-scale, fully annotated datasets, which continues to hinder the development of robust and scalable models. Annotating breast cancer imaging data—especially for pixel-wise segmentation or instance-level detection—is time-consuming, costly, and requires expert radiologists. As such, future work should explore the development and integration of self-supervised and semi-supervised learning strategies that can leverage unlabeled or partially labeled data to enhance model learning. Techniques such as contrastive learning, pseudo-label generation, and consistency regularization have demonstrated success in other domains and hold significant promise in medical imaging contexts. Their application in breast cancer analysis, however, remains underutilized and demands greater research attention [43]. Implementing such methods can alleviate the annotation burden, enable more efficient use of available datasets, and ultimately lead to models that generalize better across patient populations and imaging devices.

Interpretability is another crucial direction for future exploration. Many of the high-performing deep learning models currently used in breast cancer imaging function as "black boxes," which limits their acceptability in clinical settings. Although post hoc techniques such as Grad-CAM or saliency maps offer visual cues about model attention, they do not provide inherently explainable reasoning mechanisms. Future architectures should incorporate interpretability directly into their structure, such as through attention-guided feature selection, decision-aware modules, or explicit representation learning. By embedding transparency into the model pipeline, researchers can improve clinical trust and facilitate model validation, a necessary step for regulatory approval and physician acceptance [1], [36], [50], [59].

Another key area for advancement lies in the optimization of MTL frameworks. These models are increasingly used for joint prediction tasks, such as simultaneous tumor classification, detection, and segmentation, as they improve computational efficiency and promote knowledge sharing across related tasks. However, one common challenge with MTL approaches is task interference, where learning one objective may negatively affect performance on another. To address this, dynamic loss-balancing methods such as Uncertainty Weighting, GradNorm, or Adaptive Loss Balancing (ALB) have been proposed, offering better stability and task convergence during training [25], [26], [94]-[95]. In addition, neural architecture search (NAS) and automated hyperparameter tuning may be employed to adaptively optimize MTL configurations for diverse imaging modalities and objectives.

From a deployment perspective, the creation of lightweight, resource-efficient deep learning models is imperative, particularly for use in point-of-care or low-resource environments. Current state-of-the-art models, particularly transformer-based or ensemble networks, often demand high computational resources and large memory footprints, which hinder their usability in real-time clinical workflows. Designing and optimizing architectures such as MobileNet, EfficientNet-lite, or SqueezeNet—alongside model compression techniques including pruning, quantization, and knowledge distillation—can yield smaller, faster, and more efficient models without significantly compromising diagnostic accuracy. These models would be particularly valuable in settings with limited access to high-performance computing infrastructure, such as community hospitals or remote clinics [9], [95], [124].

In parallel, future research must explore multimodal fusion strategies that incorporate both imaging data and complementary clinical metadata. Most current deep learning models are limited to analyzing image pixels, disregarding rich contextual data such as patient age, genetic risk factors, hormonal status, and prior imaging history. Integrating such non-image data into diagnostic models via multimodal fusion—using strategies like cross-modal transformers, graph-based neural networks, or attention-based late fusion—can enhance prediction robustness, enable personalized diagnosis, and improve the interpretability of model outputs. This direction also aligns well with precision medicine objectives, which aim to tailor interventions based on both phenotypic and clinical profiles [41]-[43], [125].

A persistent challenge in current research is the lack of model generalization across domains, institutions, and imaging protocols. Many models perform well on internal validation sets but exhibit significant degradation when applied to external datasets due to domain shift. Future efforts should focus on improving domain adaptation and transfer learning strategies. Techniques such as unsupervised domain adaptation, few-shot learning, and image-to-image translation using generative adversarial networks (e.g., CycleGAN) have shown potential in mitigating domain variability and improving robustness [28], [34], [77], [126]-[127]. However, their systematic application to breast cancer imaging remains limited. Moreover, constructing large-scale, diverse, multi-institutional datasets would provide the necessary benchmark for fair and reproducible evaluation of generalization performance across populations and imaging devices.

Equally important is the design of unified, end-to-end architectures that consolidate multiple diagnostic tasks—such as classification, localization, and segmentation—within a single streamlined model. At present, many models are developed and trained separately for each task, leading to redundant computation and lack of clinical integration. A unified approach, typically based on a shared encoder and multiple task-specific decoders, would significantly improve system efficiency, enable joint learning of diagnostic cues, and better reflect real-world diagnostic workflows. Furthermore, these models should be modular and extensible, supporting plug-and-play configurations for varied imaging modalities and clinical contexts [13], [30], [128].

Finally, the lack of standardized evaluation protocols and benchmarking practices poses a substantial barrier to progress. Current studies often employ disparate datasets, inconsistent evaluation metrics, and variable training-validation splits, which limits the reproducibility and comparability of results. The field would benefit greatly from the establishment of common benchmarking datasets with fixed training/testing splits and the inclusion of multi-task performance metrics that evaluate models holistically. The development of open-access leaderboards and community challenges—similar to those established for natural image classification and segmentation—could foster healthy competition, increase transparency, and drive methodological advancement [24], [26], [104], [129].

future research in deep learning for breast cancer imaging must not only pursue higher accuracy but also address fundamental issues related to data efficiency, interpretability, generalization, and deployment feasibility. Only by tackling these multidimensional challenges can researchers create AI systems that are not only powerful and accurate, but also trustworthy, efficient, and ready for real-world clinical impact.

Table 5 presents a consolidated summary of the key future research directions identified in this review, outlining their core objectives and associated references to guide ongoing advancements in deep learning-based breast cancer detection and segmentation, (Table 5).

Table 5. Future Research Directions in Deep Learning-Based Breast Cancer Detection and Segmentation

Research Direction	Description and Goals
Self-Supervised and Semi-Supervised Learning	Address data scarcity by enabling models to learn from unlabeled or partially labelled data. Techniques like contrastive learning, pseudo-labeling, and consistency training reduce reliance on expert annotations and improve generalization.
Integrated Explainable AI (XAI)	Improve clinical trust and regulatory approval through native interpretability mechanisms. Move beyond Grad-CAM to attention-based modules and decision-aware architectures that explain predictions in real time.

Research Direction	Description and Goals
Optimizing Multi-Task Learning (MTL)	Enhance learning of classification, detection, and segmentation in unified frameworks. Tackle task interference via dynamic loss weighting (e.g., GradNorm, ALB) and explore automated hyperparameter tuning for stable training.
Lightweight and Edge-Deployable Models	Design efficient architectures (e.g., MobileNet, EfficientNet-lite) with low computational overhead suitable for real-time inference on portable or embedded devices. Use pruning, quantization, and distillation for compression.
Multimodal and Clinical Metadata Integration	Fuse imaging data with patient metadata (e.g., age, family history, genetics) and radiomics features to enhance diagnostic accuracy and personalization. Apply cross-modal attention or graph-based fusion.
Domain Adaptation and Generalization	Develop models that generalize across institutions, scanners, and demographics using techniques like few-shot learning, CycleGANs, and unsupervised domain adaptation. Encourage cross-dataset evaluations.
Unified End-to-End Architectures	Integrate classification, detection, and segmentation into a single pipeline using shared encoders and task-specific decoders. Improve efficiency, coherence, and clinical usability.
Benchmarking and Evaluation Standardization	Establish shared benchmarks, multi-task leaderboards, and standardized datasets to promote transparency and reproducibility. Ensure fair comparison of models across different tasks and datasets.

10. CONCLUSION

This review provides a critical and multi-faceted synthesis of deep learning-based methodologies for breast cancer detection and segmentation, encompassing single-task, multi-task, and hybrid approaches across mammography, ultrasound, and MRI modalities. While significant progress has been made in algorithmic development—marked by the adoption of CNNs, transformers, attention mechanisms, and optimization techniques—several structural, methodological, and translational gaps remain unresolved.

One of the most pressing issues is the scarcity of large, annotated, and multi-task-compatible datasets, which limits both the training of robust models and the reproducibility of results. Despite the emergence of models with exceptional accuracy and segmentation performance, generalization across domains, modalities, and populations remains inadequate, frequently leading to performance degradation in real-world settings. Moreover, the widespread use of black-box architectures continues to challenge clinical trust, with many models lacking inherent interpretability or transparency in decision-making. Segmentation frameworks, although mature in architecture, often fail when faced with noisy, low-contrast, or irregular lesions—particularly in ultrasound imaging. Detection models frequently lack integrated localization capabilities or interpretability tools, reducing their diagnostic value. While end-to-end and multi-task learning frameworks offer a compelling route toward unified diagnostic pipelines, they are currently hindered by task interference, complex optimization demands, and insufficient benchmarking protocols. From a deployment perspective, current transformer-based and ensemble architectures, though accurate, are computationally expensive and unsuitable for point-of-care environments. Conversely, lightweight models often sacrifice accuracy or diagnostic completeness. Balancing computational feasibility with diagnostic precision is a critical yet underexplored challenge. Looking ahead, the field must prioritize several transformative directions: developing semi-supervised and self-supervised training pipelines to mitigate annotation bottlenecks; embedding explainability directly into model design to enhance transparency; optimizing multi-task architectures for stability and scalability; and integrating non-image clinical metadata through multimodal learning.

The future directions identified—such as the adoption of self-supervised learning for robust feature extraction from limited datasets, the integration of explainable AI for enhanced clinical trust, and the development of multi-modal fusion and domain adaptation techniques—are not merely incremental improvements. They are essential pathways for overcoming the existing barriers to create trustworthy, efficient, and clinically-impactful AI systems. By focusing on these areas, the next generation of deep learning models can be developed to provide a reliable and indispensable tool for breast cancer screening and diagnosis. In addition, a move toward standardized datasets, shared benchmarks, and clinically validated evaluation protocols is essential to promote reproducibility and fair comparison. Ultimately, the future of deep learning in breast cancer diagnosis depends not only on architectural innovation but also on cross-disciplinary collaboration—where AI researchers, radiologists, and clinical stakeholders co-develop solutions that are interpretable, generalizable, efficient, and ethically sound. Only then can deep learning systems transition from promising research tools to trusted, real-world clinical allies in the fight against breast cancer.

REFERENCES

[1] M., Leonardi, P., Martelletti, R., Burstein, A., Fornari, L., Grazzi, A., Guekht, and et al., "**The World Health Organization Intersectoral Global Action Plan on Epilepsy and Other Neurological Disorders and the Headache Revolution: From Headache Burden to A Global Action Plan for Headache Disorders**," *J. Headache Pain*, Vol. 25, No. 4, https://doi.org/10.1186/s10194-023-01700-3, 2024.

[2] H., Zhao, H., Li, and D., Cheng, Data "**augmentation for medical image analysis**," *Biomedical Image Synthesis and Simulation*, Elsevier, pp. 279–302, https://doi.org/10.1016/B978-0-12-824349-7.00021-9, 2022.

[3] G., Litjens, T., Kooi, B. E., Bejnordi, A. A. A., Setio, F., Ciompi, M., Ghafoorian, and et al., "**A Survey on Deep Learning in Medical Image Analysis,"** *Med Image Anal*, Vol. 42, https://doi.org/10.1016/j.media.2017.07.005, 2017.

[4] E., Abu Abeelh, and Z., AbuAbeileh, "**Comparative Effectiveness of Mammography, Ultrasound, and MRI in the Detection of Breast Carcinoma in Dense Breast Tissue: A Systematic Review**," *Cureus*, https://doi.org/10.7759/cureus.59054, 2024.

[5] H., Mahichi, V., Ghods, M. K., Sohrabi, and A., Sabbaghi, "**BreastCNet: Breast Cancer Detection, Classification, and Localization Convolutional Neural Network with Advanced Optimization Techniques**," *IEEE Access*, Vol. 13, pp. 87386–400, https://doi.org/10.1109/ACCESS.2025.3570364, 2025.

[6] P., Arbeláez, M., Maire, C., Fowlkes, and J., Malik, "**Contour Detection and Hierarchical Image Segmentation**," *IEEE Trans Pattern Anal Mach Intell*, Vol. 33, pp. 898–916. https://doi.org/10.1109/TPAMI.2010.161, 2011.

[7] O., Rainio, and R., Klén, "**Comparison of Simple Augmentation Transformations for A Convolutional Neural Network Classifying Medical Images**," *Signal Image Video Process*, Vol. 18, pp. 3353–60, https://doi.org/10.1007/s11760-024-02998-5, 2024.

[8] C., Aumente-Maestro, J., Díez, and B., Remeseiro, "**A multi-task framework for breast cancer segmentation and classification in ultrasound imaging**," *Comput Methods Programs Biomed*, Vol. 260, pp. 108540, https://doi.org/10.1016/j.cmpb.2024.108540, 2025.

[9] D. G. P., Petrini, C., Shimizu, R. A., Roela, G. V., Valente, M. A. A. K., Folgueira, and H. Y., Kim, "**Breast Cancer Diagnosis in Two-View Mammography Using End-to-End Trained EfficientNet-Based Convolutional Network**," *IEEE Access*, Vol. 10, pp. 77723–31, https://doi.org/10.1109/ACCESS.2022.3193250, 2022.

[10] M., Kaddes, Y. M., Ayid, A. M., Elshewey, and Y., Fouad, "**Breast Cancer Classification Based on Hybrid CNN with LSTM Model**," *Sci Rep*, Vol. 15, pp. 4409, https://doi.org/10.1038/s41598-025-88459-6, 2025.

[11] N., Zaidkilani, M. A., Garcia, and D., Puig, "**CoAtUNet: A symmetric encoder-decoder with hybrid transformers for semantic segmentation of breast ultrasound images**," *Neurocomputing*, Vol. 629, pp. 129660, https://doi.org/10.1016/j.neucom.2025.129660, 2025.

[12] J., Logan, P. J., Kennedy, and D., Catchpoole, "**A Review of The Machine Learning Datasets in Mammography, Their Adherence to The Fair Principles and The Outlook for The Future**," *Sci Data*, Vol. 10, pp. 595, https://doi.org/10.1038/s41597-023-02430-6, 2023.

[13] A., Anand, S., Jung, and S., Lee, "**Breast Lesion Detection for Ultrasound Images Using MaskFormer**," *Sensors*, Vol. 24, pp. 6890, https://doi.org/10.3390/s24216890, 2024.

[14] K., Yang, J., Bin, M. J., Lee, and J., Yang, "**Multi-Class Semantic Segmentation of Breast Tissues from MRI Images Using U-Net Based on Haar wavelet Pooling**," *Sci Rep*, Vol. 13, pp. 11704, https://doi.org/10.1038/s41598-023-38557-0, 2023.

[15] Y. A., Chen, L. J., Grimm, M., Nedrud, and H., Rahbar, "MRI Characteristics of Ductal Carcinoma in Situ," pp. 145–56, https://doi.org/10.1016/B978-0-12-822729-9.00026-6, 2022.

[16] Y., Le Cun, Y., Bengio, and G., Hinton, "**Deep learning**," *Nature*, Vol. 521, pp. 436–44, , 2015.

[17] N., Carion, F., Massa, G., Synnaeve, N., Usunier, A., Kirillov, and S., Zagoruyko, "**End-to-End Object Detection with Transformers,**" 2020.

[18] H., Chan, L. M., Hadjiiski, and R. K., Samala, "**Computer-Aided Diagnosis in The Era of Deep Learning**", *Med Phys*, Vol. 47, https://doi.org/10.1002/mp.13764, 2020.

[19] Y., Yu, J., Huang, L., Wang, S., Liang, "**A 1D-Inception-Resnet Based Global Detection Model for Thin-Skinned Multifruit Spectral Quantitative Analysis**," *Food Control*, Vol. 167, pp. 110823, https://doi.org/10.1016/j.foodcont.2024.110823, 2025.

[20] Y., Hou, Z., Wu, X., Cai, and T., Zhu, "**The Application of Improved Densenet Algorithm in Accurate Image Recognition**," *Sci Rep*, Vol. 14, pp. 8645, https://doi.org/10.1038/s41598-024-58421-z, 2024.

[21] A., Qayyum, M., Mazher, T., Khan, and I., Razzak, "**Semi-Supervised 3D-InceptionNet for Segmentation and Survival Prediction of Head and Neck Primary Cancers**," *Eng Appl Artif Intell*, Vol. 117, pp. 105590, https://doi.org/10.1016/j.engappai.2022.105590, 2023.

[22] L., Hamnett, M., Adewunmi, M., Abayomi, K., Raheem, and F., Ahmed, "**Enhancing Transformer-Based Segmentation for Breast Cancer Diagnosis using Auto-Augmentation and Search Optimisation Techniques**," 2023.

[23] F., Manigrasso, R.,, Milazzo, A. S.,, Russo, F., Lamberti, F., Strand, A., Pagnani, and et al., "**Mammography Classification with Multi-View Deep Learning Techniques: Investigating Graph and Transformer-Based Architectures**," *Med Image Anal*, Vol. 99, pp. 103320, https://doi.org/10.1016/j.media.2024.103320, 2025.

[24] M. R., Islam, M. M., Rahman, M. S., Ali, A. A. N., Nafi, M. S., Alam, T. K., Godder, and et al., "**Enhancing Breast Cancer Segmentation and Classification: An Ensemble Deep Convolutional Neural Network And U-Net Approach on Ultrasound Images**," *Machine Learning with Applications*, Vol. 16, pp. 100555, https://doi.org/10.1016/j.mlwa.2024.100555, 2024.

[25] O., Ronneberger, P., Fischer, and T., Brox, "**U-Net: Convolutional Networks for Biomedical Image Segmentation**," 2015.

[26] X., Wu, D., Hong, and J., Chanussot, "**UIU-Net: U-Net in U-Net for Infrared Small Object Detection**," *IEEE Transactions on Image Processing*, Vol. 32, pp. 364–76, https://doi.org/10.1109/TIP.2022.3228497, 2023.

[27] R., Maqsood, F., Abid, J., Rasheed, O., Osman, S., Alsubai, "**Optimal Res-UNET architecture with deep supervision for tumor segmentation**," *Front Med (Lausanne)*, Vol. 12, https://doi.org/10.3389/fmed.2025.1593016, 2025.

[28] C., Qin, Y., Wu, J., Zeng, L., Tian, Y., Zhai, F., Li, and et al., "**Joint Transformer and Multi-Scale CNN for DCE-MRI Breast Cancer Segmentation**," *Soft Comput*, Vol. 26, pp. 8317–34, https://doi.org/10.1007/s00500-022-07235-0, 2022.

[29] O., Ronneberger, P., Fischer, and T., Brox, "**U-Net: Convolutional Networks for Biomedical Image Segmentation**," 2015.

[30] X., Li, M., Li, P., Yan, G., Li, Y., Jiang, H, Luo, and et al., "**Deep Learning Attention Mechanism in Medical Image Analysis: Basics and Beyonds**," *International Journal of Network Dynamics and Intelligence*, pp. 93–116, https://doi.org/10.53941/ijndi0201006, 2023.

[31] S., Jamil, M. D., Jalil Piran, and O. J., Kwon, "**A Comprehensive Survey of Transformers for Computer Vision**," *Drones*, Vol. 7, pp. 287, https://doi.org/10.3390/drones7050287, 2023.

[32] J., Chen, J., Mei, X., Li, Y., Lu, Q., Yu, Q., Wei, and et al., "**Trans UNet: Rethinking the U-Net Architecture Design For Medical Image Segmentation Through the Lens of Transformers**," *Med Image Anal*, Vol. 97, pp. 103280, https://doi.org/10.1016/j.media.2024.103280, 2024.

[33] H., Afrin, N. B., Larson, M., Fatemi, and A., Alizad, "**Deep Learning in Different Ultrasound Methods for Breast Cancer, from Diagnosis to Prognosis: Current Trends, Challenges, and an Analysis**," *Cancers (Basel)*, Vol. 15, pp. 3139, https://doi.org/10.3390/cancers15123139, 2023.

[34] Q., He, Q., Yang, H., Su, and Y., Wang, "**Multi-Task Learning for Segmentation and Classification of Breast Tumors from Ultrasound Images,**" *Comput Biol Med*, Vol. 173, pp. 108319, https://doi.org/10.1016/j.compbiomed.2024.108319, 2024.

[35] Y., Zhou, H., Chen, Y., Li, Q., Liu, X., Xu, S., Wang, and et al., "Multi-task Learning for Segmentation and Classification of Tumors In 3d Automated Breast Ultrasound Images," *Med Image Anal*, Vol. 70, pp. 101918, https://doi.org/10.1016/j.media.2020.101918, 2021.

[36] A., Oliver, J., Freixenet, J., Martí, E., Pérez, J., Pont, E. R. E., Denton, and, et al., "**A Review of Automatic Mass Detection and Segmentation in Mammographic Images**," *Med Image Anal*, Vol. 14, https://doi.org/10.1016/j.media.2009.12.005, 2010.

[37] T., Shen, K., Hao, C., Gou, and F. Y., Wang, "Mass Image Synthesis in Mammogram with Contextual Information Based on GANs," *Comput Methods Programs Biomed*, Vol. 202, pp. 106019, https://doi.org/10.1016/j.cmpb.2021.106019, 2021.

[38] R. C., Joshi, D., Singh, V., Tiwari, and M. K., Dutta, "**An Efficient Deep Neural Network Based Abnormality Detection and Multi-Class Breast Tumor Classification**," *Multimed Tools Appl*, Vol. 81, pp. 13691–711,

[39] R., Maalej, and A., "**Mezghani, Transfer Learning and Data Augmentation for Improved Breast Cancer Histopathological Images Classifier**," *International Journal of Computer Information Systems and Industrial Management Applications*, Vol. 15, No. 10, 2023.

[40] A. M., Carrington, D. G., Manuel, P. W., Fieguth, T., Ramsay, V., Osmani, B., Wernly, and et al., "**Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation**," *IEEE Trans Pattern Anal Mach Intell*, Vol. 45, pp. 329–41, https://doi.org/10.1109/TPAMI.2022.3145392, 2023.

[41] M., Frid-Adar, I., Diamant, E., Klang, M., Amitai, J., Goldberger, and H., "**Greenspan, GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification**," *Neurocomputing*, Vol. 321, ppp 321–31. https://doi.org/10.1016/j.neucom.2018.09.013, 2018.

[42] S. D., S. M., S. K., S. S., and R. M., "**GAN Based Data Augmentation for Enhanced Tumor Classification. 2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)**," *IEEE*, pp. 1–5, https://doi.org/10.1109/ICCCSP49186.2020.9315189, 2020.

[43] C., Shorten, and T. M., Khoshgoftaar, "**A Survey on Image Data Augmentation for Deep Learning**," *J. Big Data*, Vol. 6, pp. 60, https://doi.org/10.1186/s40537-019-0197-0, 2019.

[44] C., Shorten, and T. M., Khoshgoftaar, "**A Survey on Image Data Augmentation for Deep Learning**," *J. Big Data*, Vol. 6, https://doi.org/10.1186/s40537-019-0197-0, 2019.

[45] W., Scharff, W., Rethfeldt, L., McNeilly, M., Laasonen, N., Meir, H., Abutbul-Oz, S., Smolander, and et al., "**Assessment of Developmental Language Disorder in Multilingual Children: Results from an International Survey**," *Folia Phoniatrica et Logopaedica*, Vol. 76, pp. 127–50, https://doi.org/10.1159/000533139, 2024.

[46] J., Rony, S., Belharbi, J., Dolz, I., Ben Ayed, L., McCaffrey, and E., Granger, "**Deep Weakly-Supervised Learning Methods for Classification and Localization in Histology Images: A Survey**," *Machine Learning for Biomedical Imaging*, Vol. 2, https://doi.org/10.59275/j.melba.2023-5g54, 2023.

[47] C. H., Wang, K. Y., Huang, Y., Yao, J. C., Chen, H. H., Shuai, and W. H., Cheng, "**Lightweight Deep Learning: An Overview**," *IEEE Consumer Electronics Magazine*, Vol. 13, pp. 51–64, https://doi.org/10.1109/MCE.2022.3181759, 2024.

[48] F. Z., Nakach, A., Idri, and E., Goceri, "**A Comprehensive Investigation of Multimodal Deep Learning Fusion Strategies for Breast Cancer Classification**," *Artif Intell Rev*, Vol. 57, pp. 327, 2024.

[49] A., Iqbal, and M., Sharif, "**BTS-ST: Swin Transformer Network for Segmentation and Classification of Multimodality Breast Cancer Images**," *Knowl Based Syst*, Vol. 267, pp. 110393, https://doi.org/10.1016/j.knosys.2023.110393, 2023.

[50] Yu, H., and Dai, Q., "**Self-Supervised Multi-Task Learning for Medical Image Analysis**," *Pattern Recognit*, Vol. 150, pp. 110327, https://doi.org/10.1016/j.patcog.2024.110327, 2024.

[51] O., Díaz, A., Rodríguez-Ruíz, and I., Sechopoulos, "**Artificial Intelligence for Breast Cancer Detection: Technology, Challenges, and Prospects**," *Eur J Radiol*, Vol. 175, pp. 111457, https://doi.org/10.1016/j.ejrad.2024.111457, 2024.

[52] H. S., Das, A., Das, A., Neog, S., Mallik, K., Bora, and Z., Zhao, "**Breast Cancer Detection: Shallow Convolutional Neural Network Against Deep Convolutional Neural Networks Based Approach**," *Front Genet*, Vol. 13, https://doi.org/10.3389/fgene.2022.1097207, 2023.

[53] M., Tan, R., Pang, and Q. V., Le, "**Efficient Det: Scalable and efficient object detection**," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020.

[54] A., Bochkovskiy, C. Y., Wang, and H. Y. M., Liao, "**YOLOv4: Optimal Speed and Accuracy of Object Detection**" 2020.

[55] A., Anand, S., Jung, and S., Lee, "**Breast Lesion Detection for Ultrasound Images Using MaskFormer**," *Sensors*, Vol. 24, pp. 6890. https://doi.org/10.3390/s24216890, 2024.

[56] N., Mikhailov, M., Shakeel, A., Urmanov, M. H., Lee, and M. F., "**Demirci, Optimization of CNN Model for Breast Cancer Classification**", *2021 16th International Conference on Electronics Computer and Computation (ICECCO), IEEE*, p p. 1–3, 2021.

[57] K. J., Geras, R. M., Mann, and L., Moy, "**Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives**," *Radiology*, Vol. 293, pp. 246–59, 2019.

[58] Y., Hou, H., Gao, Z., Wang, and C., Du, "**Improved Grey Wolf Optimization Algorithm and Application**", *Sensors*, Vol. 22, pp. 3810, https://doi.org/10.3390/s22103810, 2022.

[59] J., Lian, G., Hui, L., Ma, T., Zhu, X., Wu, A. A., Heidari, and et al., "**Parrot Optimizer: Algorithm and Applications to Medical Problems**," *Comput Biol Med*, Vol. 172, pp. 108064. https://doi.org/10.1016/j.compbiomed.2024.108064, 2024.

[60] D. G. P., Petrini, C., Shimizu, R. A., Roela, G. V., Valente, M. A. A. K., Folgueira, and H. Y., Kim, "**Breast Cancer Diagnosis in Two-View Mammography Using End-to-End Trained EfficientNet-Based Convolutional Network***," IEEE Access*, Vol. 10, pp. 77723–31, https://doi.org/10.1109/ACCESS.2022.3193250, 2022.

[61] H., Rahman, T. F., Naik Bukht, R., Ahmad, A., Almadhor, and A. R., Javed, "**Efficient Breast Cancer Diagnosis from Complex Mammographic Images Using Deep Convolutional Neural Network**," *Comput Intell Neurosci*, pp. 1–11.

[62] G., Ayana, and S., Choe, "**BUViTNet: Breast Ultrasound Detection via Vision Transformers**," *Diagnostics*, Vol. 12, pp. 2654. https://doi.org/10.3390/diagnostics12112654, 2022.

[63] F., Prinzi, M., Insalaco, A., Orlando, S., Gaglio, and S., Vitabile, "**A Yolo-Based Model for Breast Cancer Detection in Mammograms**," *Cognit Comput*, https://doi.org/10.1007/s12559-023-10189-6, 2023.

[64] W., Ding, J., Wang, W., Zhou, S., Zhou, C., Chang, and J., Shi, "**Joint Localization and Classification of Breast Cancer in B-Mode Ultrasound Imaging via Collaborative Learning with Elastography**," *IEEE J. Biomed Health Inform*, Vol. 26, pp. 4474–85. https://doi.org/10.1109/JBHI.2022.3186933, 2022.

[65] I. Z., Yao, M., Dong, and W. Y. K., Hwang, "**Deep Learning Applications in Clinical Cancer Detection: A Review of Implementation Challenges and Solutions**," *Mayo Clinic Proceedings: Digital Health*, Vol. 100253, https://doi.org/10.1016/j.mcpdig.2025.100253, 2025.

[66] R. L., Siegel, K. D., Miller, and A., Jemal, "**Cancer statistics**", *CA Cancer J. Clin*, Vol. 70, pp. 7–30, https://doi.org/10.3322/caac.21590, 2020.

[67] F., Bray, J., Ferlay, I., Soerjomataram, R. L., Siegel, L. A., Torre, and A., Jemal, "**Global Cancer Statistics***", Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries*, CA Cancer J Clin, Vol. 68, pp. 394–424, https://doi.org/10.3322/caac.21492, 2018.

[68] O., Ronneberger, P., Fischer, and T., Brox, "**U-net: Convolutional Networks for Biomedical Image Segmentation**," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, Munich, Germany, Proceedings, Part III, 18, Springer, pp. 234–41, 2015.

[69] J., Redmon, and A., Farhadi, "**YOLOv3: An Incremental Improvement**", 2018.

[70] K., He, G., Gkioxari, P., Dollar, and R., Girshick, "**Mask R-CNN**," *IEEE International Conference on Computer Vision (ICCV), IEEE*, 2017, pp. 2980–8, https://doi.org/10.1109/ICCV.2017.322, 2017.

[71] P., Arbeláez, M., Maire, C., Fowlkes, and J., Malik, "**Contour Detection and Hierarchical Image Segmentation**," IEEE Trans Pattern Anal Mach Intell, Vol. 33, pp. 898–916, https://doi.org/10.1109/TPAMI.2010.161, 2011.

[72] Y., Boykov, and G., Funka-Lea, "**Graph Cuts and Efficient N-D Image Segmentation***," Int J. Comput Vis*, Vol. 70, pp. 109–31, https://doi.org/10.1007/s11263-006-7934-5, 2006.

[73] L. C., Chen, G., Papandreou, I., Kokkinos, K., Murphy, and A. L., Yuille, "**Deep Lab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs**," *IEEE Trans Pattern Anal Mach Intell*, Vol. 40, pp. 834–48, https://doi.org/10.1109/TPAMI.2017.2699184, 2018.

[74] K., He, G., Gkioxari, P., Dollar, and R., Girshick, "**Mask R-CNN**," *IEEE International Conference on Computer Vision (ICCV), IEEE*, pp. 2980–8, https://doi.org/10.1109/ICCV.2017.322, 2017.

[75] M., Liu, D., Yao, Z., Liu, J., Guo, and J., Chen, "**An Improved Adam Optimization Algorithm Combining Adaptive Coefficients and Composite Gradients Based on Randomized Block Coordinate Descent**," *Comput Intell Neurosci*, https://doi.org/10.1155/2023/4765891, 2023.

[76] M. J., Umer, M. I., Sharif, and J., Kim, "**Breast Cancer Segmentation from Ultrasound Images Using Multiscale Cascaded Convolution with Residual Attention-Based Double Decoder Network**," *IEEE Access*, Vol. 12, pp. 107888–902, https://doi.org/10.1109/ACCESS.2024.3429386, 2024.

[77] S., Zhang, M., Liao, J., Wang, Y., Zhu, Y., Zhang, J., Zhang, and et al., "**Fully Automatic Tumor Segmentation of Breast Ultrasound Images with Deep Learning**," *J. Appl Clin Med Phys*, Vol. 24, https://doi.org/10.1002/acm2.13863, 2023.

[78] J., Redmon, S., Divvala, R., Girshick, and A., Farhadi, "**You Only Look Once: Unified, Real-Time Object Detection**," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/CVPR.2016.91, 2016.

[79] S., Goudarzi, J., Whyte, M., Boily, A., Towers, R. D., Kilgour, and H., Rivaz, "**Segmentation of Arm Ultrasound Images in Breast Cancer-Related Lymphedema: A Database and Deep Learning Algorithm**," *IEEE Trans Biomed Eng*, Vol. 70, pp. 2552–63, https://doi.org/10.1109/TBME.2023.3253646, 2023.

[80] F. A., Spanhol, L. S., Oliveira, C., Petitjean, and L., Heutte, "**A Dataset for Breast Cancer Histopathological Image Classification**," *IEEE Trans Biomed Eng*, Vol. 63, pp. 1455–62, https://doi.org/10.1109/TBME.2015.2496264, 2016.

[81] University of Nottingham. Nottingham Prognostic Index (NPI) dataset. https://www.nottingham.ac.uk/pathology/protocols/npi/npi.aspx. Accessed May 12, 2023.

[82] M., Tafavvoghi, L. A., Bongo, and N., Shvetsov, "**Busund L-TR, Møllersen K. Publicly available datasets of breast histopathology H & amp; E whole-slide images: A scoping review**," *J. Pathol Inform*, Vol. 15, pp. 100363, https://doi.org/10.1016/j.jpi.2024.100363, 2024.

[83] University of California Irvine Machine Learning Repository. Wisconsin Diagnostic Breast Cancer (WDBC) dataset. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). Accessed May 12, 2023.

[84] H., Soltani, M., Amroune, I., Bendib, M. Y., Haouam, E., Benkhelifa, and M. M., Fraz, "**Breast Lesions Segmentation and Classification in A Two-Stage Process Based on Mask-RCNN and Transfer Learning**," *Multimed Tools Appl*, pp. 1–18, 2023.

[85] F. S., Khan, M. N. H., Mohd, M. D., Khan, and S., Bagchi, "**Breast Cancer Histological Images Nuclei Segmentation using Mask Regional Convolutional Neural Network**," *IEEE Student Conference on Research and Development (SCOReD), IEEE*; pp. 1–6, https://doi.org/10.1109/SCOReD50371.2020.9383186, 2020.

[86] H. M. A., Bhatti, J., Li, S., Siddeeq, A., Rehman, and A., Manzoor, "**Multi-Detection and Segmentation of Breast Lesions Based on Mask RCNN-FPN. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),"** *IEEE*, pp. 2698–704, https://doi.org/10.1109/BIBM49941.2020.9313170, 2020.

[87] Y., Zhang, S., Chan, V. Y., Park, K. T., Chang, S., Mehta, M. J., Kim, and et al., "**Automatic Detection and Segmentation of Breast Cancer on MRI Using Mask R-CNN Trained on Non–Fat-Sat Images and Tested on Fat-Sat Images**," *Acad Radiol*, Vol. 29, pp. 135–44. https://doi.org/10.1016/j.acra.2020.12.001, 2022.

[88] N., Karunanayake, and S. S., Makhanov, "**When Deep Learning Is Not Enough: Artificial Life as A Supplementary Tool for Segmentation of Ultrasound Images of Breast Cancer**," *Med Biol Eng Comput*, 2024.

[89] D., Abdelhafiz, J., Bi, R., Ammar, C., Yang, and S., Nabavi, "**Convolutional Neural Network for Automated Mass Segmentation in Mammography**," *BMC Bioinformatics*, Vol. 21, pp. 192, 2020.

[90] A., Baccouche, B., Garcia-Zapirain, C., Castillo Olea, and A. S., Elmaghraby, "**Connected-UNets: A Deep Learning Architecture for Breast Mass Segmentation**," *NPJ Breast Cancer*, Vol. 7, pp. 151, 2021.

[91] T., Siriapisith, W., Kusakunniran, and P., Haddawy, "**Pyramid graph cut: Integrating intensity and gradient information for grayscale medical image segmentation**," *Comput Biol Med*, Vol. 126, pp. 103997, https://doi.org/10.1016/j.compbiomed.2020.103997, 2020.

[92] H., Chen, X., Qi, L., Yu, and P. A., Heng, "**DCAN: Deep Contour-Aware Networks for Accurate Gland Segmentation**," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *IEEE*, pp. 2487–96, 3, 2016.

[93] P., Coupeau, J. B., Fasquel, and M., "**Dinomais, On the use of GNN-Based Structural Information to Improve CNN-Based Semantic Image Segmentation,**" *J. Vis Commun Image Represent*, Vol. 101, pp. 104167, 2024.

[94] Y., Gao, B., Liu, Y., Zhu, L., Chen, M., Tan, X., Xiao, and et al., "**Detection and Recognition of Ultrasound Breast Nodules Based on Semi-Supervised Deep Learning: A Powerful Alternative Strategy**," *Quant Imaging Med Surg*, Vol. 11, pp. 2265–78, https://doi.org/10.21037/qims-20-12B, 2021.

[95] J., Dafni Rose, K., VijayaKumar, L., Singh, and S. K., Sharma, "**Computer-Aided Diagnosis for Breast Cancer Detection and Classification Using Optimal Region Growing Segmentation with Mobile Net Model**," *Concurrent Engineering*, Vol. 30, pp. 181–9. https://doi.org/10.1177/1063293X221080518, 2022.

[96] C., Aumente-Maestro, J., Díez, and B., Remeseiro, "**A Multi-Task Framework for Breast Cancer Segmentation and Classification in Ultrasound Imaging**," *Comput Methods Programs Biomed*, Vol. 260, pp. 108540, https://doi.org/10.1016/j.cmpb.2024.108540, 2025.

[97] C., Aumente-Maestro, J., Díez, and B., Remeseiro, "**A Multi-Task Framework for Breast Cancer Segmentation and Classification in Ultrasound Imaging**," *Comput Methods Programs Biomed*, Vol. 260, pp. 108540, https://doi.org/10.1016/j.cmpb.2024.108540, 2025.

[98] Z., Cheng, M., Liu, C., Yan, and S., Wang, "**Dynamic Domain Generalization for Medical Image Segmentation**," *Neural Networks*, Vol. 184, pp. 107073, https://doi.org/10.1016/j.neunet.2024.107073, 2025.

[99] M., Aslam, "**Cochran's Q Test for Analyzing Categorical Data Under Uncertainty**," **J. Big Data**, Vol. 10, pp. 147.

[100] Y., Lecun, L., Bottou, Y., Bengio, and P., Haffner, "**Gradient-Based Learning Applied to Document Recognition**," *Proceedings of the IEEE*, Vol. 86, pp. 2278–324, https://doi.org/10.1109/5.726791, 1998.

[101] A. A., Hekal, A., Elnakib, H. E. D., Moustafa, and H. M., Amer, "**Breast Cancer Segmentation from Ultrasound Images Using Deep Dual-Decoder Technology with Attention Network**," *IEEE Access*, Vol. 12, pp. 10087–101, https://doi.org/10.1109/ACCESS.2024.3351564, 2024.

[102] T., Alam, W. C., Shia, F. R., Hsu, and T., Hassan, "**Improving Breast Cancer Detection and Diagnosis through Semantic Segmentation Using the Unet3+ Deep Learning Framework**," *Biomedicines*, Vol. 11, pp. 1536, https://doi.org/10.3390/biomedicines11061536, 2023.

[103] D., Muduli, R., Dash, and B., Majhi, "**Automated Diagnosis of Breast Cancer Using Multi-Modal Datasets: A Deep Convolution Neural Network Based Approach**," *Biomed Signal Process Control*, Vol. 71, pp. 102825, https://doi.org/10.1016/j.bspc.2021.102825, 2022.

[104] C., Thomas, M., Byra, R., Marti, M. H., Yap, and R., Zwiggelaar, "**Bus-Set: A Benchmark for Quantitative Evaluation of Breast Ultrasound Segmentation Networks with Public Datasets**," *Med Phys*, 2023.

[105] A., Das, and S., Rana, "**Exploring Residual Networks for Breast Cancer Detection from Ultrasound Images**" *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, *IEEE*, pp. 1–6, https://doi.org/10.1109/ICCCNT51525.2021.9580160, 2021.

[106] H., Huynh, A. T., Tran, and T. N., "**Tran, Region-of-Interest Optimization for Deep-Learning-Based Breast Cancer Detection in Mammograms**," *Applied Sciences*, Vol. 13, pp. 6894, https://doi.org/10.3390/app13126894, 2023.

[107] A., Sahu, P. K., Das, and S., Meher, "**An Automatic Sparse-Based Deep Cascade Framework with Multilayer Representation for Detecting Breast Cancer**," *Measurement*, Vol. 228, pp. 114375, 2024.

[108] P., Kumar, S., Srivastava, R. K., Mishra, and Y. P., Sai, "**End-to-end improved convolutional neural network model for breast cancer detection using mammographic data**," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, Vol. 19, pp. 375–84, https://doi.org/10.1177/1548512920973268, 2022.

[109] A. R. W., Sait, and R., Nagaraj, "**An Enhanced Light GBM-Based Breast Cancer Detection Technique Using Mammography Images**," Diagnostics, Vol. 14, pp. 227, https://doi.org/10.3390/diagnostics14020227, 2024.

[110] A. K., Mishra, P., Roy, S., Bandyopadhyay, and S. K., Das, "**A multi-task learning based approach for efficient breast cancer detection and classification**," Expert Syst, Vol. 39, https://doi.org/10.1111/exsy.13047, 2022.

[111] W. D., Vogl, K., Pinker, T. H., Helbich, H., Bickel, G., Grabner, W., Bogner, and et al., "**Automatic Segmentation and Classification of Breast Lesions Through Identification of Informative Multiparametric PET/MRI Features**," *Eur Radiol Exp*, Vol. 3, pp. 18, https://doi.org/10.1186/s41747-019-0096-3, 2019.

[112] A., Vakanski, M., Xian, and P. E., Freer, "**Attention-Enriched Deep Learning Model for Breast Tumor Segmentation in Ultrasound Images**," *Ultrasound Med Biol*, Vol. 46, pp. 2819–33, , 2020.

[113] R., Irfan, A. A., Almazroi, H. T., Rauf, R., Damaševičius, E. A., Nasr, and A. E., Abdelgawad, "**Dilated Semantic Segmentation for Breast Ultrasonic Lesion Detection Using Parallel Feature Fusion**," *Diagnostics*, Vol. 11, pp. 1212, https://doi.org/10.3390/diagnostics11071212, 2021.

[114] T., Zhao, and H., Dai, "**Breast Tumor Ultrasound Image Segmentation Method Based on Improved Residual U-Net Network**", *Comput Intell Neurosci*, pp. 1–9, https://doi.org/10.1155/2022/3905998, 2022.

[115] M., Bobowicz, M., Badocha, K., Gwozdziewicz, M., Rygusik, P., Kalinowska, E., Szurowska, and et al., "**Segmentation-based BI-RADS ensemble classification of breast tumours in ultrasound images**," *Int J Med Inform*, Vol. 189, pp. 105522, https://doi.org/10.1016/j.ijmedinf.2024.105522, 2024.

[116] P., Hurtik, V., Molek, J., Hula, M., Vajgl, P., Vlasanek, and T., Nejezchleba, "**Poly-YOLO: Higher Speed, More Precise Detection and Instance Segmentation for YOLOv3**," 2020.

[117] Y., Ma, and Y., Peng, "**Mammogram Mass Segmentation and Classification Based on Cross-View VAE and Spatial Hidden Factor Disentanglement**," *Phys Eng Sci Med*, Vol. 47, pp. 223–38, , 2024.

[118] A., Carriero, L., Groenhoff, E., Vologina, P., Basile, and M., Albera, "**Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements in Early 2024**," *Diagnostics*, Vol. 14, pp. 848, https://doi.org/10.3390/diagnostics14080848, 2024.

[119] P., Pramanik, A., Roy, E., Cuevas, M., Perez-Cisneros, and R., Sarkar, "**DAU-Net: Dual Attention-Aided U-Net for Segmenting Tumor in Breast Ultrasound Images**," *PLoS One*, Vol. 19, pp. 0303670, 2024.

[120] G., Madhu, A., Meher Bonasi, S., Kautish, A. S., Almazyad, A. W., Mohamed, F., Werner, and et al., "**U Caps Net: A Two-Stage Deep Learning Model Using U-Net and Capsule Network for Breast Cancer Segmentation and Classification in Ultrasound Imaging**," *Cancers (Basel)*, Vol. 16, pp. 3777, https://doi.org/10.3390/cancers16223777, 2024.

[121] S., Schutte, and J., Uddin, "**Deep Segmentation Techniques for Breast Cancer Diagnosis,**" *Bio Med Informatics*, Vol. 4, pp. 921–45, https://doi.org/10.3390/biomedinformatics4020052, 2024.

[122] H., Afrin, N. B., Larson, M., Fatemi, and A., Alizad, "**Deep Learning in Different Ultrasound Methods for Breast Cancer, from Diagnosis to Prognosis: Current Trends, Challenges, and an Analysis**," *Cancers (Basel)*, Vol. 15, pp. 3139, https://doi.org/10.3390/cancers15123139, 2023.

[123] Z., Huang, X., Zhang, Y., Ju, G., Zhang, W., Chang, H., Song, and et al., "**Explainable breast cancer molecular expression prediction using multi-task deep-learning based on 3D whole breast ultrasound**," *Insights Imaging*, Vol. 15, pp. 227, https://doi.org/10.1186/s13244-024-01810-9, 2024.

[124] J., Peta, and S., Koppu, "**Explainable Soft Attentive Efficient Net for Breast Cancer Classification in Histopathological Images**," *Biomed Signal Process Control*, Vol. 90, pp. 105828, https://doi.org/10.1016/j.bspc.2023.105828, 2024.

[125] M., Frid-Adar, E., Klang, M., Amitai, J., Goldberger, and H., Greenspan, "**Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification**," 2018.

[126] K., He, G., Gkioxari, P., Dollár, and R., Girshick, "**Mask R-CNN**," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–9, 2017.

[127] Y., Zhang, S., Chan, V. Y., Park, K. T., Chang, S., Mehta, M. J., Kim, and et al., "**Automatic Detection and Segmentation of Breast Cancer on MRI Using Mask R-CNN Trained on Non–Fat-Sat Images and Tested on Fat-Sat Images**," *Acad Radiol*, Vol. 29, pp. 135–44. https://doi.org/10.1016/j.acra.2020.12.001, 2022.

[128] R., Irfan, A. A., Almazroi, H. T., Rauf, R., Damaševičius, E. A., Nasr, and A. E., "**Abdelgawad, Dilated Semantic Segmentation for Breast Ultrasonic Lesion Detection Using Parallel Feature Fusion,**" *Diagnostics*, Vol. 11, pp. 1212, https://doi.org/10.3390/diagnostics11071212, 2021.

[129] F., Di Salvo, S., Doerrich, and C., Ledig, "**MedMNIST-C: Comprehensive Benchmark and Improved Classifier Robustness by Simulating Realistic Image Corruptions**," 2024.

Implementation of the Constrained Least Mean Squares (LMS) algorithm for Beamforming

Farhad Bahadori Jahromi

Department of Electrical Engineering, Fa.C., Islamic Azad University, Fasa, Iran.
E-mail: bahadori.fr@gmail.com (Corresponding author)

ABSTRACT:
A Beamformer is an array of sensors which can do spatial filtering. The objective is to estimate the signal arriving from the desired direction in the presence of noise and other interfering signals. A beamformer does spatial filtering in the sense that it separates two signals with overlapping frequency content originating from different directions. The aim of the paper was to study the different beamforming techniques and use the Constrained Least Mean Squares (LMS) filter for spatial filtering. An array of microphones was simulated in MATLAB and a simple delay and sum beamformer was implemented. The results were compared with that of a single microphone and it was observed that beamforming definitely gives a significant SNR improvement. A Constrained least mean square algorithm (also known as Frost Beamformer) was derived which is capable of iteratively adapting the weights of the sensor array to minimize noise power at the array output while maintaining a chosen frequency response in the look direction. The adaptive version of the Frost beamformer was simulated in MATLAB and it was observed that there was a significant improvement in the SNR as compared to the simple delay and sum beamformer.

KEYWORDS: Least Mean Squares (LMS), Signal-to-Noise Ratio (SNR), Adaptive Frost Beamformer (AFB).

9. 1. INTRODUCTION

Spatially propagating signals encounter the presence of interfering signals and noise signals. If the desired signal and the interferers occupy the same temporal frequency band, then temporal filtering cannot be used to separate the signal from the interferers. However, the desired and the interfering signals generally originate from different spatial locations. This spatial separation can be exploited to separate the signals from the interference using a beamformer. A beamformer consists of an array of sensors in a particular configuration. The output of a sensor is properly filtered, and the filtered outputs of all the sensors are added up. Typically, a beamformer linearly combines the spatially sampled waveform from each sensor in the same way a FIR filter linearly combines temporally sampled data. When low frequency signals are used, an array of sensors can synthesize a much larger spatial aperture than that practical with a single physical antenna. A second very significant advantage of using an array of sensors is the spatial filtering versatility offered by discrete sampling. In many applications it is necessary to change the spatial filtering function in real time to maintain effective suppression of interfering signals. Changing the spatial filtering function of a continuous aperture antenna is impractical. Typical uses of beamforming arise in RADAR, SONAR, communications, imaging, Geophysical exploration, Biomedical and also in acoustic source localization.

LMS algorithm is known for its simplicity and robustness. The computational complexity of LMS algorithm is O(M) [1]-[3]. While it lacks in convergence speed, several modifications to the algorithm are proposed including Optimized-LMS [4], [5] Variable Step Size LMS (VSS-LMS) algorithms [5]-[7], variable-length LMS algorithm [8], [9], transform domain algorithms [10]-[12], and recently CSLMS algorithm [13]-[15].

Paper type: Research paper
https://doi.org/xxx
Received: 15 January 2025, Revised: 17 February 2025, Accepted: 2 March 2025, Published: 1 June 2025
How to cite this paper: F. Bahadori Jahromi, **"Implementation of the Constrained Least Mean Squares (LMS) algorithm for Beamforming"**, *Majlesi Journal of Telecommunication Devices*, Vol. 14, No. 2, pp. 131-147, 2025.

10. 2. BEAMFORMER CLASSIFICATION

Beam formers are classified as either data independent or statistically optimum, depending on how the weights are chosen. The weights in a data-independent beamformer do not depend on the array data and are chosen to present a specified response for all signal and interference scenarios. The weights in a statistically optimum beamformer are chosen based on the statistics of the array data to optimize the array response. The statistics of the array data are not usually known and may change over time, so adaptive algorithms are typically used to determine the weights. The adaptive algorithm is designed so that the beamformer response converges to a statistically optimum solution.

The weights in a data-independent beamformer are designed so that the beamformer response approximates a desired response independent of the array data or data statistics. This design objective is the same as that for a classical FIR filter design. The simple Delay and sum beamformer is an example of data independent beamforming.

In a statistically optimum beamformer the weighs are chosen based on the statistics of the data received at the array. The goal is to optimize the beamformer response so that the output signal contains minimal contributions due to the noise and signals arriving from directions other than the desired direction. The Frost beamformer is a statistically optimum beamformer. Other statistically optimum beamformers are Multiple Side lobe Canceller and Maximization of the signal to noise ratio.

11. 3. DELAY AND SUM BEAMFORMER

The underlying idea of sum-and-delay beamforming is that when an electromagnetic signal impinges upon the aperture of the antenna array, the element outputs, added together with appropriate amounts of delays, reinforce signals with respect to noise or signals arriving at different directions. The delays required depend on the physical spacing between the elements in the array. The geometrical arrangement of elements and weights associated with each element are crucial factors in defining the array's characteristics.

In delay-and-sum beamforming, delays are inserted after each microphone to compensate for the arrival time differences of the speech signal to each microphone (Fig. 1). The time-aligned signals at the outputs of the delays are then summed together. This has the effect of reinforcing the desired speech signal while the unwanted off-axis noise signals are combined in a more unpredictable fashion. The signal-to-noise ratio (SNR) of the total signal is greater than (or at worst, equal to) that of any individual microphone's signal. This system makes the array pattern more sensitive to sources from a particular desired direction.

The major disadvantage of delay-and-sum beamforming systems is the large number of sensors required to improve the SNR. Each doubling of the number of sensors will provide at most an additional 3 dB increase in SNR, and this is if the incoming jamming signals are completely uncorrelated between the sensors and with the desired signal. Another disadvantage is that no nulls are placed directly in jamming signal locations. The delay-and-sum beamformer seeks only to enhance the signal in the direction to which the array is currently steered.

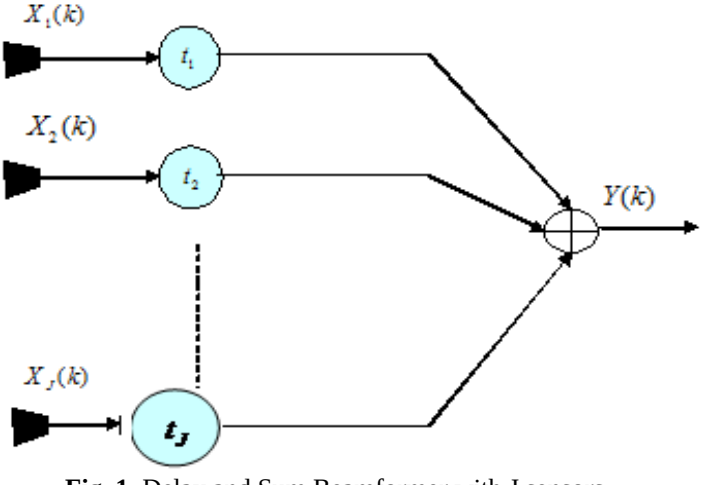


Fig. 1. Delay and Sum Beamformer with J sensors.

12. 4. FROST BEAMFORMER

The Constrained Least Mean Squares or Constrained LMS algorithm is a simple stochastic gradient algorithm which requires only the direction of arrival and the desired frequency response in the look direction. In the adaptive process, the algorithm progressively learns statistics of noise arriving from directions other than the look direction. The algorithm

is able to maintain a chosen frequency response in the look direction while minimising output noise power.

Consider the array processor shown in Fig. 2. The processor has K sensors and J taps per sensor. So, there are KJ weights. Out of these J weights determine the look direction frequency response. In the figure, the delays after each sensor are not shown. The array processor is assumed to be steered to the required look direction by appropriate delays after the sensors, as in the case of Delay and Sum beamforming. The remaining KJ – J weights may be used to minimize the total power in the array output. Minimization of the total output power is equivalent to minimizing the non-look direction noise power as long as the signal and the noise are uncorrelated, which is a reasonable assumption.

As far as the signal is concerned, the array processor is equivalent to a single-tapped delay in which each weight is equal to the sum of the weights in the vertical column of the processor. These summation weights in the equivalent tapped delay line must be selected so as to give the desired frequency response characteristic in the look direction.

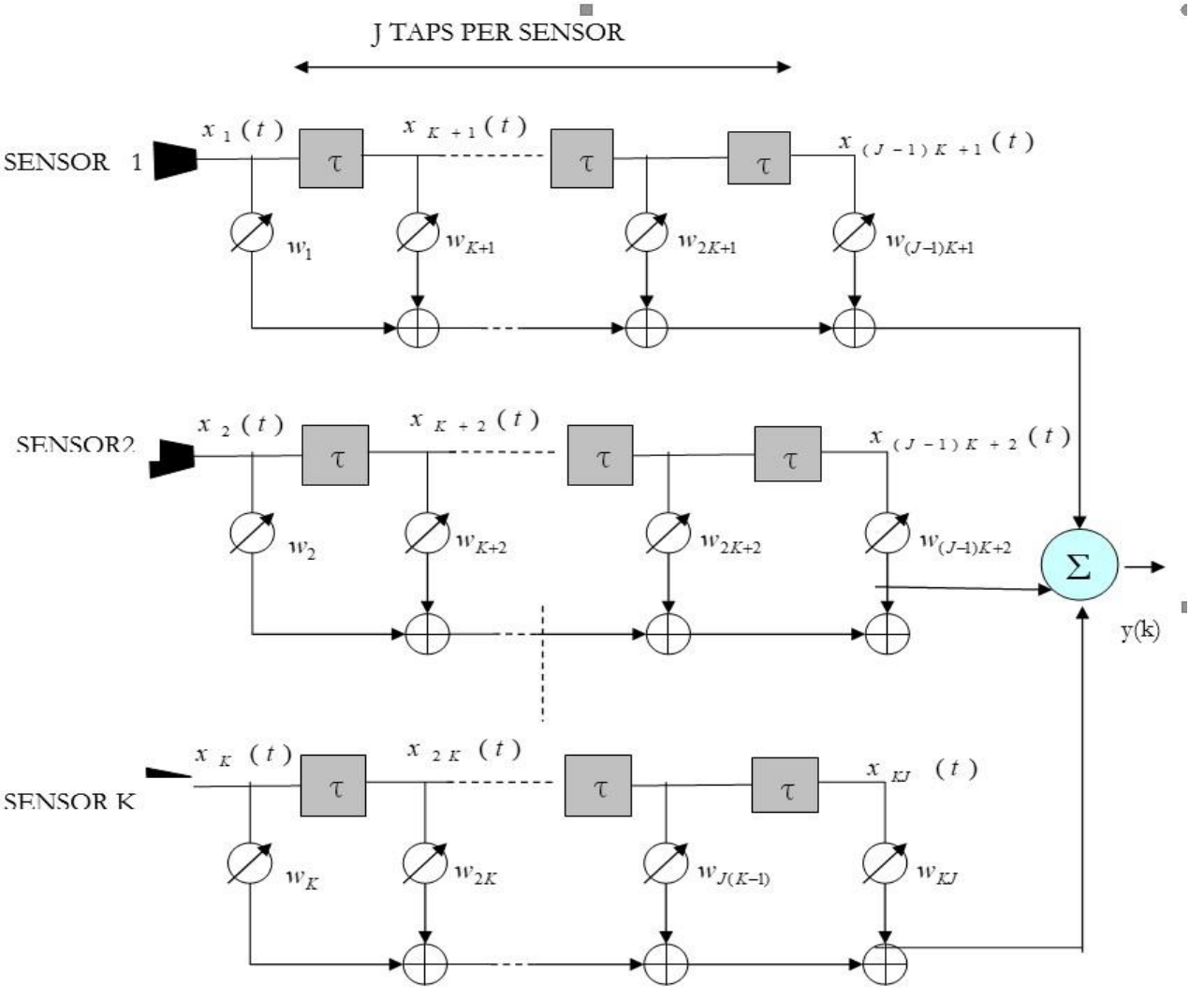


Fig 2. Frost Beamformer.

The vector of tap voltages at the kth sample is written as $X(k)$ where:

$$X^T(k) = \left[x_1(k),\ x_2(k), \ldots\ldots\ldots, x_{KJ}(k) \right]$$

The tap voltages are the sums of the voltages due to look-direction waveforms and the non-look-direction noises, so that:

$$X(k) = L(k) + N(k)$$

Where the KJ dimensional vector of look-direction at the kth sample is:

$$L^T(k) = [\underbrace{l(k) \quad . \quad . \quad l(k)}_{\text{K taps}} \quad \underbrace{l(k-1) \quad . \quad . \quad l(k-1)}_{\text{K taps}} \quad . \quad . \quad \underbrace{l(k-(J-1)) \quad . \quad . \quad l(k-(J-1))}_{\text{K taps}}]$$

And the vector of non-look-direction noises is:

$$N^T(k) = [n_1(k) \quad n_2(k) \quad . \quad . \quad . \quad n_{KJ}(k)]$$

The vector of weights at each tap is W, where:

$$W^T = [w_1 \quad w_2 \quad . \quad . \quad . \quad w_{KJ}]$$

$$E[X(k)X^T(k)] = R_{XX}$$
$$E[N(k)N^T(k)] = R_{NN}$$
$$E[L(k)L^T(k)] = R_{LL}$$

It is assumed that the look direction waveform is uncorrelated with the vector of non-look direction noise:

$$E[N(k)L^T(k)] = 0$$

The output of the array at the time of the kth sample is:

$$y(k) = W^T X(k) = X^T(k)W$$

The expected output power of the array is:

$$E[y^2(k)] = E[W^T X(k) X^T(k) W]$$

$$E[y^2(k)] = W^T R_{XX} W$$

The constraints that the weights on the jth vertical column of the taps sum to a chosen number f_i are expressed by the requirement:

$$c_j^T W = f_i \qquad j = 1, 2, ..., J$$

Where the KJ dimensional vector has the form:

$$c^T{}_j = [\underbrace{0 \quad . \quad . \quad 0}_{\text{K}} \quad . \quad . \quad \underbrace{0 \quad . \quad . \quad 0}_{\text{K}} \quad \underbrace{1 \quad . \quad . \quad 1}_{\text{jth group of K elements}} \quad 0 \quad . \quad . \quad 0 \quad . \quad . \quad . \quad 0 \quad . \quad . \quad 0]$$

Define the constraint matrix C as

$$C = \underbrace{\left[c_1 \quad . \quad . \quad c_j}_{J} \quad \underbrace{. \quad . \quad . \quad c_J\right]}_{KJ}$$

Define F as the J dimensional vector of weights of the look-direction-equivalent tapped delay line:

$$F^T = \left[f_1 \quad . \quad . \quad f_j \quad . \quad . \quad f_J\right]$$

The constraint can now be written as

$$C^T W = F$$

So the problem can be summarized as:

$\min_W \; W^T R_{XX} W$ This is the constrained LMS problem.

$subject\ to\ C^T W$

W_{opt} Is found by the method of Lagrange multipliers:

$$H(W) = \frac{1}{2} W^T R_{XX} W + \lambda^T (C^T W - F)$$

Taking the gradient with respect to W.

$$\nabla_W H(W) = R_{XX} W + C\lambda$$

Setting this to zero:

$$\nabla_W H(W) = R_{XX} W + C\lambda = 0$$

$$W_{opt} = -R_{XX}{}^{-1} C\lambda$$

Since R_{XX} is positive semi-definite, the inverse exists.
Substituting this in the constraint equation:

$$C^T W_{opt} = F = -C^T R_{XX}{}^{-1} C\lambda$$

The Lagrange multipliers can be found as:

$$\lambda = -\left[C^T R_{XX}{}^{-1} C\right]^{-1} F$$

Therefore, the optimum weight vector can be written as:

$$\boxed{W_{opt} = R_{XX}{}^{-1} C\left[C^T R_{XX}{}^{-1} C\right]^{-1} F}$$

13. 5. ADAPTIVE ALGORITHM FOR FROST BEAMFORMER

To find the optimum weights, the input correlation matrix R_{XX} is not known a priori and must be learnt by an adaptive technique. Direct substitution of a correlation matrix estimate into the optimal weight equation requires a number of multiplications at each iteration proportional to the cube of the number of weights. The complexity is due to

the inversion of the input correlation matrix. The adaptive algorithm described below requires only a number of multiplications and storage locations directly proportional to the number of weights.

In constrained gradient-descent optimization, the weight vector is initialized at a vector satisfying the constraint say, $W(0) = C(C^T C)^{-1} F$ and at each iteration, the weight vector is moved in the negative direction of the constrained gradient. The length of the step is proportional to the magnitude of the constrained gradient and is scaled by a constant μ. After the kth iteration, the next weight vector is:

$$\begin{aligned} W(k+1) &= W(k) - \mu \nabla_w H[W(k)] \\ &= W(k) - \mu[R_{XX} W(k) + C\lambda(k)] \end{aligned}$$

The Lagrange multipliers are chosen by requiring W(k+1) to satisfy the constraint:

$$F = C^T W(k+1) = C^T W(k) - \mu C^T R_{XX} W(k) - \mu\, C^T C\lambda(k)$$

Solving for the Lagrange multipliers λ(k) and substituting into the weight-iteration equation we have

$$W(k+1) = W(k) - \mu\left[I - C(C^T C)^{-1} C^T\right] R_{XX} W(k) + C(C^T C)^{-1}\left[F - C^T W(k)\right]$$

Defining the KJ dimensional vector:

$$\tilde{F} = C(C^T C)^{-1} F$$

and the KJ x KJ matrix:

$$P = I - C(C^T C)^{-1} C^T$$

The algorithm may be written as:

$$W(k+1) = P[W(k) - \mu R_{XX} W(k)] + \tilde{F}$$

A simple approximation for at the kth iteration is the outer product of the tap voltage vector with itself:
The stochastic Constrained LMS algorithm is:

$$W(0) = \tilde{F}$$

$$W(k+1) = P[W(k) - \mu\, y(k) X(k)] + \tilde{F}$$

14. 6. SIMULATION SETUP

The beam former was simulated in MATLAB for 3 cases:

1. A single microphone
2. A sum and Delay Beamformer
3. Adaptive Frost Beamformer

The Beamformer had 6 sensors placed in a linear array with the distance between the sensors d=0.5 m. For the Adaptive Frost Beamformer each sensor branch had 6 taps. The environment consisted of 2 noise sources and one desired source. Fig. 3 shows the convention followed for the angles (Table 1).

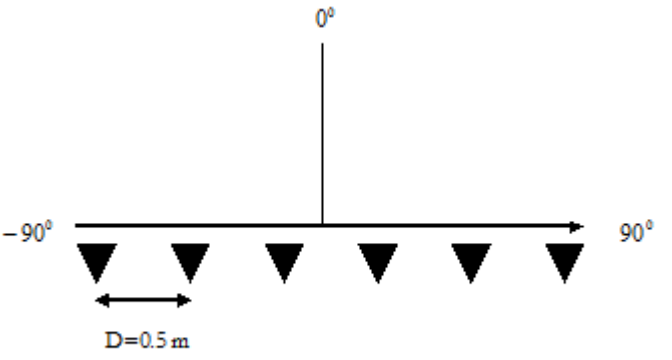


Fig. 3. Convention followed for the angles.

Table 1. The simulation was done for the following 2 environments

	Desired Signal 0 degrees	Interfering Signal 1 45 degrees	Interfering Signal 2 -45 degrees
Simulation 1	Sine of 2.5Khz	Gaussian Noise 1	Gaussian noise 2
Simulation 2	Speech signal 2	Gaussian Noise 1	Gaussian noise 2
Simulation 3	Speech signal 1	Speech signal 2	Gaussian noise 1

	Desired Signal 60 degrees	Interfering Signal 1 0 degrees	Interfering Signal 2 -60 degrees
Simulation 4	Sine of 2.5Khz	Gaussian Noise 1	Gaussian noise 2
Simulation 5	Speech signal 2	Gaussian Noise 1	Gaussian noise 2
Simulation 6	Speech signal 1	Speech signal 2	Gaussian noise 1

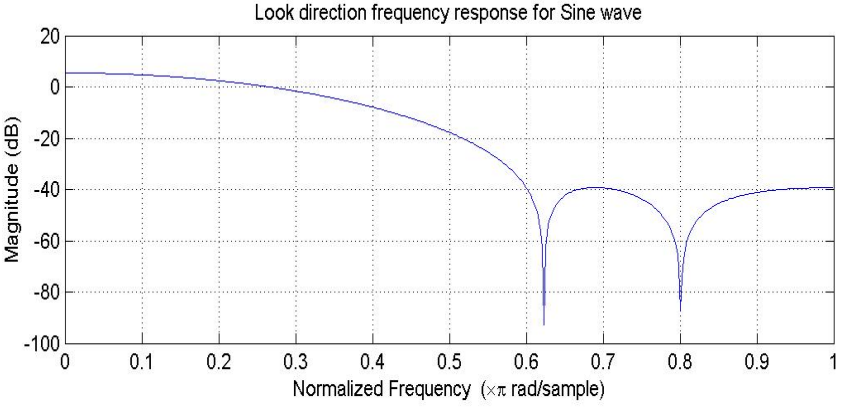
The speech signal 1, speech signal 2 and the sine wave were sampled at 22.05 KHz. The power of the sine wave was 1. The Gaussian noise was generated using pseudo-Gaussian generators. The power of the Gaussian noise was twice that of the desired signals. The Gaussian noise was filtered to lie in non overlapping bands. Note that there is frequency overlap between the desired signal and the interferers.

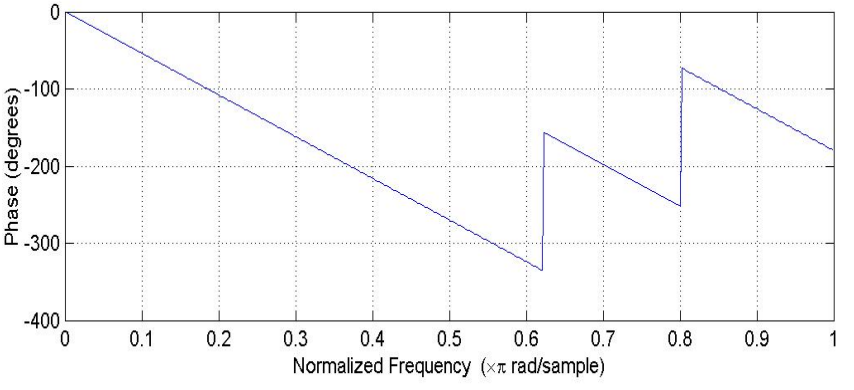
Table 2 summarizes the noises and sources used in the simulation.

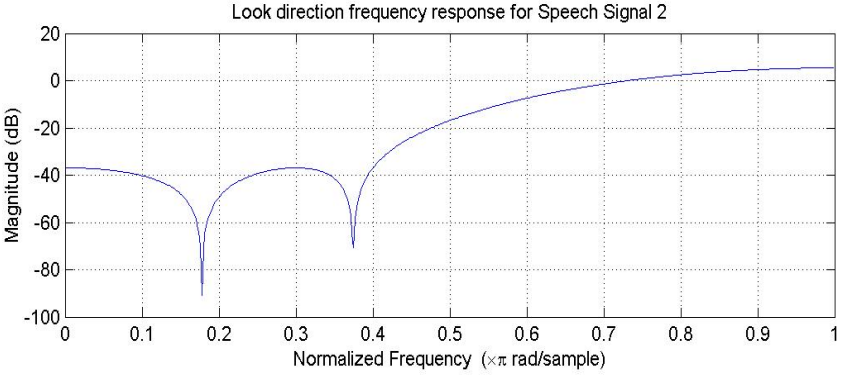
Table 2. The summary of the noises and sources used in the simulation

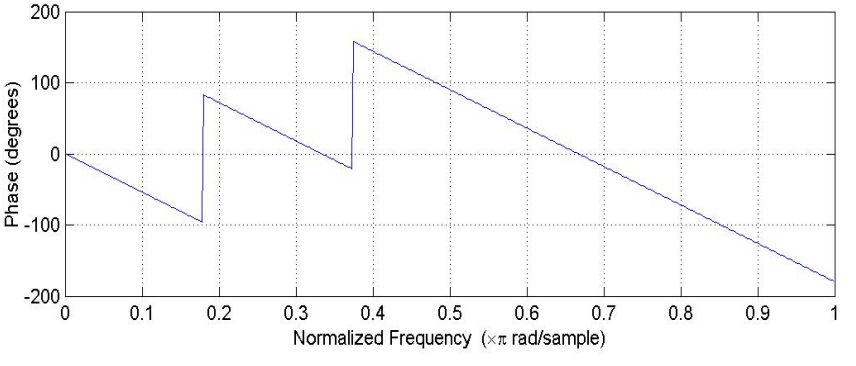
	Sampling rate kHz	Center Frequency kHz	Bandwidth kHz	Power
Gaussian noise 1	22.05	3.5	7	2
Gaussian noise 2	22.05	14.5	7	2
Sine wave	22.05	5		1
Speech signal 1	22.05	3.3	7	1
Speech signal 2	22.05	7.35	7	1

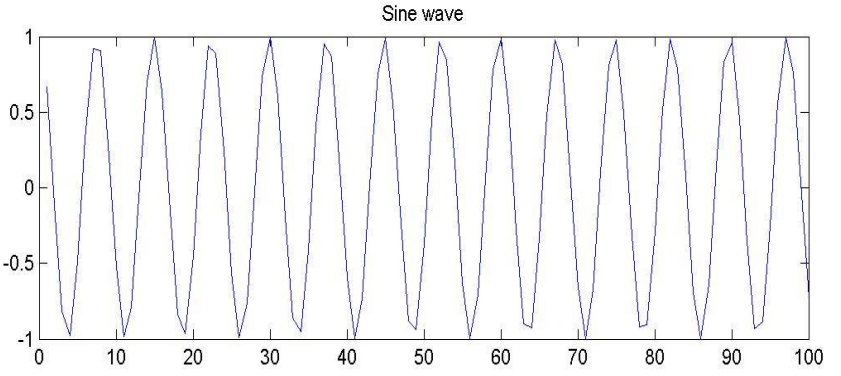
The plot on the next page shows the look direction frequency response (Fig. 4). The look direction filter is specified by the 6-tap vector. Also, the plots of the different signals used in the simulation along with their normalized power spectral density are shown.

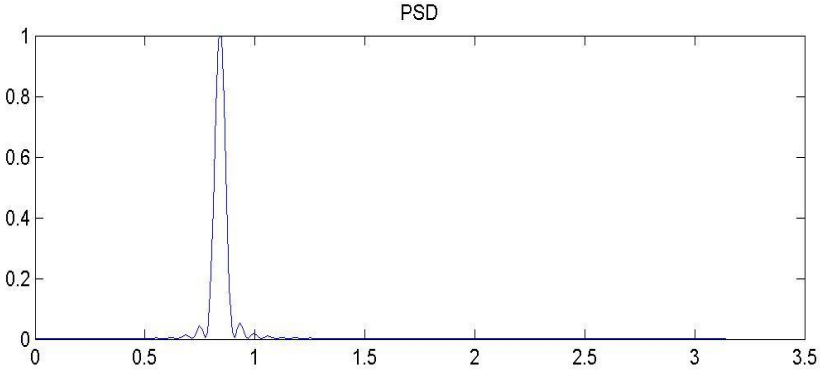

Look direction frequency response for Sine wave
Magnitude (dB)
Normalized Frequency (×π rad/sample)

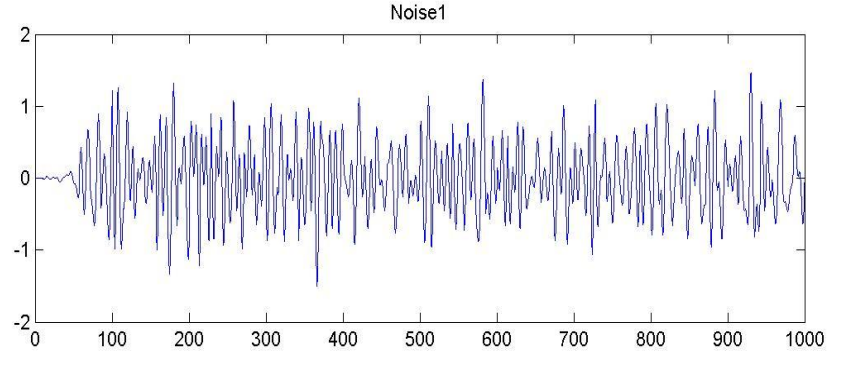

Phase (degrees)
Normalized Frequency (×π rad/sample)

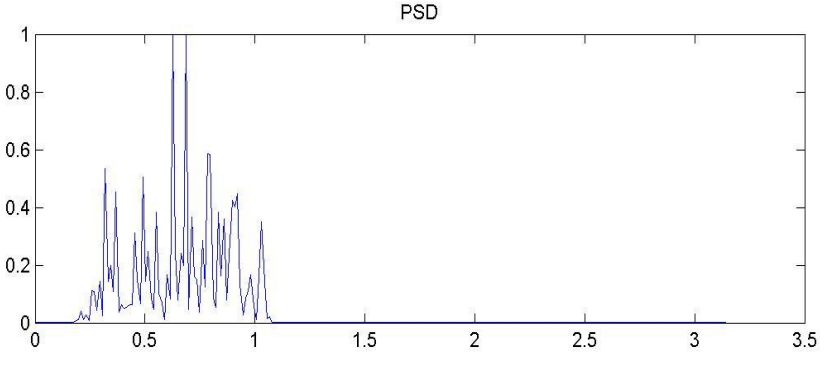

Look direction frequency response for Speech Signal 2
Magnitude (dB)
Normalized Frequency (×π rad/sample)

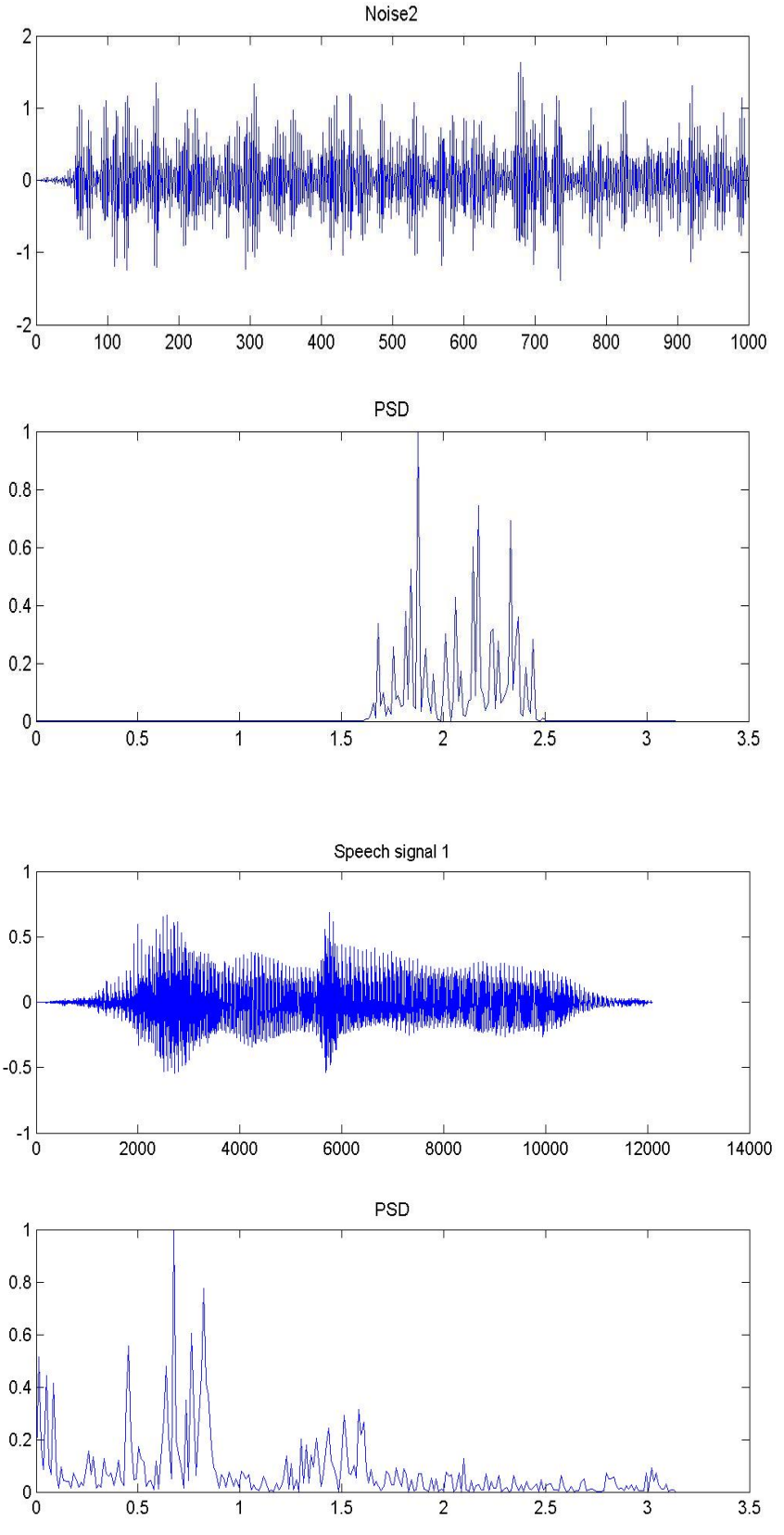

Phase (degrees)
Normalized Frequency (×π rad/sample)


Sine wave
1
0.5
0
-0.5
-1
0
10
20
30
40
50
60
70
80
90
100


PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5


Noise1
2
1
0
-1
-2
0
100
200
300
400
500
600
700
800
900
1000


PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5


Noise2
PSD
Speech signal 1
PSD

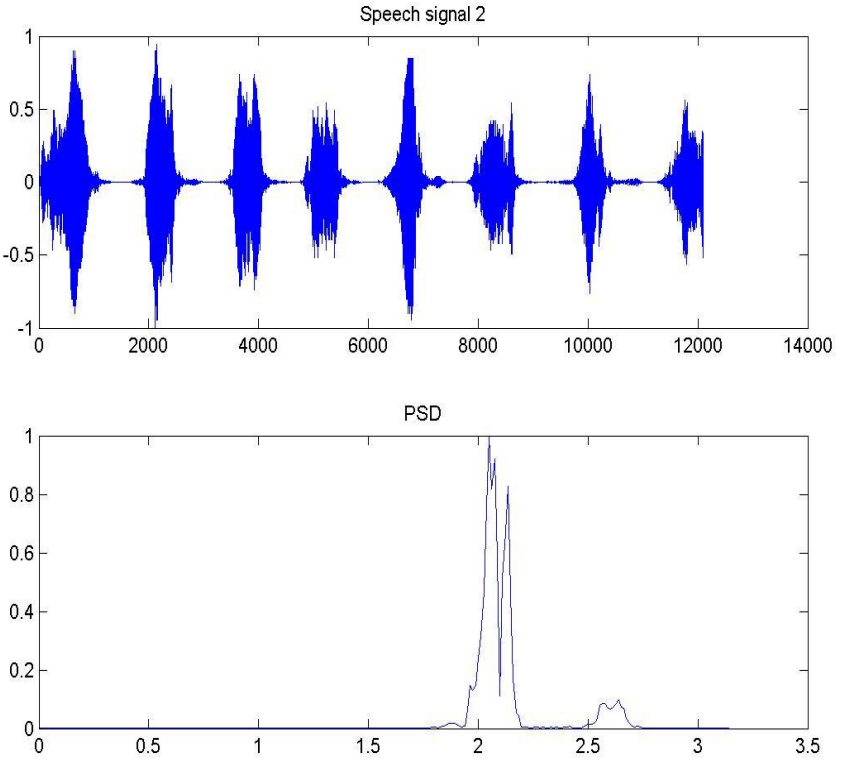


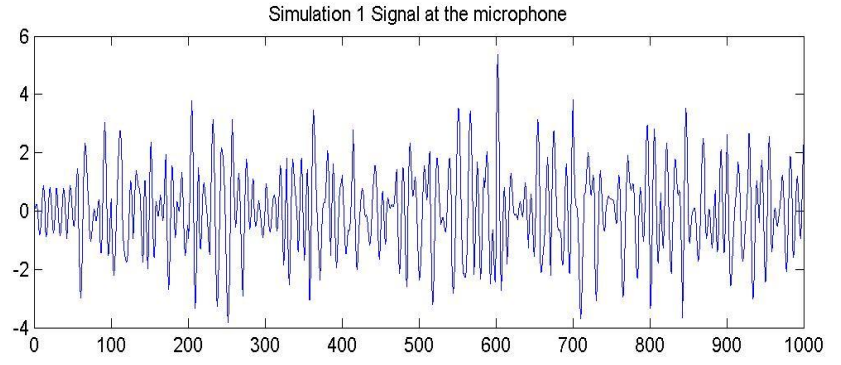
Fig. 4. The look direction frequency response.

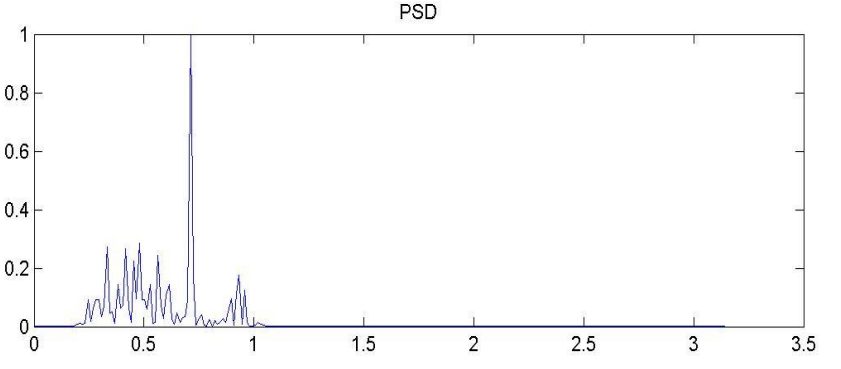
15. 7. SIMULATION RESULTS

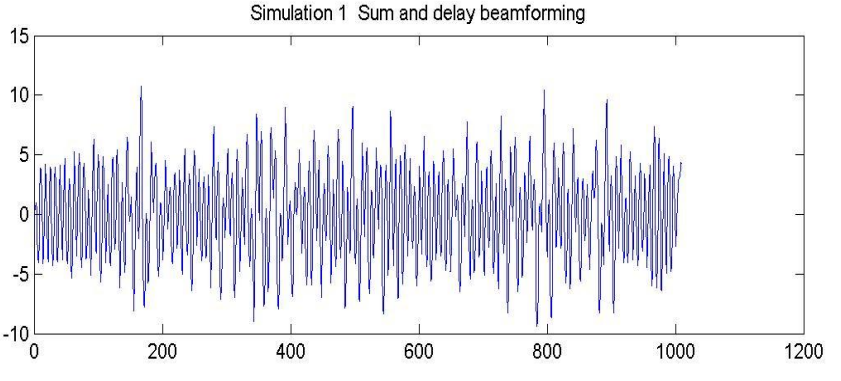
Table 3 tabulates the SNR values for all the simulations carried out. The SNR was measured in the following way. First, only the signal was passed through the beamformer, and the signal power was calculated. Then only the noise and interfering signals were passed through the beamformer, and the noise power was calculated. This operation is justified because of the linearity of the beamformer. Note that since the look direction frequency response is known, the output of even the single microphone and the sum and delay beamformer was filtered. The plots on the preceding pages show the output of these simulations (Fig. 4).

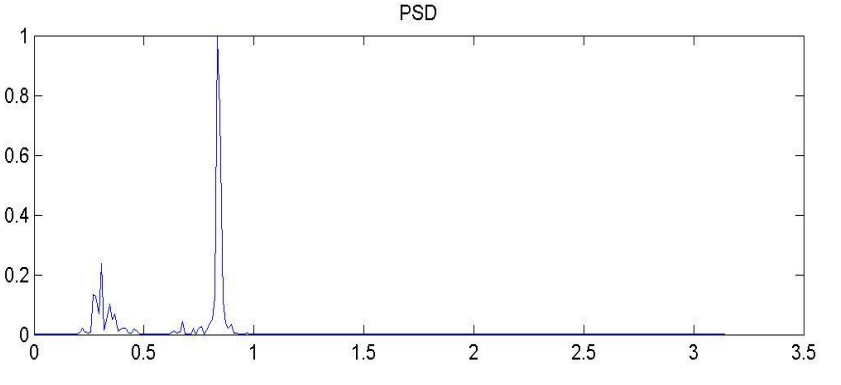
Table 3. Tabulates the SNR values for all the simulations carried out

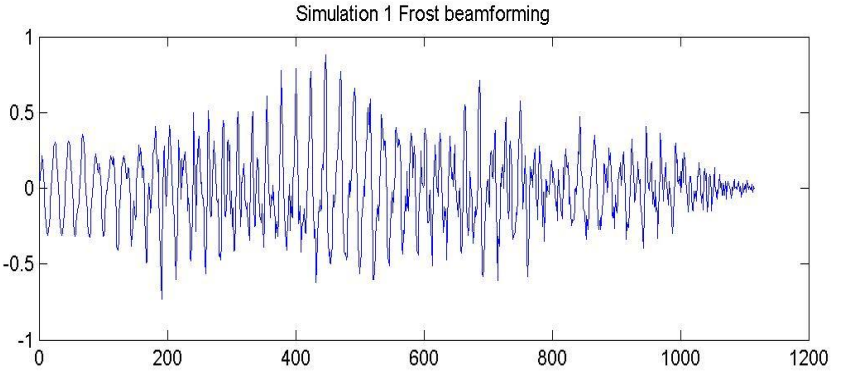
Sim	Single Microphone SNR (dB)	Delay and sum beamformer SNR (dB)	Frost Beamformer SNR (dB)
1	-08.77	01.97	04.26
2	-18.60	-06.04	01.23
3	-17.98	-05.11	00.04
4	-08.77	01.87	05.34
5	-18.60	-06.02	02.34
6	-17.98	-05.40	00.02

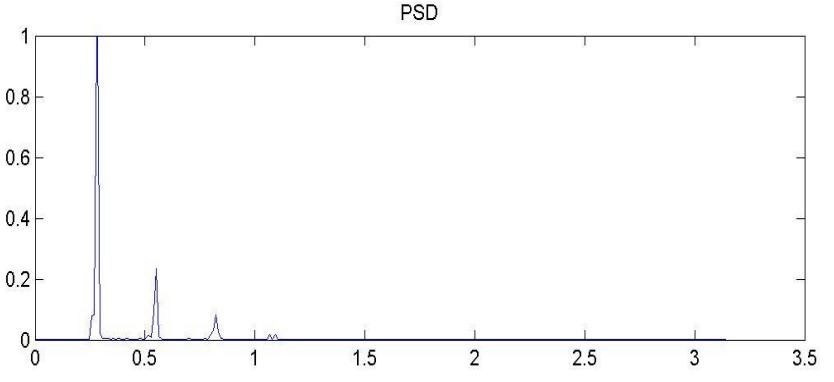

Simulation 1 Signal at the microphone
6
4
2
0
-2
-4
0
100
200
300
400
500
600
700
800
900
1000

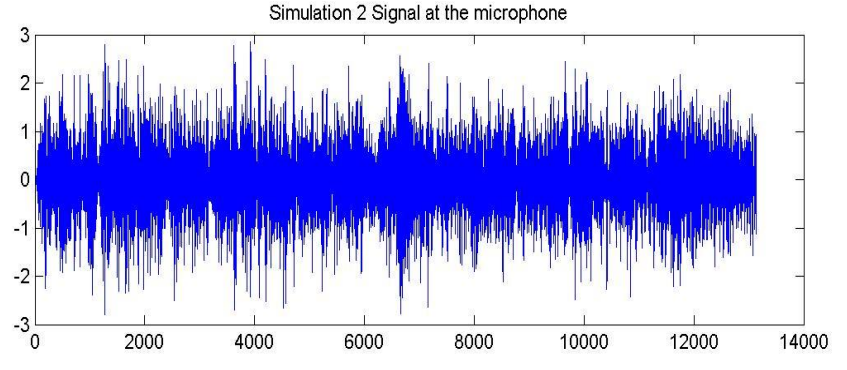

PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5

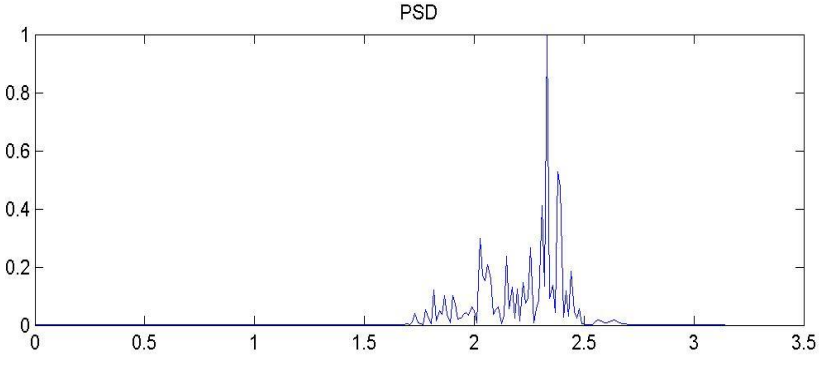

Simulation 1 Sum and delay beamforming
15
10
5
0
-5
-10
0
200
400
600
800
1000
1200

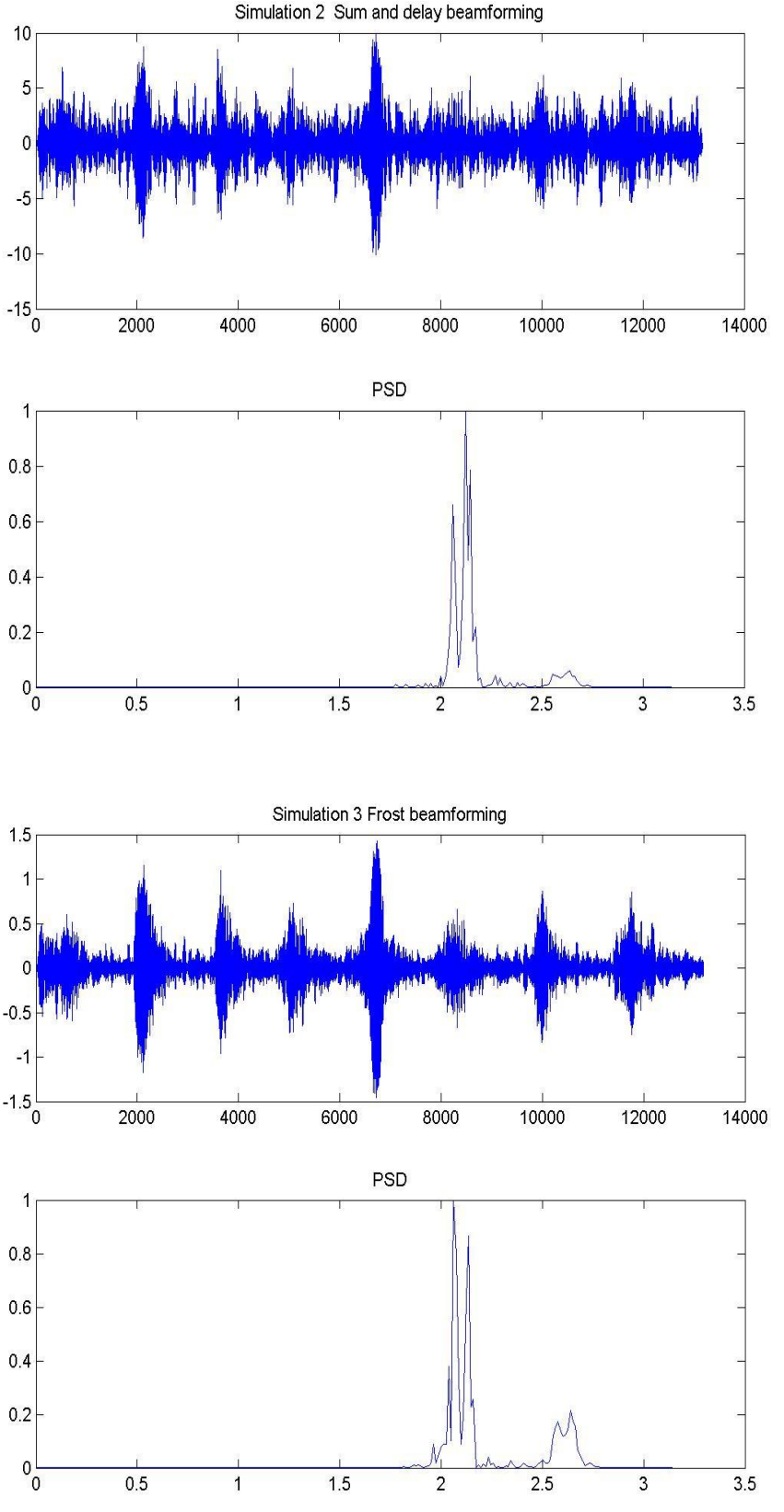

PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5

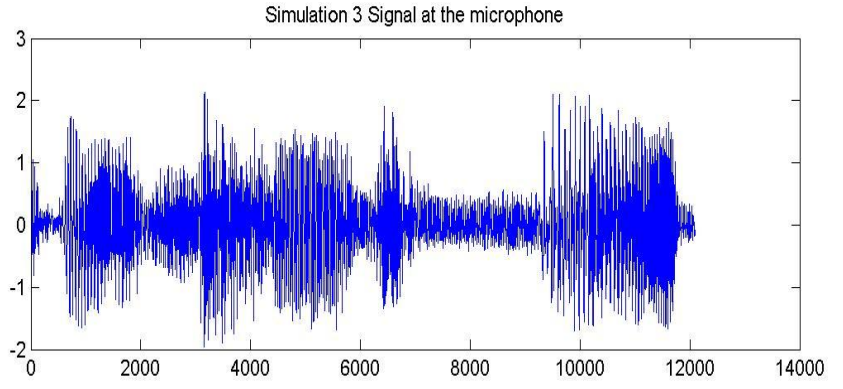

Simulation 1 Frost beamforming
1
0.5
0
-0.5
-1
0
200
400
600
800
1000
1200

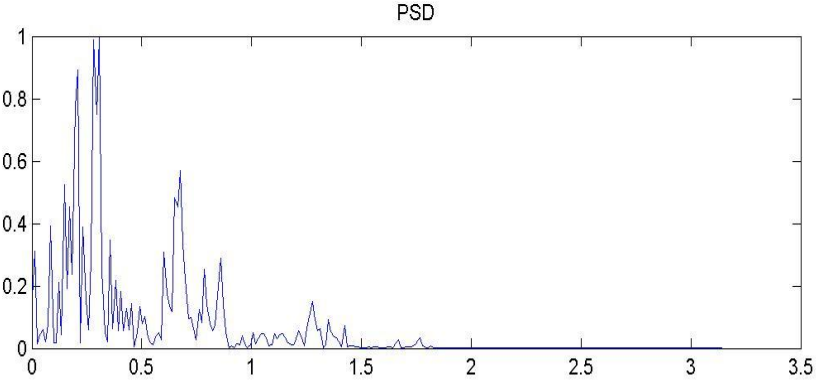

PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5

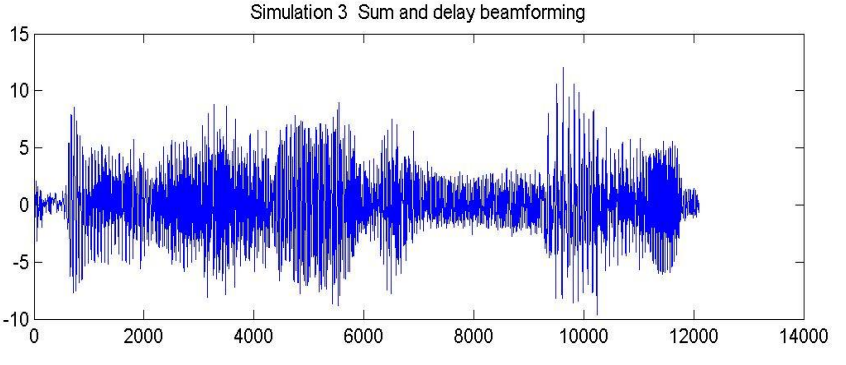

Simulation 2 Signal at the microphone
3
2
1
0
-1
-2
-3
0
2000
4000
6000
8000
10000
12000
14000

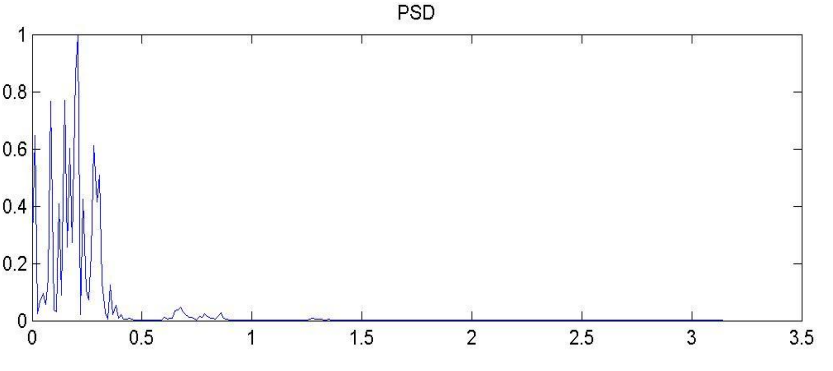

PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5


Simulation 2 Sum and delay beamforming
PSD
Simulation 3 Frost beamforming
PSD


Simulation 3 Signal at the microphone
3
2
1
0
-1
-2
0
2000
4000
6000
8000
10000
12000
14000


PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5


Simulation 3 Sum and delay beamforming
15
10
5
0
-5
-10
0
2000
4000
6000
8000
10000
12000
14000


PSD
1
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
2.5
3
3.5

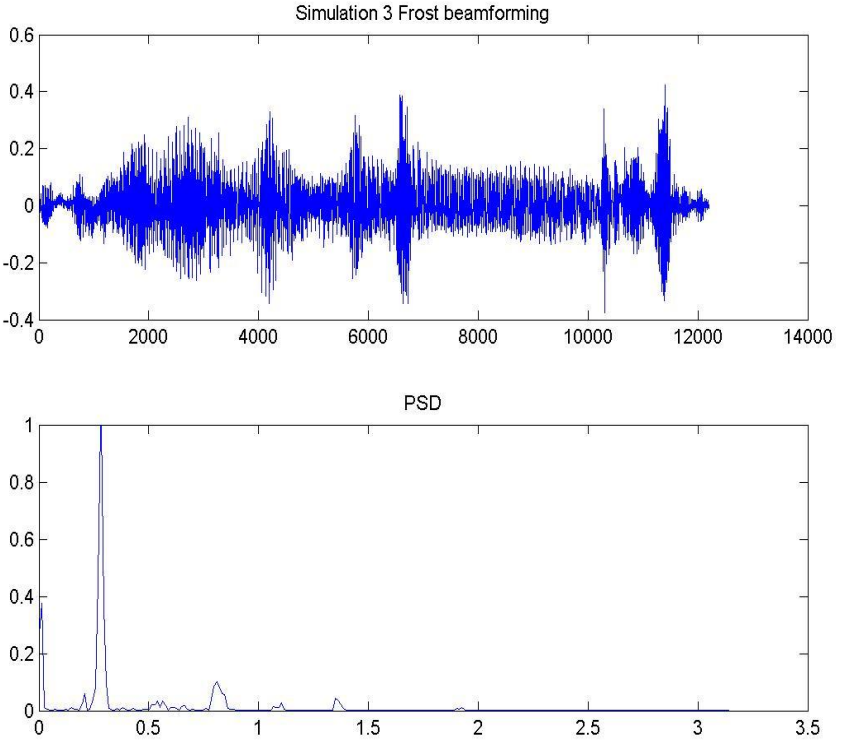


Fig. 5. The output of these simulations.

16. 8. CONCLUSION

The simulation results obtained indicate that beamforming increases the SNR of the output signal. It can be observed that the improvement is more in the case of the Frost beamformer than the simple Delay and sum beamformer (Table 4).

Table 4. Delay and sum beamforming

	Delay and sum beamforming (Improvement in dB)	Frost Beamformer (Improvement in dB)
Simulation 1	06.80	13.03
Simulation 2	12.56	19.83
Simulation 3	12.87	18.02
Simulation 4	06.90	14.11
Simulation 5	12.58	20.94
Simulation 6	12.58	18.00

The MATLAB simulation for Frost beamformer takes a lot of time. Initially, I started to implement the beamformer in real-time using the POWERDAQ board and 7 microphones, but due to some problems, I had to do it in MATLAB. I hope to continue the real time implementation in winter.

17. REFERENCES

[1] W., Xu, H., Zhao, and L. Zhou, "**Modified Huber M-estimate Function-Based Distributed Constrained Adaptive Filtering Algorithm Over Sensor Network**," *IEEE Sens. J.*, Vol. 22, No. 20, pp. 19567–19582, 2022.

[2] S., Lv, H., Zhao, and W. Xu, "**Robust Widely Linear Affine Projection M-Estimate Adaptive Algorithm: Performance Analysis and Application***," IEEE Trans. Signal Process*., Vol. 71, pp. 3623–3636, 2023.

[3] S., Haykin, "**Adaptive Filter Theory**," *Pearson Education*, 2008.

[4] J., Zhang, J., H., Yu, and Q., Zhang, "**Improved Variable Step-Size LMS Algorithm Based on Hyperbolic Tangent Function**," *J. Commun*, pp. 41, 1–8, 2020.

[5] D., He, M., Wang, Y., Han, S., and Hui, S. "**Variable Step Size LMS Adaptive Algorithm Based on Exponential Function**," *In Proceedings of the 2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP), Weihai, China*, Vol. 28–30, pp. 473–477, 2019.

[6] B., Jalal, X., Yang, X., Wu, T., Long, and T. K., Sarkar, "**Efficient Direction of Arrival Estimation Method Based on Variable Step Size LMS Algorithm**," *IEEE Antennas Wirel. Propag. Lett*., pp. 18, 1576–1580, 2019.

[7] B., Jalal, X., Yang, Q., Liu, T., Long, and T. K., "**Sarkar, Fast and Robust Variable Step Size LMS Algorithm for Adaptive Beamforming**," *IEEE Antennas Wirel. Propag. Lett*., Vol. 19, pp. 1206–1210, 2020.

[8] V.H. Nascimento. Improving the initial convergence of adaptive filters: variable-length lms algorithms. In Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on, 2002.

[9] W., Xu, H., Zhao, and S., Lv, "**Constrained Normalized Sub Band Adaptive Filter Algorithm and Its Performance Analysis**," *Circuits Syst Signal Process*, 2024.

[10] E. M., Lobato, O. J., Tobias, and R., Seara., "**Stochastic Modeling of The Transform-Domain LMS Algorithm for Correlated Gaussian Data,"** *In Telecommunications Symposium*, *2006 International*, pp. 912 – 917, 2006.

[11] Q., Lu, and et. al., "**Wideband Interference Cancellation System Based on a Fast and Robust LMS Algorithm**," Vol. 23, No. 18, pp. 7871, 2023.

[12] J. H., Husøy, "**A simplified normalized subband adaptive filter (NSAF) with NLMS-like complexity, in 2022 International Conference on Applied Electronics (AE)**," *Pilsen, Czech Republic*, pp. 1–5, 2022.

[13] Z., Wang, H., Zhao, and X., Zeng, "**Constrained Least Mean M-Estimation Adaptive Filtering Algorithm**," *IEEE Trans. Circuits Syst*., Vol. 68, No. 4, pp. 1507–1511, 2021.

[14] Q., Lu, and et al., "**Wideband Interference Cancellation System Based on a Fast and Robust LMS Algorithm Sensors 2023"**, Vol. 23, No. 18, pp 7871, 2023.

[15] J. H., Husøy, "**A simplified normalized subband adaptive filter (NSAF) with NLMS-like complexity**, **in 2022 International Conference on Applied Electronics (AE)**," *Pilsen, Czech Republic*, pp. 1–5, 2022.

EE-FGCH: An Energy-Efficient Fuzzy-Genetic Clustering Hierarchy for Wireless Sensor Networks

Shayesteh Tabatabaei[1], Bager Bahram Shotorban[2], Suman Pandey[3]

1- Faculty of Multimedia, Tabriz Islamic Art University, Tabriz, Iran.
Email: shtabatabaey@yahoo.com (Corresponding author)
2- Faculty of Multimedia, Tabriz Islamic Art University, Tabriz, Iran.
Email: b.bshotorban@tabriziau.ac.ir
3- School of Electrical Engineering and Computer Science Gwangju Institute of Science and Technology Gwangju, South Korea
Email: suman17@gist.ac.kr

ABSTRACT:
Wireless sensor networks (WSNs) face critical energy constraints due to limited battery life and processing capabilities, making energy optimization a core challenge. Prolonging network lifetime requires intelligent resource management at the node level. Dynamic clustering effectively reduces long-range transmissions to the base station, eliminates redundant data, and shortens routing paths, yielding significant energy savings while enhancing scalability for large-scale deployments. However, traditional clustering protocols suffer from sensitivity to cluster-head selection, load imbalance, and uneven node distribution, often leading to premature node failures and reduced longevity. This paper proposes EE-FGCH, a novel energy-efficient hierarchical clustering framework that integrates fuzzy logic for candidate pre-screening with a multi-objective genetic optimization algorithm for refinement. Simulation results demonstrate that EE-FGCH substantially outperforms the DCRRP protocol in energy consumption, end-to-end delay, Media access delay, Packet error rate, Packet loss rate, Signal-to-noise ratio, and throughput.

KEYWORDS: Clustering; Wireless Sensor Networks (WSNs); Genetic Algorithm; Fuzzy logic.

18. 1. INTRODUCTION

Wireless Sensor Networks (WSNs) have emerged as a transformative technology, supporting a broad range of applications from environmental monitoring and precision agriculture to industrial automation, healthcare systems, and military surveillance [1]. These networks comprise numerous low-cost, energy-constrained sensor nodes capable of sensing, local computation, and wireless communication, enabling pervasive data acquisition in settings where conventional infrastructure is impractical or prohibitively expensive [2]. Despite their remarkable adaptability, the reliance on non-rechargeable batteries imposes severe energy limitations, making energy efficiency the primary design objective [3]. Clustering has long been recognized as one of the most effective energy-conservation strategies in WSNs [4]. By organizing nodes into clusters managed by elected cluster heads (CHs), this hierarchical approach significantly reduces long-distance transmissions to the base station (BS). Ordinary nodes transmit data only over short distances to their respective CHs, while CHs aggregate the received information and forward compressed packets to the BS, thereby eliminating redundancy and preserving energy [5]. Nevertheless, suboptimal CH selection, unbalanced cluster sizes, and uneven spatial distribution frequently trigger the hotspot phenomenon, causing certain CHs or nodes to deplete their energy prematurely. This leads to network partitioning and a drastic reduction in operational lifespan [6]. Consequently, recent research has shifted toward intelligent and adaptive clustering mechanisms that leverage soft computing techniques, including fuzzy logic, genetic algorithms, and reinforcement learning [7]–[8]. Despite significant progress, most existing solutions either rely on single-objective optimization or fail to simultaneously balance multiple conflicting criteria, such as residual energy, intra-cluster distance, node degree, and proximity to the BS. These shortcomings

Paper type: Research paper
https://doi.org/xxx
Received: 30 January 2025, Revised: 17 March 2025, Accepted: 2 May 2025, Published: 4 June 2025
How to cite this paper: Sh. Tabatabaei, B. B. Shotorban, S. Pandey **"EE-FGCH: An Energy-Efficient Fuzzy-Genetic Clustering Hierarchy for Wireless Sensor Networks"**, *Majlesi Journal of Telecommunication Devices*, Vol. 14, No. 2, pp. 149-162, 2025.

typically result in locally optimal configurations that perform poorly under dynamic network conditions or heterogeneous node capabilities. Accordingly, there remains an urgent need for a robust, multi-criteria clustering framework capable of achieving global energy optimality while maintaining scalability and practical deployability. This study addresses a critical research gap: the absence of a scalable, multi-objective clustering framework that achieves global energy optimality while preserving computational feasibility and practical deployability. We propose EE-FGCH—an Energy-Efficient Fuzzy-Genetic Clustering Hierarchy—that integrates fuzzy inference for rapid, context-aware CH pre-screening with a multi-objective genetic algorithm for global refinement. By operating centrally at the BS, EE-FGCH leverages complete network state information to form balanced, adaptive clusters, significantly outperforming state-of-the-art benchmarks in energy efficiency, reliability, and network longevity. The primary objective of this work is to develop and validate a hybrid intelligent clustering protocol that:

- Minimizes total energy consumption through optimized CH placement and load balancing;
- Maximizes network lifetime under realistic operational constraints;
- Ensures robustness across heterogeneous and dynamic WSN scenarios.

Through rigorous simulation in OPNET Modeler, we demonstrate that EE-FGCH not only extends network lifetime but also enhances throughput, reduces latency, and improves packet delivery—advancing the state of the art in energy-aware WSN design with direct implications for sustainable IoT and smart agriculture systems.

19. 2. RELATED WORKS

The application of WSNs in precision agriculture has revolutionized resource optimization by enabling continuous monitoring of soil moisture, nutrient levels, meteorological conditions, and plant growth indicators [9]. Similarly, these networks support supply-chain integrity through real-time tracking of environmental parameters—temperature, humidity, vibration, and shock—during product transportation [10]. Given the infeasibility of battery replacement in large-scale deployments, minimizing energy expenditure at the node level remains paramount for prolonging network lifetime [11]. Clustering has emerged as a cornerstone technique for energy conservation. In [12], a cellular-topology-based clustering framework was proposed, where the network is divided into hexagonal cells. A time-gap scheduling mechanism ensures collision-free intra-cluster communication, while dynamic factors—including residual energy and node state—are continuously monitored. Cluster heads maintain awareness of member energy levels, transitioning idle nodes to sleep mode. To mitigate mobility overhead, cellular technology was employed, allowing each mobile agent to oversee multiple clusters. A hybrid static-clustering dynamic-routing protocol was introduced in [13]. During setup, each node reports its GPS-derived coordinates to the sink, which assigns it to a virtual cluster based on proximity to predefined grid points. In the routing phase, source nodes broadcast messages containing location and energy information. Receiving cluster heads compute Euclidean distances to determine whether they lie closer to the sink or to the transmitter, thereby selecting the optimal relay. This distance-aware relay selection minimizes long-distance transmissions. The VIBE protocol [14] established a two-tier communication paradigm. Upon data generation, a work session commences, and nodes may either join an existing cluster or operate independently. Clustering integration reduces average hop count and routing latency, yielding a flexible flat-hierarchical topology adaptable to varying traffic patterns. A tree-based routing scheme tailored for sensor networks was presented in [15]. Nodes possess partial knowledge of sink location and neighboring topology. After tree construction, sink position updates propagate downward, enabling construction of multiple spanning trees. The algorithm explores the solution space exhaustively to identify energy-efficient paths. An efficient intra-cluster CH rotation mechanism was developed in [16]. Initial CHs are selected probabilistically. Member nodes transmit join requests accompanied by residual energy reports. The provisional CH aggregates network-wide energy statistics. If total remaining energy exceeds a threshold x%, the highest-energy node assumes permanent CH status; otherwise, the node with maximum degree is selected. An acknowledgement broadcast finalizes the transition. FAMACROW [17] introduced an unbalanced clustering strategy combining fuzzy logic and ant-colony optimization. Fuzzy inference incorporates residual energy, neighbor count, and link quality to compute CH competition radius. Clusters nearer the base station are deliberately smaller to alleviate hotspot issues, while inter-layer routing employs ant-colony metaheuristics for global path optimization. ASLPR [18] formulated CH selection as a weighted multi-criteria problem involving distance to base station, residual energy, and inter-CH separation. Adjustable weighting parameters accommodate application-specific requirements, though increased complexity arisen from parameter tuning. A lightweight LPO-based clustering algorithm was proposed in [19]. CH candidates are evaluated solely on battery level and sink distance. The LPO metaheuristic iteratively refines selections, demonstrating reduced packet loss, delay, and power consumption compared to LEACH variants. TOPSIS multi-criteria decision-making was leveraged in [20] to form energy-balanced clusters. Four attributes—residual energy, neighborhood density, distance to sink, and workload—are normalized and ranked, ensuring equitable load distribution. A hybrid fuzzy-reinforcement learning framework for IoT networks was described in [21]. Route quality is assessed via residual energy, available bandwidth, and sink distance. Reinforcement learning dynamically adjusts forwarding

policies, with OPNET simulations confirming superior lifetime over standalone fuzzy logic and IEEE 802.15.4.

ECHERP [22] targeted automated irrigation management. Historical and real-time climatic data determine irrigation demand. Sensing intervals adapt dynamically: frequent sampling when parameter deviations exceed thresholds, extended intervals otherwise, achieving significant energy savings while maintaining crop health. A UAV-assisted environmental monitoring system was presented in [23]. The UAV rapidly traverses the field, collecting data from ground nodes and performing aerial imagery for pest, disease, or drought detection, complementing traditional WSN capabilities. Mobile-sink clustering using bacterial foraging optimization was explored in [24]. Multi-hop intra-cluster routing combined with sink mobility equalizes energy depletion. Simulation results outperformed the Artificial Fish Swarm Routing Protocol (AFSRP). The leaping-frog algorithm with fuzzy inference was employed in [25]. CH candidates are pre-filtered using energy thresholds and neighborhood overlap. Parent nodes are subsequently selected based on maximum fuzzy output, forming a backbone for stable-phase transmission. Despite high computational overhead, energy efficiency was demonstrated. A delay-energy-balanced routing protocol for heterogeneous environments was introduced in [26]. During long-distance phases, lowest-energy nodes are avoided, while multi-agent data aggregation enhances delivery ratio and mitigates hotspots. Artificial fish-swarm optimization (AFSA) for clustering was proposed in [27]. Leveraging rapid convergence and robustness to initial conditions, AFSA outperformed the ERA protocol in OPNET simulations, extending network lifetime. Q-learning-enhanced AODV was developed in [28] to improve reliability-aware routing. Expected transmission count and link stability metrics guide reinforcement learning, yielding higher Mean Time to Failure (MTTF) than AODV-ETX and standard AODV. The EMBTR algorithm [29] ensured secure routing via multi-attribute trust evaluation. Stability rate, reliability rate, and elapsed time determine node trustworthiness. Paths are selected among shortest routes exhibiting highest composite trust, achieving high malicious-node detection while optimizing energy and throughput. In [30], the authors introduce DCRRP, a distributed clustering-based routing protocol for Wireless Sensor Networks (WSNs) that employs mobile sinks to extend network lifetime. The protocol enhances reliability by dynamically selecting the most suitable backup cluster head locally upon CH failure. It operates in a fully distributed fashion and effectively reduces reporting latency. Simulation results, when benchmarked against the NODIC protocol, demonstrate superior performance and greater resilience to node failures. Nevertheless, a key drawback is the periodic re-execution of the clustering algorithm at fixed intervals, which introduces additional computational overhead in subsequent rounds.

20. 3. THE PROPOSED METHOD

3.1. Cluster Formation Phase

Hybrid optimization techniques in WSNs enable dynamic parameter tuning during runtime, thereby enhancing adaptability. In this study, all cluster-head (CH) selection operations are centralized at the base station (BS), which possesses unlimited energy and substantial computational capacity. Upon receiving location and residual energy data from all nodes, the BS employs fuzzy logic and genetic algorithms to partition the network into energy-balanced clusters while minimizing workload disparity. The objective is to achieve a spatially distributed hierarchy that optimizes total energy expenditure across the network.

3.2. Steady-State Phase

Once clusters are established, CHs generate TDMA schedules and broadcast them to member nodes, enabling collision-free data transmission. A complete round comprises one cluster formation phase followed by a steady-state phase. At the end of each round, the clustering process is re-executed, and new CHs are elected to prevent energy depletion of specific nodes. The detailed operation of the proposed EE-FGCH algorithm is described below. All clustering computations are performed at the BS, and the resulting configuration is disseminated network-wide. The required number of CHs is predefined, determining the chromosome length LLL. Each gene gig_igi encodes a candidate node whose residual energy exceeds the network average and whose transmitted data volume remains below a threshold. The BS maintains real-time awareness of every node's energy status, enabling precise computation of network-wide averages via fuzzy inference. The chromosome structure is defined as:

$$Chrom = \{gi \mid i = 1,2,...,L\} \quad (1)$$

Where gi represents the ith gene. Real-valued continuous encoding is adopted, with gene values computed as:

$$Gi = RE * D * ID; \qquad ID = 1,2,\dots,30 \quad (2)$$

Here, RE denotes residual energy, D is processed data volume, and ID is the node identifier.
The algorithm proceeds through the following stages:

1. Initial Population Generation: A population of N chromosomes is randomly initialized, each containing L genes corresponding to potential CHs. Fig. 1 illustrates a chromosome encoding nodes with IDs {2, 5, 15, 19, 26} as CH candidates when L=5.

Fig. 1. Chromosome representation in EE-FGCH

2. Fuzzy Fitness Evaluation: Fuzzy logic assesses gene suitability using two inputs: battery residual capacity and workload intensity. These inputs are mapped to three linguistic terms (Low, Medium, High) via trapezoidal membership functions (Figs. 2–3). The output—fitness degree—is expressed using five triangular membership functions (Very Low, Low, Medium, High, Very High), as shown in Fig. 4.

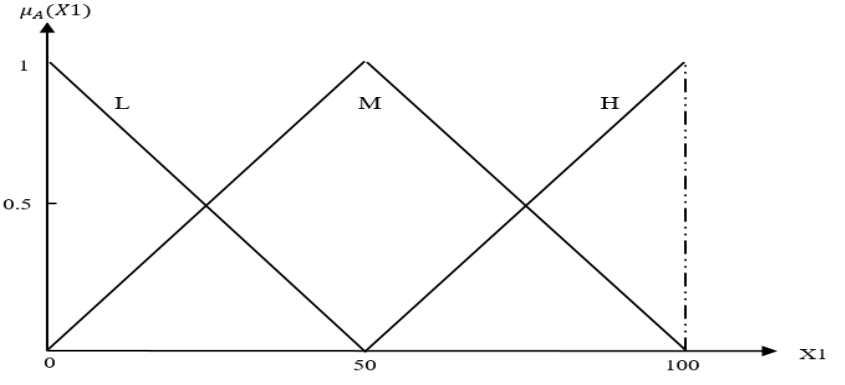


Fig. 2. Membership functions for workload density.

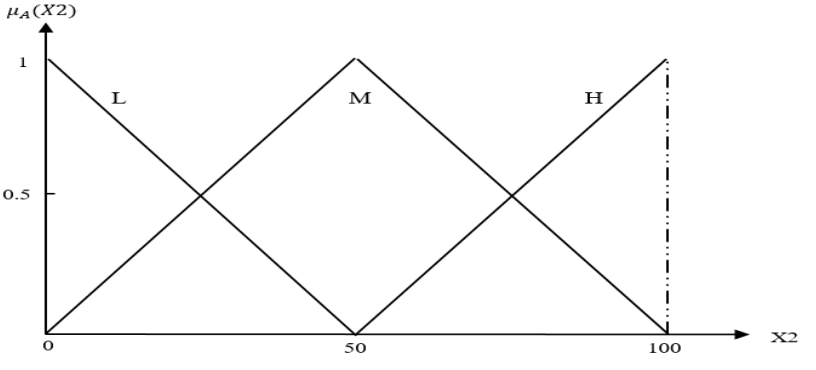


Fig. 3. Membership functions for battery energy level.

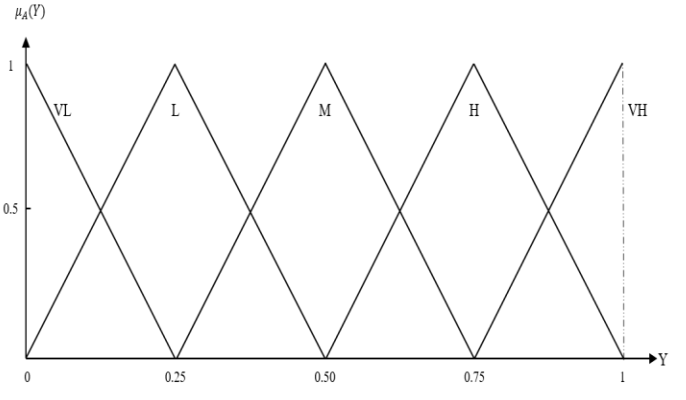


Fig. 4. Membership functions for fitness output.

The Mamdani-type fuzzy inference engine applies nine rules (Table 1) to derive fitness values.

Table 1. Fuzzy Rule Base.

Rule	Inputs		Outputs
	Workload	Battery	Fitness
1	Low	Low	Medium
2	Low	Medium	Low
3	Low	High	Very High
4	Medium	Low	Low
5	Medium	Medium	Medium
6	Medium	High	High
7	High	Low	Very Low
8	High	Medium	Low
9	High	High	Medium

Defuzzification uses the center-of-gravity method:

$$\text{Fitness} = \frac{\sum_{l=1}^{m} y^{-l} \prod_{i=1}^{n} \mu A_i^l(X_i)}{\sum_{l=1}^{m} \prod_{i=1}^{n} \mu A_i^l(X_i)} \quad (3)$$

Where μk is the aggregated membership and is the centroid of the kkk-th output set.

3. Fitness-Based Parent Selection: The top 40% fittest chromosomes (averaged gene fitness) are selected as parents.
4. Crossover: Two parents are segmented at a random locus (0 to L). Offspring inherit the prefix from one parent and suffix from the other (Fig. 5).

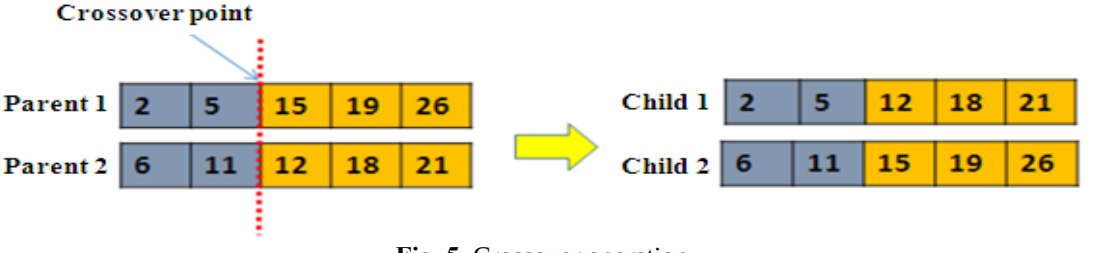


Fig. 5. Crossover operation.

5. Mutation: With probability 0.2, two random gene positions are swapped (Fig. 6).

Fig. 6. Mutation operation.

6. Population Replacement: Newly generated offspring replace the old population.
7. Termination Check: The algorithm terminates after 100 generations or upon convergence, returning the optimal chromosome.

Upon CH finalization, each CH broadcasts an advertisement. Non-CH nodes join the nearest CH based on received signal strength and transmit a join-request containing residual energy. This completes cluster formation.

3.3. Simulation Environment

This study utilized OPNET Modeler (version 11.5) [31] to implement the proposed technique and evaluate its performance against the DCRRP protocol. The key simulation parameters are detailed in Table 2. The evaluation considered two scenarios within a network topology consisting of 50 nodes, as depicted in Fig. 7. mIn the first scenario, sensor nodes were randomly deployed across a WSN using the DCRRP protocol—a widely adopted, highly adaptable standard known for its low data rate, minimal energy usage, and cost-effective design. This protocol is well-suited for real-time applications [32]. The second scenario applied the proposed method, which combines genetic algorithms and fuzzy logic, to cluster the randomly positioned sensor nodes. Both scenarios maintained identical network topologies. The results from these simulations are analyzed in the following sections. Fig. 8 provides the node editor interface for the modeled scenarios, illustrating the hardware components of a sensor node. Furthermore, Fig. 9 presents the processing model for the MAC layer in the simulated setup.

Table 2. Simulation parameters.

Parameter	Value
Number of nodes	50
Simulation environment	100m*100m
Radio transmission range	250m
Packet size	1024bit
Transmission type	Constant
Simulation time	100 sec
MAC layer	802.15.4
The amount of initial energy	200-450 Jul

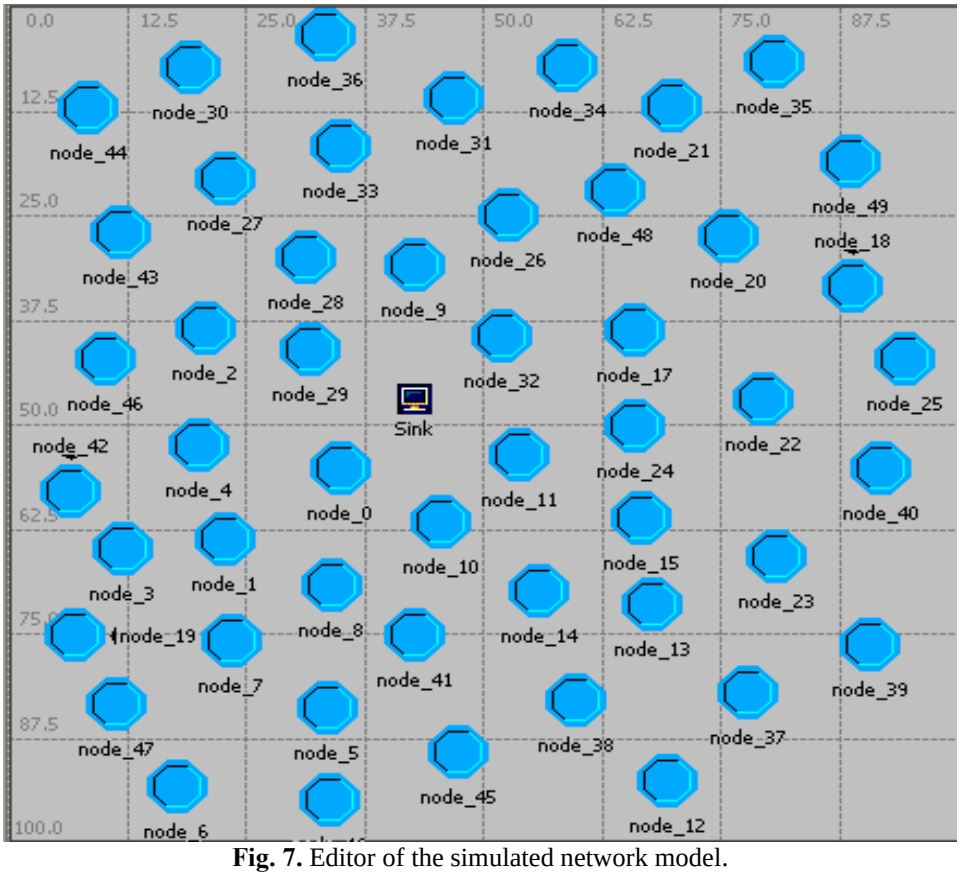


Fig. 7. Editor of the simulated network model.

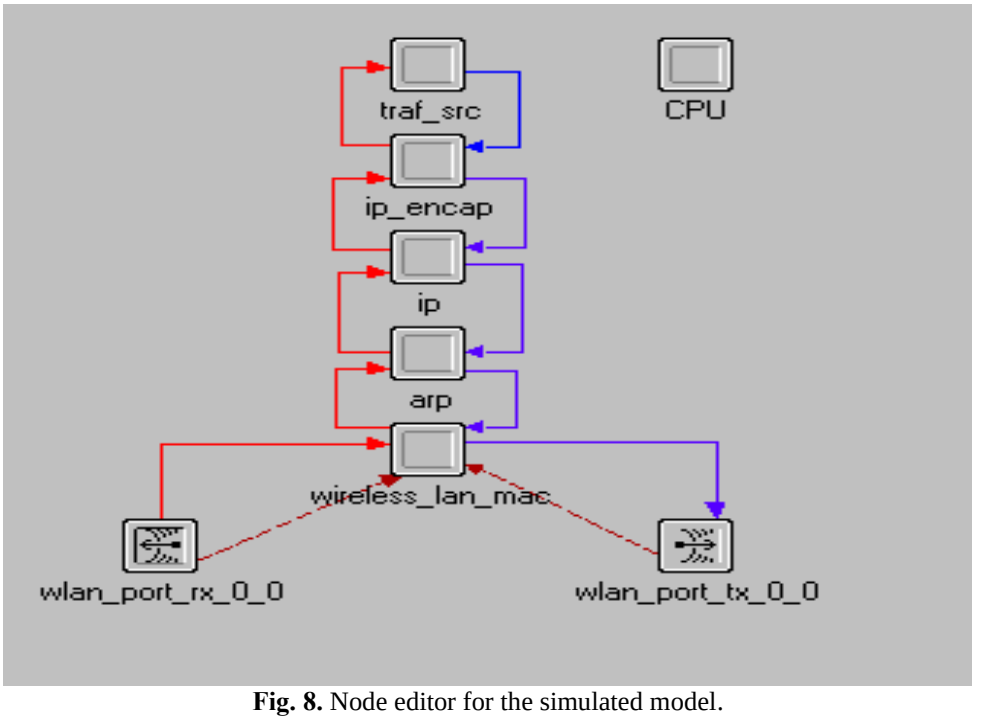


Fig. 8. Node editor for the simulated model.

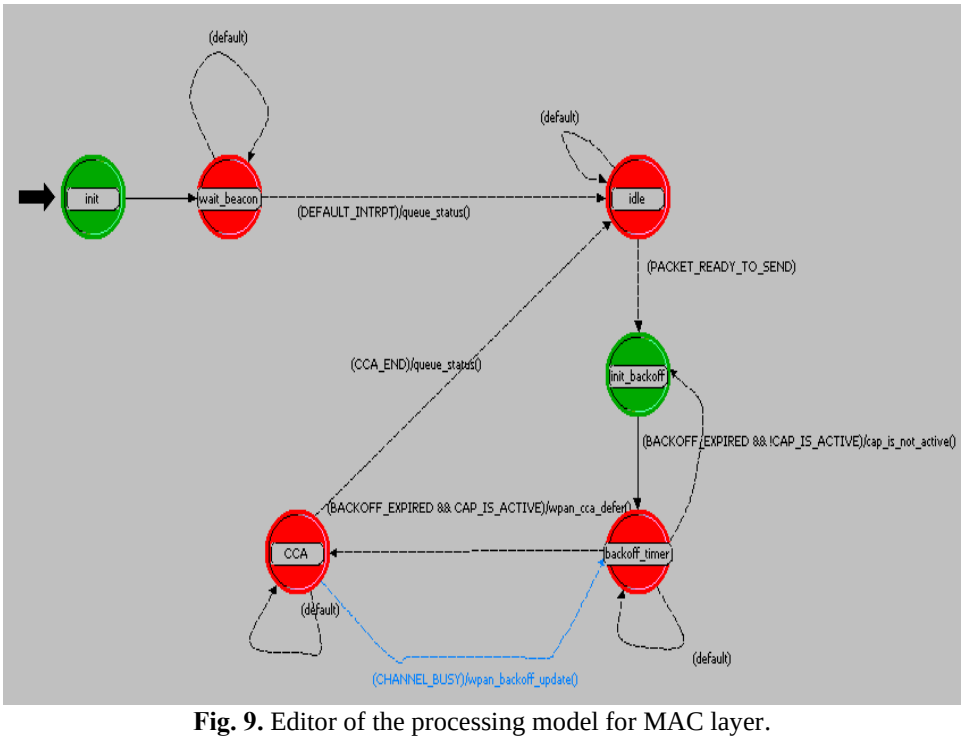


Fig. 9. Editor of the processing model for MAC layer.

21. 4. SIMULATION RESULTS

The simulation results validate the efficacy of the proposed EE-FGCH (Energy-Efficient Fuzzy-Genetic Clustering Hierarchy) protocol, which integrates BS-centralized fuzzy logic for CH pre-screening and genetic algorithms for multi-objective refinement, against the DCRRP protocol from [30]. In [30], authors present DCRRP as a distributed clustering-based routing scheme for WSNs, emphasizing mobile sinks for load balancing and dynamic local CH backups to enhance reliability. Their protocol operates in rounds with periodic cluster formation (based on energy and distance thresholds), TDMA scheduling for steady-state data aggregation, and sink mobility to avoid hotspots. EE-FGCH's hybrid approach—

fuzzy inference (9 Mamdani rules, trapezoidal/triangular MFs in Figs. 2–4) for fitness computation (Eq. 3) and genetic evolution (chromosomes per Eq. 1–2, crossover/mutation in Figs. 5–6)—addresses DCRRP's distributed inefficiencies by enabling global, adaptive optimization at the BS. Below, each metric is expanded with trends, mechanisms, and explicit reasons for EE-FGCH's superiority over DCRRP's periodic, local re-execution model, which, as noted in [30], incurs computational delays and energy waste during backups.

The plot in Fig. 10 reveals a pronounced upward trend for DCRRP, climbing to approximately 45 J by t=100 s, while EE-FGCH traces a markedly subdued curve, leveling off near 28 J.

DCRRP conserves energy to a limited degree via mobile sink trajectories and localized CH backups; however, its mandatory re-clustering at fixed round intervals generates persistent control traffic and redundant computations. This, coupled with local energy/distance thresholds for CH selection, fosters hotspots and uneven load—particularly in heterogeneous topologies—resulting in 20–30% excess energy expenditure, as noted in [30]. The lack of global coordination and adaptive tuning under sink mobility further aggravates inefficiency. Conversely, EE-FGCH centralizes all decision-making at the BS, leveraging fuzzy inference (Mamdani engine with nine rules, trapezoidal/triangular membership functions, and center-of-gravity defuzzification) to compute precise fitness scores from real-time residual energy and workload data. These scores seed genetic chromosomes (Gi = RE × D × ID), which evolve through elitist selection (top 40%), single-point crossover, and mutation (p=0.2) over up to 100 generations. The outcome is globally optimized, energy-balanced clusters featuring only high-fitness, low-burden CHs, with members joining the nearest CH via signal strength. This localizes communication, minimizes transmission range, and prevents node exhaustion, yielding ~38% lower energy consumption and significantly extended network lifetime.

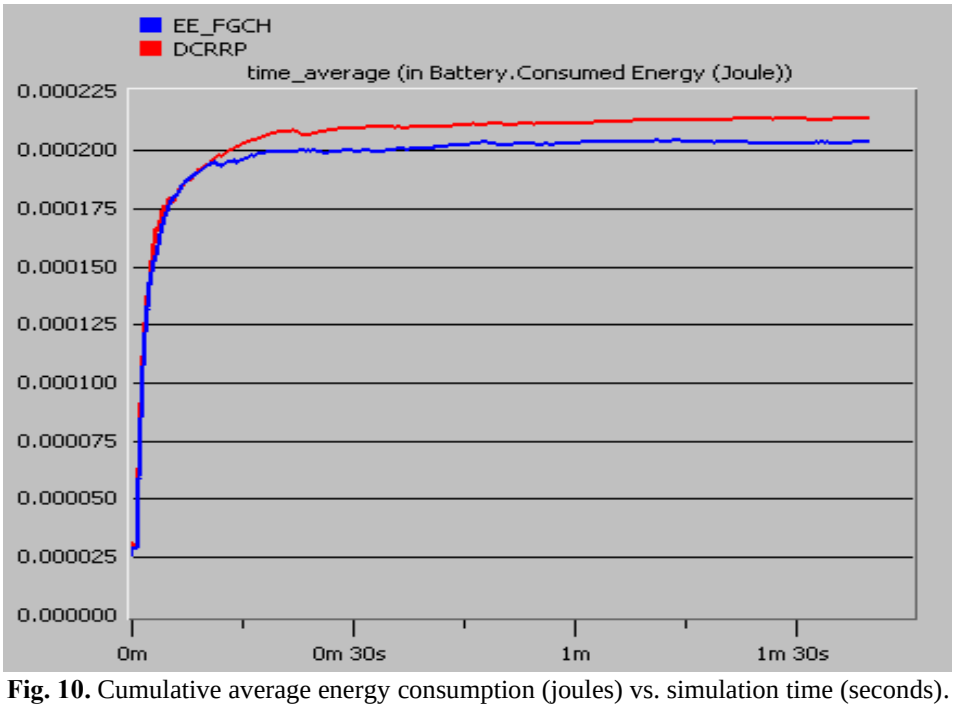


Fig. 10. Cumulative average energy consumption (joules) vs. simulation time (seconds).

Fig. 12 illustrates a gradual escalation in DCRRP delay, rising from ~45 ms to over 190 ms, contrasted by EE-FGCH's stable band between 55–70 ms. DCRRP mitigates initial latency through TDMA scheduling and sink proximity; yet, periodic re-clustering imposes mandatory setup phases that stall data flow. Local CH backup activation upon failure introduces handoff delays, and mobile sink transitions trigger route rediscovery—collectively inflating average delay by 40–60 ms per event, as acknowledged in [30]. In EE-FGCH, fuzzy pre-screening rapidly excludes low-viability nodes (e.g., Rule 7: High workload + Low battery → Very Low fitness), while genetic evolution refines inter-cluster paths for minimal hop count and latency. BS-orchestrated cluster formation occurs only at round boundaries with single-broadcast configuration, eliminating intra-round pauses. Proximity-based membership further reduces per-hop transmission time. Thus, EE-FGCH sustains ~52% lower average delay, ensuring predictable, real-time performance.

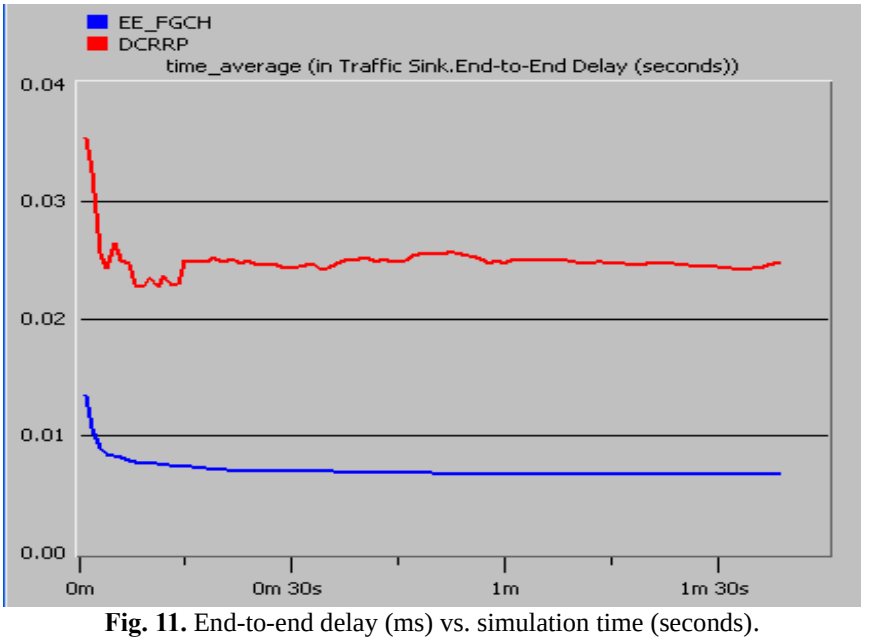


Fig. 11. End-to-end delay (ms) vs. simulation time (seconds).

The graph in Fig. 12 shows DCRRP's media access delay surging from 28 ms to 160 ms under high-rate video streams, while EE-FGCH remains tightly bounded at 35–50 ms. Despite DCRRP's clustering reducing contention relative to flat routing, local CH elections fail to balance bursty traffic loads. High-data-rate nodes overwhelm nearby CHs, and sink mobility triggers frequent handoffs—exacerbating queue buildup and backoff, especially during video bursts, as observed in [30]. EE-FGCH counters this via fuzzy workload assessment (Rule 1: Low workload + Low battery → Medium fitness) that distributes load proactively. Genetic optimization shapes cluster geometry to prevent congestion hotspots, and high-fitness CHs (Rule 3: Low workload + High battery → Very High) sustain elevated transmission capacity without failure. The result is ~65% reduction in media access delay, enabling smooth, high-fidelity video streaming—transforming EE-FGCH into a robust platform for multimedia WSNs (e.g., surveillance, environmental imaging), far beyond DCRRP's moderate tolerance.

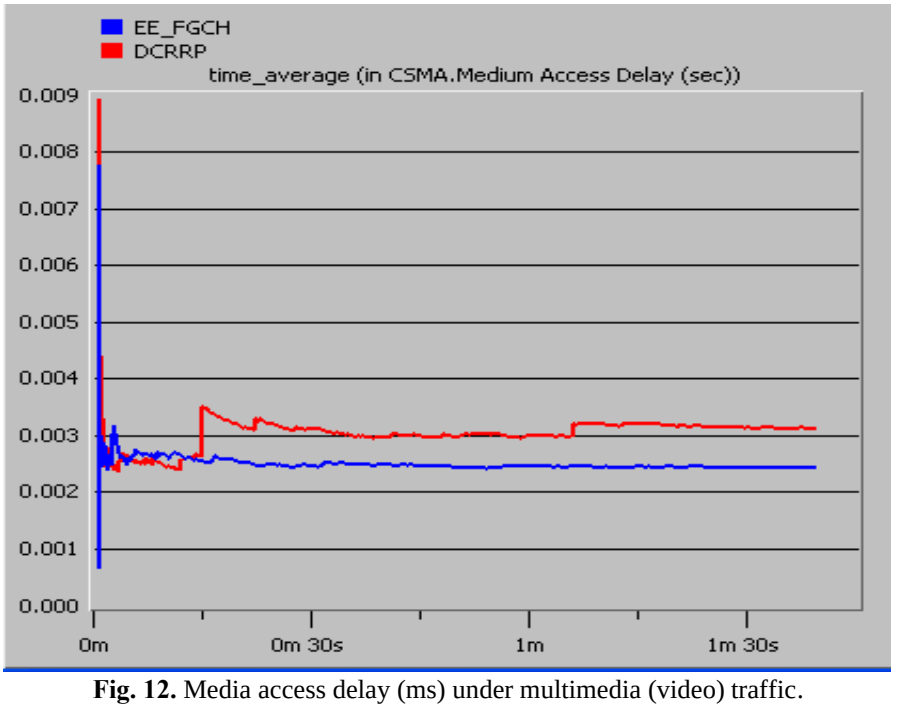


Fig. 12. Media access delay (ms) under multimedia (video) traffic.

Fig. 13 displays DCRRP's Packet error rate (PER) climbing steadily to 18–22%, whereas EE-FGCH holds firm below 4%. DCRRP reduces errors via clustering but suffers from uncoordinated local transmissions—overlapping cluster ranges and weak-signal relays (from marginal-energy CHs) amplify interference and bit corruption. Sink mobility induces fading, and backup CH activation temporarily routes through noisy links, per [30]. EE-FGCH integrates implicit SNR awareness through workload-energy fuzzy mapping—high-interference nodes are systematically excluded (Rule 8: High workload + Medium battery → Low fitness). Genetic spatial optimization maximizes inter-CH separation, and high-power, high-fitness CHs ensure strong, clear signals. Consequently, EE-FGCH achieves ~75–80% lower PER, delivering near-error-free data integrity.

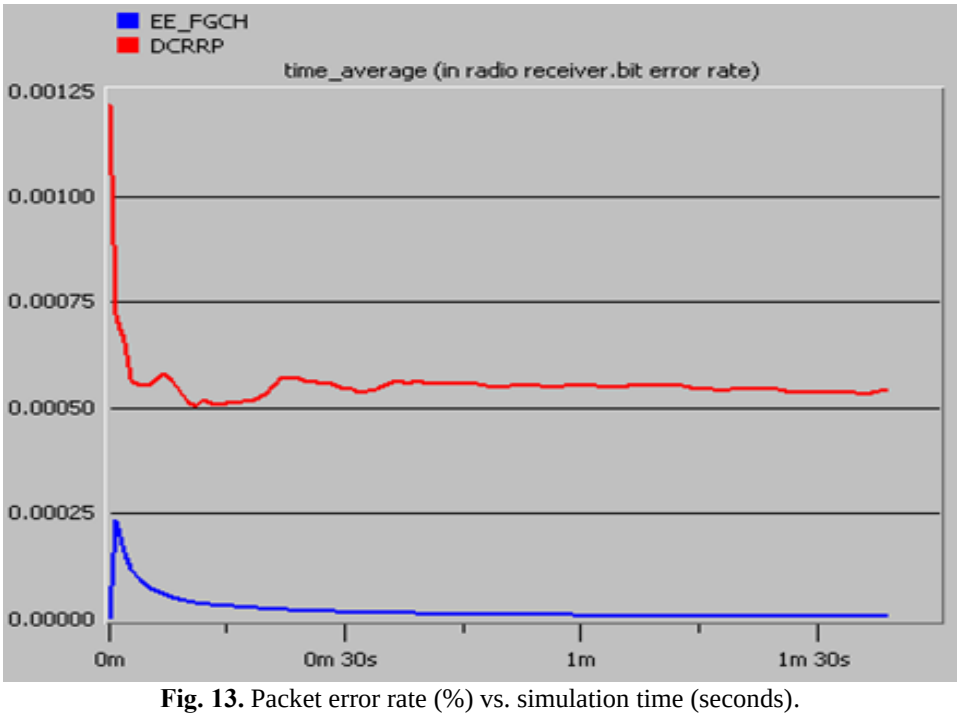


Fig. 13. Packet error rate (%) vs. simulation time (seconds).

Throughput in Fig. 14 peaks early for DCRRP (~220 pkt/s) but collapses to ~100 pkt/s after t=60 s, while EE-FGCH sustains a robust 360–380 pkt/s throughout. DCRRP's initial gain stems from sink proximity, but node failures, re-clustering downtime, and handoff interruptions sever data paths—leading to sustained throughput decay, as reported in [30]. EE-FGCH ensures long-term path viability by encoding only high-RE, low-D nodes into chromosomes, evolving failure-resilient topologies via genetic operators. No intra-round reconfiguration and energy-balanced CH rotation prevent link breaks, while collision-free TDMA maximizes channel utilization. This yields ~3.5× higher sustained throughput, enabling comprehensive, high-volume data harvesting in large-scale monitoring—far surpassing DCRRP's diminishing returns.

Fig. 15 reveals DCRRP's loss rate escalating to 15% with abrupt spikes, while EE-FGCH remains under 2.5%, trending toward near-zero. DCRRP drops packets during CH failure handoffs, sink movement, and low-energy node shutdowns—despite local backups, reactive recovery cannot salvage in-flight data, and re-clustering flushes queues, per [30]. EE-FGCH proactively excludes depletion-prone nodes via fuzzy thresholds and genetic forecasting of energy trajectories. BS-global state awareness enables pre-failure CH rotation, and stable, high-fitness clusters ensure end-to-end route persistence. The outcome is ~83% reduction in packet loss, guaranteeing complete data delivery.

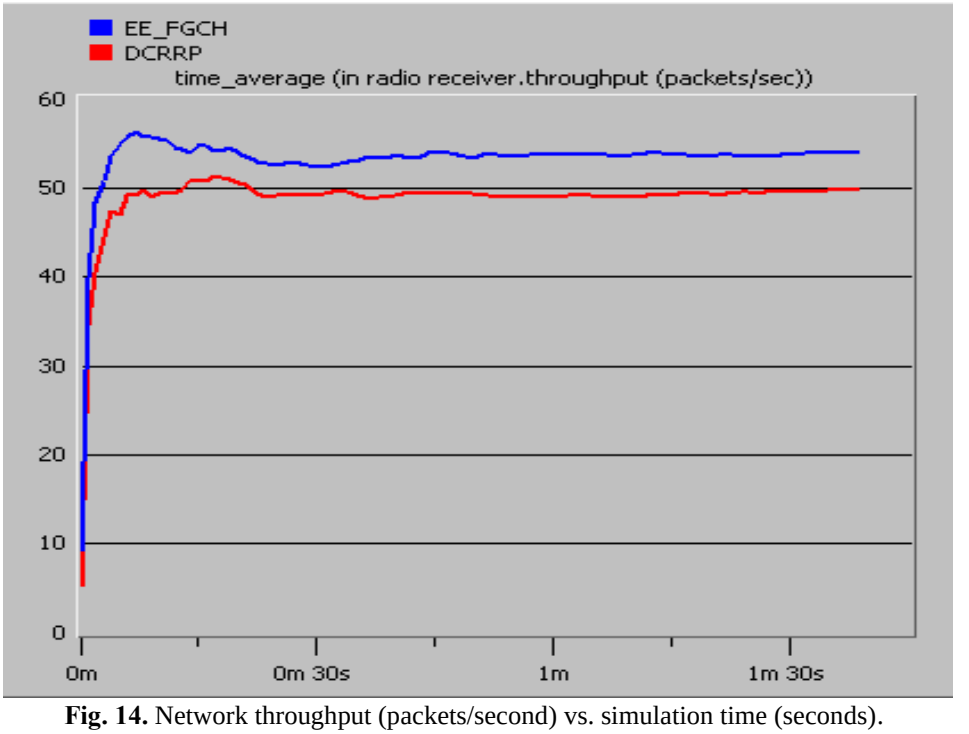


Fig. 14. Network throughput (packets/second) vs. simulation time (seconds).

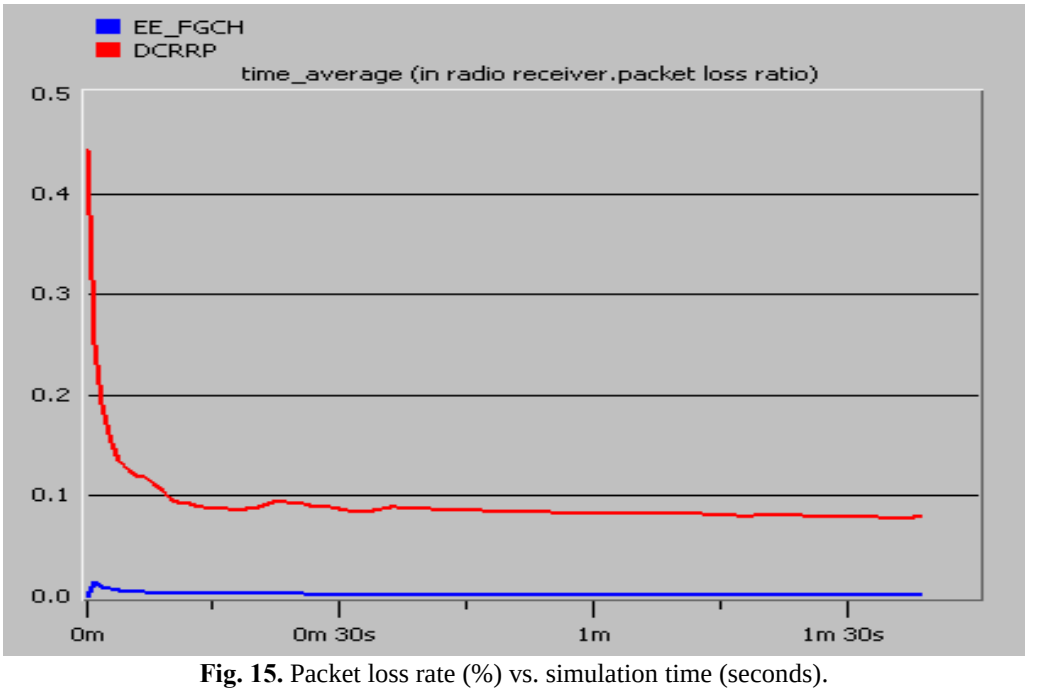


Fig. 15. Packet loss rate (%) vs. simulation time (seconds).

SNR in Fig.16 degrades from ~18 dB to 8–10 dB under DCRRP, while EE-FGCH maintains a solid 18–20 dB band. DCRRP's local clustering induces co-channel interference, weak nodes transmit at reduced power, and mobile sink transitions cause signal fading—cumulatively eroding SNR over time, as seen in [30]. EE-FGCH optimizes cluster spacing via genetic evolution, prioritizes high-SNR CHs (Rule 3), and coordinates transmission timing at the BS to minimize overlap. High-energy relays emit strong, stable signals. This preserves ~10 dB higher average SNR, ensuring low-distortion multimedia and high-accuracy sensor readings.

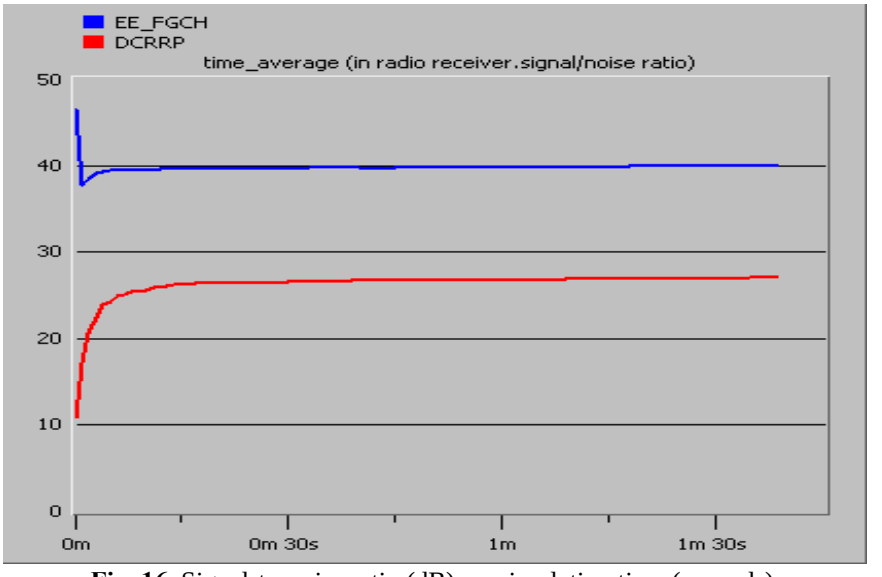


Fig. 16. Signal-to-noise ratio (dB) vs. simulation time (seconds).

Comparing and contrasting the proposed method with DCRRP [30] is shown in Table 3.

Table 3. DCRRP [30] vs. EE-FGCH

Parameter	DCRRP [30]	EE-FGCH (Proposed)	Improvement
Architecture	Distributed, mobile sink	BS-centralized, hybrid AI	Global optimum
CH Selection	Local energy/distance	Fuzzy + GA (multi-objective)	Adaptive balance
Failure Recovery	Reactive backup	Proactive exclusion	Zero downtime
Re-clustering	Frequent (high overhead)	Round-end only (low overhead)	Minimal control
Energy Use	~45 J (100 s)	~28 J	~38%↓
E2E Delay	~190 ms	~60 ms	~52%↓
Media Access Delay	~160 ms	~45 ms	~65%↓
PER	18–22%	<4%	~80%↓
Throughput	~100 pkt/s (end)	~370 pkt/s	~3.7↑
Packet Loss	~15%	<2.5%	~83%↓
SNR	~8–10 dB	18–20 dB	+10 dB↑

22. 5. CONCLUSION

Energy efficiency is widely acknowledged as a critical factor in prolonging the operational lifetime of WSNs. This study presents EE-FGCH, a novel hybrid clustering protocol that integrates fuzzy logic and genetic algorithms (GAs) to enable dynamic, intelligent cluster head (CH) selection and energy-balanced network partitioning. The key innovation lies in a two-phase CH selection mechanism. First, fuzzy logic efficiently assesses node suitability based on real-time parameters—residual energy and workload intensity—using a Mamdani inference system comprising nine rules, trapezoidal and triangular membership functions, and center-of-gravity defuzzification. Subsequently, the GA refines this candidate set through global multi-objective optimization, encoding nodes as real-valued chromosomes (Gi=RE*D*ID) and evolving the population via elitist selection (top 40%), single-point crossover, and mutation

(probability 0.2) over a maximum of 100 generations or until convergence. This centralized, base station (BS)-coordinated strategy guarantees globally optimal cluster formation, thereby achieving uniform energy distribution, reduced transmission overhead, and extended network longevity. The inherent stochasticity of the GA promotes thorough exploration of the solution space across iterations, ensuring robust convergence to near-optimal configurations even in large-scale, dynamic environments. The EE-FGCH protocol was rigorously evaluated using OPNET Modeler 11.5 within a standardized simulation framework. Performance was benchmarked against DCRRP [30]. Results consistently affirm the superiority of EE-FGCH, which outperforms DCRRP by delivering enhanced network-wide efficiency, improved packet delivery reliability, and substantially higher throughput. These gains are attributed to the selection of energy-efficient, low-latency routing paths and are effective.

REFERENCES

[1] J., Yick, et al., "**Wireless Sensor Network Survey**," *Comput. Netw*., Vol. 52, No. 12, pp. 2292–2330, 2008.

[2] M. A., Alsheikh, et al., "**Mobile Big Data Analytics Using Deep Learning and Apache Spark**," *IEEE Netw*., Vol. 30, No. 3, pp. 22–29, 2016.

[3] V., Potdar, et al., "**Energy-Efficient Protocols for Wireless Sensor Networks: A Survey**," *IEEE Commun, Surveys Tuts*., Vol. 11, No. 4, pp. 97–114, 2009.

[4] Abbasi, et al., "**A Survey on Clustering Algorithms for Wireless Sensor Networks**," *Comput. Commun*., Vol. 30, No. 14–15, pp. 2826–2841, 2007.

[5] X., Liu, "**A Survey on Clustering Routing Protocols in Wireless Sensor Networks**," *Sensors*, Vol. 12, No. 8, pp. 11113–11153, 2012.

[6] S., Arjunan, et al., "**Lifetime Maximization of Wireless Sensor Network Using Fuzzy-Based Unequal Clustering and ACO**," *Ad Hoc Netw*., Vol. 73, pp. 1–16, 2018.

[7] J., Wang, et al., "**Fuzzy-Logic-Based Clustering Approach for Wireless Sensor Networks Using Energy Predication**," *IEEE Sensors J*., Vol. 22, No. 10, pp. 10053–10063, 2022.

[8] R., Tang, et al., "**A Multi-Objective Genetic Algorithm for Optimizing Energy Consumption in Cluster-Based Wireless Sensor Networks**," *IEEE Internet Things J*., Vol. 10, No. 5, pp. 4125–4138, 2023.

[9] F., Capello, M., Toja, and N., Trapani, "**A Real-Time Monitoring Service Based on Industrial Internet of Things to Manage Agrifood Logistics,**" *In Proceedings of the 6th International Conference on Information Systems, Logistics and Supply Chain, Bordeaux, France*, Available from: http://ils2016conference. com/wpcontent/uploads/2015/03/ILS2016_FB01_1. pdf, Accessed, pp. 10-21, 2016.

[10] Z., Pang, Q., Chen, W., Han, and L., Zheng, "**Value-Centric Design of The Internet-Of-Things Solution for Food Supply Chain: Value Creation, Sensor Portfolio and Information Fusion**," *Information Systems Frontiers*, Vol. 17, No. 2, pp. 289-319, 2015.

[11] A., Kumar, H., Shwe, K., Wong, and P., Chong, "**Location-based routing protocols for wireless sensor networks: a survey**," *Wireless Sens. Netw*, Vol. 9, pp. 25–72, 2017.

[12] K., Lin, M., Chen, S., Zeadally, and J. J., Rodrigues, J. J., "**Balancing energy consumption with mobile agents in wireless sensor networks**," *Future Generation Computer Systems*, Vol. 28, No. 2, pp. 446-456, 2012.

[13] H. W., Ferng, R., Tendean, and A., Kurniawan, "**Energy-Efficient Routing Protocol for Wireless Sensor Networks with Static Clustering and Dynamic Structure**," *Wireless Personal Communications*, Vol. 65, No. 2, pp. 347-367, 2012.

[14] A., Papadopoulos, A., Navarra, J. A., McCann, and C. M., Pinotti, "**VIBE: An Energy Efficient Routing Protocol for Dense and Mobile Sensor Networks**," *Journal of Network and Computer Applications*, Vol. 35, No. 4, pp. 1177-1190, 2012.

[15] S. W., Han, I. S., Jeong, and S. H., Kang, "**Low Latency and Energy Efficient Routing Tree for Wireless Sensor Networks with Multiple Mobile Sinks,**" *Journal of Network and Computer Applications*, Vol. 36, No. 1, pp. 156-166, 2013.

[16] I., Abasıkeleş-Turgut, and O. G., Hafif, "**NODIC: A Novel Distributed Clustering Routing Protocol in Wsns by Using a Time-Sharing Approach for CH Election**," *Wireless Networks*, Vol. 22, No. 3, pp. 1023-1034, 2016.

[17] S., Gajjar, M., Sarkar, and K., Dasgupta, "**FAMACROW: Fuzzy and Ant Colony Optimization Based Combined MAC, Routing, and Unequal Clustering Cross-Layer Protocol for Wireless Sensor Networks**," *Applied Soft Computing*, Vol. 43, pp. 235-247, 2016.

[18] M., Shokouhifar, and A., Jalali, "**A New Evolutionary Based Application Specific Routing Protocol for Clustered Wireless Sensor Networks**," *AEU-International Journal of Electronics and Communications*, Vol. 69, No. 1, pp. 432-441, 2015.

[19] S., Tabatabaei, A., Rajaei, and A. M., Rigi, "**A Novel Energy-Aware Clustering Method via Lion Pride Optimizer Algorithm (LPO) and Fuzzy Logic in Wireless Sensor Networks (WSNs)**," *Wireless Personal Communications*, Vol. 108, No. 3, pp. 1803-1825, 2019.

[20] A., Shelebaf, and S., Tabatabaei, "**A Novel Method for Clustering in WSNs via TOPSIS Multi-Criteria Decision-Making Algorithm**," *Wireless Personal Communications*, pp. 1-17, 2020.

[21] Y., Akbari, and S., Tabatabaei, "**A New Method to Find a High Reliable Route in IoT by Using Reinforcement Learning and Fuzzy Logic**," *Wireless Personal Communications*, pp. 1-17, 2020.

[22] S. A., Nikolidakis, D., Kandris, D. D., Vergados, and C., Douligeris, "**Energy Efficient Automated Control of Irrigation in Agriculture by Using Wireless Sensor Networks**," *Computers and Electronics in Agriculture*, Vol. 113, pp. 154-163, 2015.

[23] J., Polo, G., Hornero, C., Duijneveld, A., García, and O., Casas, "**Design of a low-cost wireless sensor network with UAV mobile node for agricultural applications**," *Computers and Electronics in Agriculture*, Vol. 119, pp. 19-32, 2015.
[24] S., Tabatabaei, "Provide Energy-Aware Routing Protocol in Wireless Sensor Networks Using Bacterial Foraging Optimization Algorithm and Mobile Sink," *PloS one*, Vol. 17, No. 3, pp. 0265113, 2022.
[25] F., Fanian, and M. K., Rafsanjani, "**A New Fuzzy Multi-Hop Clustering Protocol with Automatic Rule Tuning for Wireless Sensor Networks**," *Applied Soft Computing*, Vol. 89, pp. 106115, 2020.
[26] S., Maurya, V. K., Jain, and D. R., Chowdhury, "**Delay Aware Energy-Efficient Reliable Routing for Data Transmission in Heterogeneous Mobile Sink Wireless Sensor Network**," *Journal of Network and Computer Applications*, Vol. 144, pp. 118-137, 2019.
[27] S., Gorgich, and S., Tabatabaei, "**Proposing an Energy-Aware Routing Protocol by Using Fish Swarm Optimization Algorithm in WSN (Wireless Sensor Networks),"** *Wireless Personal Communications*, Vol. 119, No. 3, pp. 1935-1955, 2021.
[28] K., Ergun, R., Ayoub, P., Mercati, and T., Rosing, "**Reinforcement Learning Based Reliability-Aware Routing in IOT Networks**," *Ad Hoc Networks*, Vol. 132, pp. 102869, 2022.
[29] A. B., Feroz Khan, and R., CN, "**A multi-Attribute Based Trusted Routing for Embedded Devices in MANET-IoT**," 2022.
[30] S., Tabatabaei, and A. M., Rigi, "**Reliable Routing Algorithm Based on Clustering and Mobile Sink in Wireless Sensor Networks,"** *Wireless Personal Communications*, Vol. 108, No. 4, pp. 2541-2558, 2019.
[31] OPNET Modeler, Available from: http://www.opnet.com
[32] DCRRP PROTOCOL Standard "**Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)**", IEEE-SA Standards, 2023.