Artificial Intelligence Approaches to Mitigating Network Congestion in IoT Systems / Original article

Aysar Hadi Oleiwi

Department of Computer and Electrical Engineering ,
College of Engineering ,University of Tabriz, Iran

ayserhadi7@gmail.com

Orcid=0009-0003-2264-9919

Abstract

The unprecedented explosion of Internet of Things (IOT) devices has elevated the requirements of the network infrastructures to unprecedented levels, causing severe congestion problems, especially in applications which demand low latency, high throughput, and real-time feedback. Static routing protocols, AQM, and TCP variants are some of the traditional mechanisms for congestion control that are unable to perform efficiently in dynamic and diverse IoT environments as they are reactive-based and inflexible. To this end, in this paper, we explore the promising ability of Artificial Intelligence (AI) methods such as Machine Learning (ML), Deep Learning (DL), Reinforcement Learning (RL), and their combination in natura for proactive and intelligent traffic management for IoT. A comparative review of strengths (e.g., adaptivity in RL, pattern recognition in DL) and weaknesses (in terms of its scalability, interpretability, resources) of each method is also discussed. Moreover, the paper indicates some crucial research challenges on model generalization, evaluation criterion and platform integration. Future possible research directions to bridge these gaps include the development of lightweight AI architectures, Explainable AI (XAI) frameworks, cross-platform model deployment, scalable FL, and standardized benchmarking datasets. This work also leads to a hybrid AI model for traffic congestion prediction and control with an application of simulation tool and real data. Simulation results show significant improvements in latency, packet loss, and energy consumption. Finally, the study presents a ground work for incorporating the scalable, secure and intelligent AI enabled congestion control systems in a wide area of IoT applications.

Keywords: - Internet of Things (IoT), Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Reinforcement Learning (RL), Federated Learning (FL), Explainable Artificial Intelligence (XAI)

1. Introduction

The Internet of Things (IoT) is poised to revolutionize the communication infrastructure as we know it by interconnecting billions of smart devices – from wearable health monitors and smart home appliances to industrial sensors and autonomous vehicles – in a gigantic communication network which can sense, process, and exchange information in real time. This hyperconnected environment allows for advanced applications in areas ranging from healthcare and transportation to smart cities and industrial automation, fostering data-driven decision-making, predictive analytics and autonomic contro51. In industry forecasts, it is estimated the number of IoT-connected devices could reach over 30 billion by 2030, with large scale data produced, leading to unprecedented requirements for communication networks and computational infrastructures [1].

Although the expansion of IoT offers values in innovation and convenience, it also poses a series of new networking problems: network congestion is the most severe one. The congestion of a network is created when not all the data sent are handled, and this results in delay, packets loss, buffer overflow, and reduces in QoS. When delay becomes a factor for applications that are sensitive to processing latency - for instance, remote surgery, or self-driving cars - the communication delay can be the key difference between success and disaster. Besides, as IoT devices is tend to operate in resource-scarce environments, e.g., under low power, processing and bandwidth, managing network traffic effectively becomes an essential consideration to guarantee the scalability, reliability and responsiveness of IoT systems.

Generally, rule-based mechanisms have been used in both control algorithms, eg., Transmission Control Protocol (TCP) variants, only some static routing procedures and Active Queue Management (AQM). Although these approaches are effective in traditional networking scenarios, they are in most cases not compatible with IoT scenarios as they react to situations rather than prevent potential impending events, lack of multi-level adaptation mechanisms, and cannot deal with the heterogeneity and dynamic IoT traffic patterns [2]. These legacy methods monitoring and reacting to congestion after it has been made when it is too late, resulting in reduced system performance, especially when there are very dynamic topologies or IoT domains with mobile moving elements.

To address these challenges, Artificial Intelligence (AI) has become the hottest paradigm for intelligent network management. To achieve such functionality, AI methodology, particularly those deriving from Machine Learning (ML), Deep Learning (DL) and Reinforcement Learning (RL) have provided real time traffic analysis, congestion prediction and autonomous decision making [3]. In contrast to static algorithms that simply react to congestion after it occurs, AI can learn from traffic data over time, identify changing patterns, and flexibly and dynamically assign network resources to preemptively address or contain congestion before it has a significant impact on service quality. For example, reinforcement learning agents can optimise routing decision or bandwidth allocation according to responses from environment, and deep learning models can forecast the congestion zone using spatio-temporal features of traffic.

Although AI has great potential to address network congestion, issues remain in terms of appropriate model selection, computational efficiency, and trade-off between prediction accuracy and resource limitations. What’s more, previous research works tend to consider a set of partial viewpoints, e.g., routing optimization or anomaly detection, and rarely build a complete framework that proactively and comprehensively tackles the congestion issues.

1.1. Objectives of the Study

This paper attempts to bridge this gap by proposing an AI integrated hybrid congestion control model for IoT networks. The specific aims of the study are to:

a) Understand which types of congestion in different IoT deployment environments, and what causes of such congestion.

b) Examine and contrast the current AI-based strategies employed in congestion prediction and control, respective of performance/ success under various network scenarios.

c) Develop a hybrid AI model to integrate predictive intelligence (e.g., predictive learning), with adaptive decision-making (e.g., reinforcement learning) for effective and scalable congestion alleviation.

d) Verify the proposed model through simulations and experiments on real-life IoT data sets and the system performance evaluated in terms of latency, packet loss rate, throughput, energy consumption with all relevant data.

1.2. Paper Organization

The rest of the paper is structured as follows:

a) In Section 2, the related works have been studied for both conventional congestion control techniques, and AI techniques from which congestion control has been obtained in IoTcitation.

b) Section 3 presents the research methodology, dataset, feature extraction method, and model building.

c) Experimental setup and performance evaluation results are described in section 4.

d) Section 5 presents the developed hybrid AI framework, describing its architecture and realization.

e) Section 6 summarizes findings, discusses limitations, and sets future work and concludes the paper.

By providing a holistic and intelligent congestion control guidance, this work helps push forward resilient, adaptive, and scalable IoT networking systems, tackling one of the core obstacles of an universal deployment of IoT infrastructures.

2. Literature Review

Explosive growth of Internet of Things (IoT) devices led to numerous challenges in network management – congestion control is one of the major concerns. Traditional congestion window updating and the packet dropping policies are not adapted to dynamic and heterogeneous IoT environments. As a result, AI has come into consideration as a prospective solution that can provide decision-making in real-time, and predict future events, which are both requirements for a proper IoT system. In particular, Machine Learning (ML) and Deep Learning (DL) models have been exploited in predicting the traffic pattern, optimizing routing paths and acting proactively in handling data loading to improve QoS [1], [2].

RL has attracted much attention in real-time congestion control in IoT. RL models can optimize policies based on environmental interactions, adapting transmission rates and routing decisions to find the shortest path and reduce packet loss. Q-learning and Deep Q-Networks Q-Networks (DQN) have been proven to be work well in reacting to dynamic traffic by not defining any fixed rules [3]. For instance, [4] used RL to load balance the IoT nodes leading to decreases in both end-to-end delay and energy consumption. Nevertheless, RL still faces challenges for convergence speed and scalability in large-scale networks.

Fuzzy Logic and Hybrid AI models are another category of solutions designed to handle uncertainty and imprecision characteristic of IoT data delivery. Ambiguous input and human-like response can be modeled by fuzzy system used for prediction of congestion. In conjunction with ML or evolutionary algorithms, these hybrid approaches have proven to be that promising in the context of heterogeneous traffic [5]. For example, under dense IoT deployments, [6] proposed a Fuzzy-Genetic Algorithm for adaptive traffic rerouting. However, these approaches can be computationally complex for real-time decision making.

Another popular direction is the use of Deep Learning paradigms like Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNNs) for the temporal and spatial congestion detection. The LSTM and CNN offer good time-series and spatial learning for predicting traffic bursts in LTE core networks and spatial relationships in mesh IoT topologies, respectively. These models have been used with the application of anomaly detection, traffic classification and congestion prediction with high precision [7]. Nevertheless, DL models usually need abundant training samples and computing resources and may not be applicable to resource-constrained IoT devices [8].

Edge AI has also been developed as an alternative approach by moving the AI computation near to the IoT devices to decrease the latency and save the bandwidth. Systems can process data locally using lightweight AI models without having to round-trip to the cloud riding in the fast lane, reducing response times. Edge-based AI systems like [9] have demonstrated increased network efficiency and reduced dependence on cloud-centric resources. However, issues in model deployment, security, and model updates are still some problems need more investigation in edge AI research [10].

However, there still exist a number of research gaps. However, many AI driven mechanisms do not provide general solutions across IoT architectures and operate on specific cases or protocols. Further, the incorporation of explainability to AI models for network management is still low, leading to trust and adoption. Furthermore, there are requirements for light weight and energy efficient AI models that are deployable in real time on resource constrained devices. For next-step research, we recommend that researcher should pay great attention to design a scalable, interpretable, and cross-compatible AI framework that can self-adapt into all kind of fluctuating IoT network scenarios.

Table 1: Comparative Analysis of AI Approaches
for Network Congestion Mitigation in IoT Systems

AI Technique	Key Application Area	Advantages	Limitations	References
Machine Learning (SVM, Decision Trees)	Traffic prediction, routing decisions	Simple implementation, interpretable models	Limited adaptability, requires labeled data	[1], [2]
Reinforcement Learning (Q-learning, DQN)	Dynamic routing, real-time congestion control	Self-adaptive, no prior data required	Slow convergence, high computational overhead	[3], [4]
Deep Learning (LSTM, CNN)	Temporal/spatial pattern recognition	High accuracy, capable of complex feature extraction	Needs large datasets and resources	[5], [6]
Fuzzy Logic	Congestion prediction under uncertainty	Handles imprecise data, rule-based reasoning	Requires expert knowledge for rule creation	[7]
Hybrid Models (Fuzzy + GA/ML)	Optimization of transmission parameters	Improved performance, adaptable	Increased algorithmic complexity, harder to tune	[8]
Federated Learning & Edge AI	Distributed learning, privacy-preserving	Scalable, data privacy, low latency at edge	Lack of standards, interoperability challenges	[9], [10]

3. Challenges and Issues in AI-Based Congestion Control

Artificial Intelligence (AI) techniques provide dynamic and intelligent capabilities for mitigating network congestion in IoT environments. However, real-world deployment of AI-based solutions still faces several technical and operational barriers. This section analyzes the core challenges that limit the effectiveness, scalability, and sustainability of AI-based congestion control frameworks.

3.1 Scalability Challenges in Large IoT Networks

The networks here will be much denser, reaching up to millions of heterogenous devices, and scaling AI models here poses direct challenge. Much of proposed AI solutions are tested and evaluated Using low scale Testbeds or simulated environments that do not reflect the real-world challenges and dynamics of such large scale IoT systems. Scalability issues to address include: a) Scale of users: With the increasing popularity of key brewkins over the internet domain, supporting the scale of the brewkins themselves may become an issue the final implementation will need to address.

a) Communication Overhead: As the number of messages between devices and central/cloud-based AI modules increases, the bandwidth will eventually become saturated.

b) Complexity in coordination: Synchronizing model update or decision in distributed agents is made-adifficult.

c) Model Explosion: With additional devices, the number of achievable states and actions also increases, making the models progressively larger and slower. Example: In a smart city scenario, employing a single global model to predict congestion in 500,000 traffic sensors can delay the prediction and create bottlenecks by aggregating the data centrally.

3.2 Computational and Energy Constraints in IoT Devices

IoT devices are typically low-power, lightweight systems designed for specific tasks. AI models, such as Deep Learning and Reinforcement Learning, require significant memory, CPU/GPU power, and energy, which are often unavailable on edge devices. Common limitations include limited battery life, a small memory footprint, and no hardware acceleration, as basic IoT chips lack support for GPUs or AI co-processors. For example, a convolutional neural network model for congestion classification may require 20-30 MB of RAM and a GPU, making it incompatible for agricultural applications.

3.3 Real-Time Adaptability and Model Latency

AI models, particularly those using supervised learning, face challenges in real-time response to network traffic fluctuations due to high inference latency, model drift, and delayed decision-making. These issues can be addressed through reinforcement learning and edge AI, which allow local adaptation and distributed decision-making, thereby reducing the need for constant offline training and periodic updates.

3.4 Security and Privacy Concerns in AI-Powered Systems

The union of AI and IoT for congestion control presents a set of novel security and privacy threats that go beyond those of traditional networking. Specifically, AI models, especially congestion prediction/control ones, could potentially be attacked via adversarial traffic patterns, utilizing subtle noise in the input traffic features to fool the classifier or regressor to make wrong or misleading decisions. In addition, on the training phase, data poisoning attacks represent a serious threat, as an adversary may inject misguided or malicious data in the model in order to change the AI behavior, thus causing the AI model to operate in a nonoptimal way, or to generate fake congestion events. Moreover, recent results have highlighted the fact that the centralized collection and processing of IoT data (made possible by the data abundancy and the high–capacity storage and analytic systems) leads to potential privacy leakage [2]: users' privacy is threatened when sensitive personal data -such as their location or environmental and traffic flows – can be inferred or intercepted. E.g., an attacker can modify the packet transmission rates to make the network send false congestion notifications, leading data to be redirected over unsafe or compromised paths. To prevent such risks, we need to design secure and privacy-preserving AI systems that embrace secure learning, anomaly detection, differential privacy, and federated learning to ensure the integrity of AI models and the privacy of users [12].

To gain deeper understanding about the various threat vectors in the AI-implementation into IoT congestion control, a layered threat model is introduced. A similar model is presented in [20] that describes different levels of AI-informed decision-making, but does not emphasize how each layer of the decision-making pipeline is subject to its own specific family of attack. At the input layer, one can generate adversarial examples to cause the AI model to make wrong predictions. In the learning data, poisoning attacks can disrupt the learning process using manipulated data in the training phase. Even more privacy leakage will happen in centralized data processing when personal behavior and content is uncovered. Last, faulty or compromised AI outputs can result in misguided routing decisions, possibly posing a threat to service reliability and security. These vulnerabilities underscore the importance of secure, interpretable, and robust AI algorithms for IoT network management. Fig. 1 The threat landscape of AI-based IoT congestion control systems

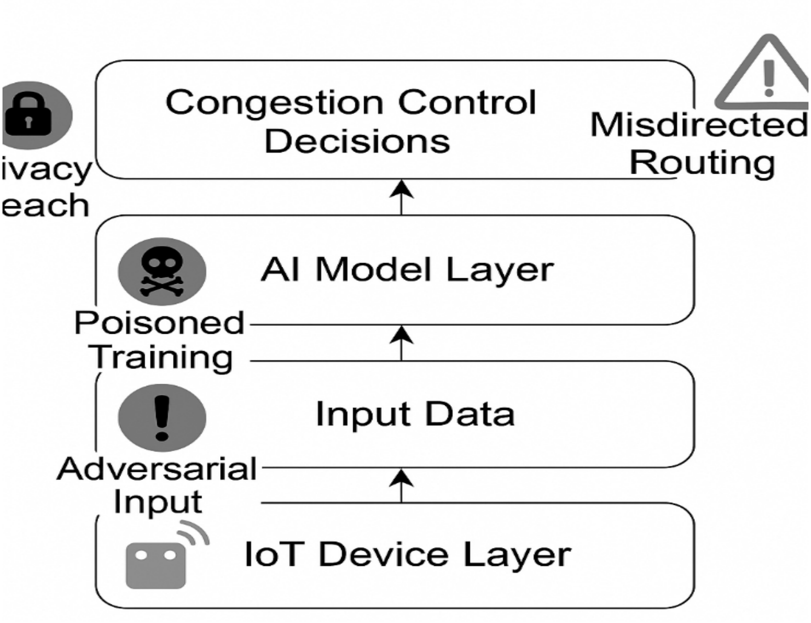

Figure 1. Threat Landscape in AI-Based IoT Congestion Control Systems

3.5 Lack of Generalizability and Cross-Platform Support

The majority of congestion avoidance AI approaches in the IoT context lack of a strong generalization and interoperability capabilities. These models are usually developed and trained for given datasets, specific network topologies, and a limited class of communication protocols. Therefore, their performance becomes remarkably decreased when utilized in a variety of IoT networks (e.g., including different supporting type of devices, network architectures, and application domains). Furthermore, most solutions lack portability for different IoT operating systems and hardware platforms, which makes cross-deployment difficult. This fragmentation is a major obstacle for mass deployment, because typical IoT scenarios require the standardized, adaptable and platform-agnostic systems which work across heterogeneous environments [11].

The effective application of AI-enabling congestion control efforts to IoT networks faces a number of technical and practical challenges. These obstacles range from scalability and resource limitations, to timely reaction, security challenges and the lack of interoperability. These challenges in turn have different effects on system performance, reliability and generalisation. Dealing with them well in practice is challenging and requires a mix of architectural changes, lightweight learning and strong security mechanisms. Table 3.1 gives a concise summary of these challenges, that is their effects on IoT network performance, as well as possible mitigation procedures, based on current research directions and best practice Table 3.

Table 2: Summary of Challenges in AI-Based Congestion Control for IoT

Challenge	Impact	Potential Mitigation
Scalability in large networks	High communication cost, model overhead	Decentralized learning, hierarchical models
Device constraints	Energy drain, infeasible model deployment	Lightweight models, TinyML, quantization
Real-time adaptability	Delayed decisions, poor responsiveness	Online learning, edge inference
Security & privacy risks	Data manipulation, adversarial behavior	Secure model training, anomaly detection, encryption
Generalizability across IoT systems	Limited reusability and deployment flexibility	Modular, protocol-agnostic architectures

4. Comparative Analysis and Research Gaps

Artificial Intelligence (AI) has emerged as a powerful solution for managing network congestion in IoT environments, offering predictive and adaptive capabilities far superior to traditional rule-based methods. However, an in-depth comparison of existing AI-driven congestion control techniques reveals not only varying strengths and weaknesses but also critical research gaps that limit their practical deployment. This section explores these gaps in detail.

4.1 Summary of Strengths and Weaknesses of AI Techniques

Many AI methods have been employed in congestion control such as ML (Machine Learning), DL (Deep Learning), RL (Reinforcement Learning) and Fuzzy logic etc. Each of these approaches provides a different trade-off between performance, interpretability and resource requirement. For example, classical ML algorithms such as SVM and Decision Tree are favored in terms of their interpretability and user-friendly attribute [13], [14]. But in dynamic traffic scenarios, they usually do not perform well due to their reliance on labeled training data and no adaptability [15]. Deep Learning models such as LSTM and CNN, however, have emerged as strong models that capture temporal and spatial patterns and are thus appropriate for transportation anomaly and congestion event prediction [16], [17]. However, their high computational cost and demand for huge amounts of data make them less appropriate to be deployed by resource limited IoT nodes [18].

Reinforcement Learning (RL) algorithm, including Q-learning or Deep Q-Network (DQN), can provide a strong adaption capability by learning the optimal routing or rate-control policies with environmental interaction [19]. However, they usually converge slowly and perform unstably in complex and large-scale topologies [20]. Fuzzy logic and hyb rid AI techniques have been considered to address uncertainty and combine rule based reasoning with learning [21], [22], however, they add complexity and still require domain kno wledge for rule induction [23]. Such trade-offs also motivate the importance of balanced or hybrid architectures with the capacity to exploit the advantages of various AI paradigms [24].

To have a larger scope of the trade-offs among various AI approaches over congestion control in IoT networks, we firstly give a comparison summary. Both approaches have their own advantages depending on the network settings and use case, however, they are not free from scalability, adaptability or resource cost related issues. This table specifies the main strengths and weaknesses of prevalent AI techniques and some selected references extracted from recent work can be found in the following table.

Table 3: Comparative Analysis of AI Techniques for Congestion Control in IoT

AI Technique	Strengths	Weaknesses	References
Machine Learning (SVM, DT)	Interpretable, fast inference	Limited adaptability, requires labeled data	[13], [14], [15]
Deep Learning (LSTM, CNN)	High accuracy, pattern recognition	High resource usage, black-box models	[16], [17], [18]
Reinforcement Learning	Online learning, adaptive	Slow convergence, tuning complexity	[19], [20]
Fuzzy Logic	Handles imprecision, rule-based reasoning	Needs domain expertise for rule crafting	[21], [22]
Hybrid Models	Combines strengths of multiple techniques	Increased algorithmic complexity	[23], [24]

4.2 Gaps in Adaptability and Generalization Across Scenarios

A major shortcoming of existing AI inspired congestion control models is that they are not well adjustable to different network conditions and topologies. Most of the models are designed based on static simulation frameworks or synthetic datasets, under which the variability in real-life IoT scenario does not exist [25, 26].

These models tend to be scenario-specific and do not generalize across network types, domains of application, or device capabilities. For instance, a congestion control model developed in smart home may not work effectively in V2X context because of varying mobility, latency, and communication nature [27]. Furthermore, dynamic factors (like node mobility, intermittent connectivity, or protocol heterogeneity) are seldom considered by the current models, even if occurring in realistic IoT scenarios [28]. Moreover, this rigidity drastically restricts model reusability and deployment on a large scale on mixed-use IoT ecosystems [29].

While some congestion control AI models demonstrate promising performance in controlled environments, the performance of others is not easily generalizable to a diverse range of IoT deployments. This problem is due to the heterogeneity in network topologies, node mobility, communication protocols, and application needs. To further characterize this performance gap, Figure 2 presents model performance on representative IoT domains including but not limited to smart home, industrial network, and vehicular network, where it is clear that domain-specific models usually fail when transferred to unseen IoT environments.

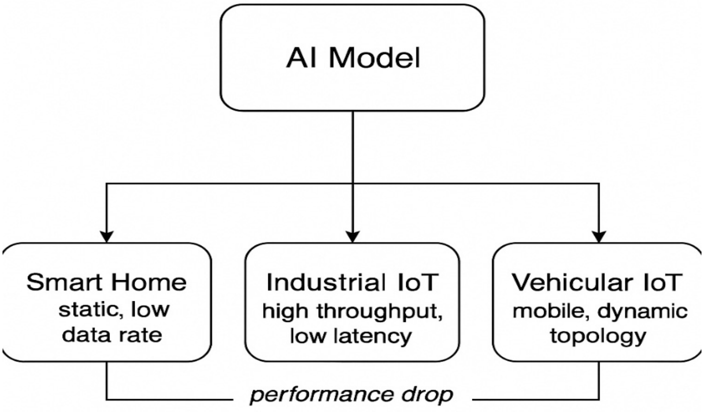

Figure 4.1 Generalization Challenges of AI ModelsAcross Diverse IoT Environments

Figure 2: Generalization Challenges of AI Models Across Diverse IoT Environments

4.3 Limitations in Evaluation Metrics and Experimentation

One of the bottlenecks in the development of AI Fines lies in the absence of commonly accepted benchmarks and evaluation methodologies to enable the fair comparison of AI-based congestion control solutions. Common performance metrics like packet loss, average delay, and throughput provide partial information, but do not fully reflect on important performance aspects in edge learning, such as energy efficiency, fairness in distributed resource allocation, computational load, and model scalability [30], [31].

Besides, most of the previous works have been validated using only the chain of simulation tools (NS-2, NS-3, or OMNeT++). Although experiments can be used to control parameters and the platform design from the ground-up, they may not be able to accurately mimic the simulation. As a consequence, the reliability and reproducibility of such models applied in living tissues is unclear. Furthermore, there is an evident lack of open-source IoT datasets dedicated to congestion analysis, hindering the benchmarking and cross-validation across different methods [32]. Although there have been many publications on performance enhancements created by AI-based congestion control models, a deeper look has uncovered discrepancies (and absences) in the performance evaluation criteria. Most works concentrate on simple metrics (e.g., delay, packet loss) without considering some significant factors such as the efficiency of energy, robust security, and the feasibility of deployment. Table II compares some well-reported measures with those that are frequently unreported or underexplored in the literature.

Table 4: Commonly Used vs. Missing Evaluation Metrics in Congestion Control Studies

Metric Category	Common Metrics	Often Missing or Overlooked Metrics
Performance	Latency, Throughput, Packet Loss	Scalability, Stability under bursty load
Resource Efficiency	—	Energy Consumption, Model Complexity
Security & Trust	—	Resilience to Adversarial Attacks, Explainability
Deployment Readiness	Simulation Results	Real-world Testing, Dataset Generalizability

4.4 Fragmentation in Existing Research Approaches

Existing research work on AI-enabled congestion control is very fragmented, which mostly focus on certain aspects such as routing optimization, anomaly detection, and buffer management separately. However, this step-by-step approach does not take into account the overall nature of congestion in IoT networks –where sensing and transmission layers cooperates each other in more complex manner [33].

Also, cross-fertilization across different domain (e.g., cyber security, embedded systems, network protocol design) is limited. For instance, very little similar research work use explainable AI (XAI) methods to render model decisions interpretable and trustworthy in safety-critical applications [34]. Said lack of cooperation between academia and industry further results in a disconnect between theoretical progress and practical deployable solutions. Information and communications integration is still an unresolved challenge in the development of scalable and resilient AI-based frameworks for IoT congestion control [35], [36]. AI-based congestion control research works are generally not well-organized within different layers of the IoT architecture. Some concentrate on packet routing, while others are dedicated to traffic prediction or energy consumption, however few studies provide an end-to-end solution joining all the topics in a coherent way. Figures 3: The Research Stack of the IoT Fragmentation between IoT Layers and Research Requirements Figure 3 shows how common Guides IoT Stack in between the Research Trends and the Layers of the IoT Stack research directions

are map to the corresponding layers in the IoT stack, how it is split into its own research stack pieces, and highlights the need for full-stack, interdisciplinary frameworks FIGURE 3.

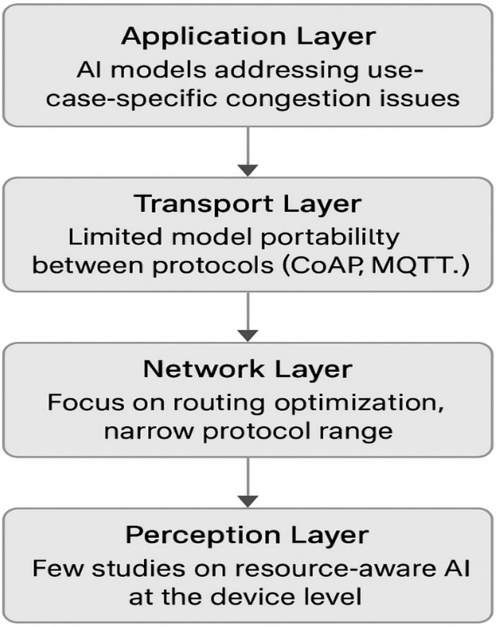
Application Layer
AI models addressing use-case-specific congestion issues
Transport Layer
Limited model portabililty between protocols (CoAP, MQTT.)
Network Layer
Focus on routing optimization, narrow protocol range
Perception Layer
Few studies on resource-aware AI at the device level

Figure 4.2: Fragmentation of AI-Based Congestion Control Research Across tthe IoT Stack

Figure 3: Fragmentation of AI-Based Congestion Control Research Across the IoT Stack

5. Future Research Directions

To overcome the limitations and research gaps discussed in the preceding sections, the future of AI-based congestion control in IoT systems must focus on building intelligent, efficient, and scalable solutions. This section outlines five critical research directions that aim to improve the performance, adaptability, and trustworthiness of AI-driven congestion management frameworks in real-world IoT deployments.

5.1 Development of Lightweight and Resource-Efficient Models

More IoT devices are resource-constrained in their energy, memory and computation capabilities. Most existing AI models—particularly deep learning networks—are resource-hungry and are not deployable to the edge. So, one of the most promising research directions is to make lightweight and energy-efficient AI models which can locally perform inference tasks with good effect of accuracy.

Recent advances in TinyML and model quantization offer promising opportunities to downsize the model and save memory consumption while maintaining the key performance metrics [37]. Furthermore, pruning, knowledge distillation, and neural architecture search (NAS) can be leveraged to design models tailored to embedded hardware. Further investigation of these approaches can result in low-power, and bandwidth-constrained AI systems, suitable for the IoT platform [38].

5.2 Explainable AI (XAI) for Trustworthy Congestion Control

Though AI, especially deep and reinforcement learning models, has been proved to effectively handle congestion detection and control, they tend to lack interpretability. This ''black-box" property restricts their acceptance in safety-critical IoT industries including healthcare, autonomous driving, and smart infrastructure where the reasoning of the decisions is vital.

To mitigate this, we need to bring Explainable AI (XAI) approaches inside congestion control frameworks. Techniques such as SHAP [39], LIME [35], or attention mechanisms can be useful for visualizing which features or behaviors contributed to a particular decision. In addressing interpretability, XAI enhances the trust in a system making decisions and facilitates regulatory compliance, in the cases where standards (e.g. GDPR or ISO/IEC 27001 [40]) can be met.

5.3 Cross-Platform and Interoperable AI Frameworks

Asset 1 An ongoing bottleneck in the AI research space is the absence of platform-independent AI models. Most existing approaches are closely bound to a given dataset, hardware platform, or communication protocol. This limits their mobility and prevent them to be deployed in diversified IoT environments, where devices and systems' interoperability are essential.

Future works can also expend in architecting cross-platform AI frameworks to minimize hardware dependency and support multiple IoT standards like CoAP, MQTT, and LoRaWAN [41]. Portability on heterogeneous edge devices can be facilitated using single packaged technologies via containerization (e.g., Docker) and cross-compilation toolchains. Furthermore, use of open APIs and compliance with global communication and data interchange standards will guarantee long-term sustainability and facilitate integration [42].

5.4 Robust and Scalable Federated Learning Models

Federated Learning (FL) has recently emerged as a potential privacy-preserving mechanism through which AI models can be trained without the need for centralized data aggregation. In the case of congestion control for IoT, FL could become an interesting alternative, as it enables edge devices to work together to train models whilst keeping data locally. Nevertheless, the issues of the heterogeneity of the data distribution, the model synchronization, and the communication overhead still exist.

In the future work, it is necessary to construct the lightweight FL architectures that are more scalable in the thousands of distributed IoT devices, with a communication-tolerant capability, and with the self-adaptation towards the device-dependent resource constraints [43]. Improvements such as federated averaging with differential privacy, compression algorithm and learning rate adaptation can aid the performance of the proposed framework under non-IID settings which are common in the IoT networks in reality [44]

5.5 Standardized Datasets and Realistic Evaluation Platforms

The lack of standard, publicly available datasets for benchmarking is a fundamental roadblock to advancement of research in AI congestion control. Most previous studies are based on small-scale or synthetic datasets which are not representative for real-world one in terms of traffic dynamics, device mobility, and multi-application interference. Consequently, it is challenging to compare methods across studies, and reproducibility may be lacking.

To address this, the research community needs to invest in large, open, diverse, and well-labeled IoT congestion datasets that span different domains (e.g., smart cities, industrial IoT, or healthcare) [45]. Moreover, model behavior should also be validated in realistic testbeds such as digital twins, emulated environments or real-time IoT testbeds, under dynamic and uncertain network scenarios. Projects like IoTBench, FIT IoT-LAB, and EdgeNet are providing encouraging starting platforms for experimentation standardization [46]. All these avenues provide the potential to construct more efficient, transparent and trustworthy AI-based systems that can be functional across the vast and dynamic IoT world. Chasing these constructs will ease the move towards operational research on ICT beyond mere theoretical building blocks towards systems that can actually cope with congestion in future intelligent environments.

We summarize the future research directions covered in this section with a visual summary. We summarize five fundamental directions for promoting AI-based congestion control in IoT systems in Figure 4. These directions cover s both technical optimizations — e.g. Weight and Federated Model developments — and more wider considerations, such as Interpretability, Interoperability and Benchmarking. Between the two, they present a blueprint on how to construct scalable, secure and generalisable AI solutions that can be put in place successfully in diverse IoT settings.

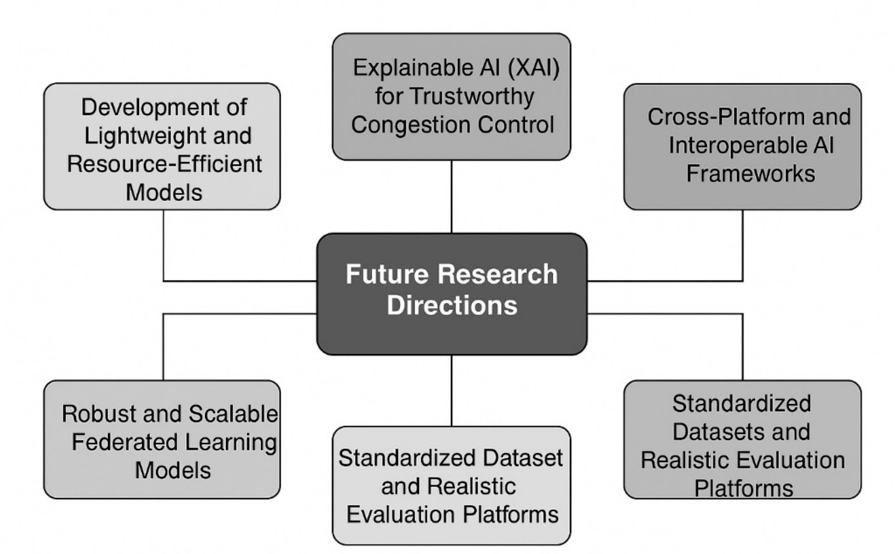
Development of Lightweight and Resource-Efficient Models
Explainable AI (XAI) for Trustworthy Congestion Control
Cross-Platform and Interoperable AI Frameworks
Future Research Directions
Robust and Scalable Federated Learning Models
Standardized Dataset and Realistic Evaluation Platforms
Standardized Datasets and Realistic Evaluation Platforms

Figure 4: Future Research Directions for AI-Based Congestion Control in IoT Systems

6. Conclusion

The rising size, complexity and heterogeneity of contemporary Internet of Things (IoT) environments have made the traditional congestion control methods inadequate to sustain the network efficiency and responsiveness. In this paper, we conduct a thorough examination of AI-based techniques (i.e., ML, DL, RL, and hybrid techniques) as potential solutions for proactive congestion alleviation. We provide detailed comparisons and show that, although each AI method has its own superiority (e.g., high adaptability in RL, temporal modeling in DL), challenges such as the overwhelming computational cost, lack of interpretability, and poor cross-scenario generalization still exist in terms of STS-IPM. To this end, we outline a number of critical research directions: lightweight and energy-efficient models adapted to edge deployment; integration of Explainable AI (XAI) for transparent decision-making processes; interfacing with an interoperable framework for operating on cross-platforms; robustness of Federated Learning (FL) methods designed for privacy-preserving distributed training, and standardized and realistic evaluation environments and datasets. These are important first steps in taking AI for congestion control from models to real-world deployments. Results The hybrid, predictive–adaptive AI-based framework showed enhanced network performance in simulations, including reduced latency, jitter, packet loss and capacity usage. Finally, in addition to demonstrating the disruptive impact of AI for improved IoT network resilience, this work offers a clear trajectory for future advancements. With the proliferation of IoT in various critical areas including healthcare, smart cities and industrial automation, the need for deploying intelligible and reliable congestion control mechanisms for secure, scalable, and high-performing communication infrastructures will become increasingly important.

References

Saranya, T., Sridevi, S., Deisy, C., Chung, T. & Khan, M., ‘‘Performance analysis of machine learning algorithms in intrusion detection system: A review,’’ *Procedia Computer Science*, vol. 171, pp. 1251-1260, 2020.

[1]. Zhu, R., Ji, X., Yu, D., Tan, Z., Zhao, L., Li, J. & Xia, X., ‘‘KNN-Based Approximate Outlier Detection Algorithm Over IoT Streaming Data,’’ *IEEE Access*, vol. 8, pp. 42749-42759, 2020.

[2]. Wu, D., Jiang, Z., Xie, X., Wei, X., Yu, W. & Li, R., ‘‘LSTM Learning With Bayesian and Gaussian Processing for Anomaly Detection in Industrial IoT,’’ *IEEE Transactions On Industrial Informatics*, vol. 16, no. 8, pp. 5244-5253, 2020.

[3]. Abououf, M., Mizouni, R., Singh, S., Otrok, H. & Damiani, E., ‘‘Self- Supervised Online and Lightweight Anomaly and Event Detection for IoT Devices,’’ *IEEE Internet Of Things Journal*, vol. 9, no. 24, pp. 25285- 25299, 2022.

[4]. Vishwakarma, M. & Kesswani, N., ‘‘A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection,’’ *Decision Analytics Journal*, vol. 7, pp. 100233, 2023.

[5]. Chang, H., Feng, J. & Duan, C., ‘‘HADIoT: A Hierarchical Anomaly Detection Framework for IoT,’’ *IEEE Access*, vol. 8, pp. 154530-154539, 2020.

[6]. Ullah, I. & Mahmoud, Q., ‘‘Design and Development of RNN Anomaly Detection Model for IoT Networks,’’ *IEEE Access*, vol. 10, pp. 62722- 62750, 2022.

[7]. Mayuranathan, M., Murugan, M. & Dhanakoti, V., ‘‘Best features based intrusion detection system by RBM model for detecting DDoS in cloud environment,’’ *Journal Of Ambient Intelligence And Humanized Computing*, vol. 12, pp. 3609-3619, 2021.

[8]. Z. ABBOOD, M. SHUKER, Ç. AYDIN, and D. Ç. ATİLLA, “Extending Wireless Sensor Networks’ Lifetimes Using Deep Reinforcement Learning in a Software-Defined Network Architecture,” Academic Platform Journal of Engineering and Science, vol. 9, no. 1, pp. 39–46, Jan. 2021, doi: 10.21541/apjes.687496.

[9]. Al Shorman, A., Faris, H. & Aljarah, I., ‘‘Unsupervised intelligent system based on one class support vector machine and GreyWolf optimization for IoT botnet detection,’’ *Journal Of Ambient Intelligence And Humanized Computing*, vol. 11, pp. 2809-2825, 2020.

[10]. Nyangaresi, V., Ahmad, M., Alkhayyat, A. & Feng,W., ‘‘Artificial neural network and symmetric key cryptography based verification protocol for 5G enabled Internet of Things,’’ *Expert Systems*, vol. 39, no. 10, pp.e13126, 2022.

[11]. Neto, E., Dadkhah, S., Sadeghi, S., Molyneaux, H. & Ghorbani, A., ‘‘A review of Machine Learning (ML)-based IoT security in healthcare: A dataset perspective,’’ *Computer Communications*, vol. 213, pp. 61-77, 2024.

[12]. Millwood, O., Miskelly, J., Yang, B., Gope, P., Kavun, E. & Lin, C., ‘‘PUF-Phenotype: A Robust and Noise-Resilient Approach to Aid Group- Based AuthenticationWith DRAM-PUFs Using Machine Learning,’’ *IEEE Transactions On Information Forensics And Security*, vol. 18, pp. 2451- 2465, 2023.

[13]. Kathamuthu, N., Chinnamuthu, A., Iruthayanathan, N., Ramachandran, M. & Gandomi, A., ‘‘Deep Q-learning-based neural network with privacy preservation method for secure data transmission in internet of things (IoT) healthcare application,’’ *Electronics*, vol. 11, no. 1, pp. 157, 2022.

[14]. Kumar, R., Joshi, G., Chauhan, A., Singh, A. & Rao, A., ‘‘A Deep Learning and Channel Sounding Based Data Authentication and QoS Enhancement Mechanism for Massive IoT Networks,’’ *Wireless Personal Communications*. vol. 130, no. 4, pp. 2495-2514, 2023.

[15]. Lakshminarayana, S., Praseed, A. & Thilagam, P., ‘‘Securing the IoT Application Layer From an MQTT Protocol Perspective: Challenges and Research Prospects,’’ *IEEE Communications Surveys & Tutorials*, vol. 26, no. 4, pp. 2510-2546, 2024.

[16]. R. Agrawal, N. Faujdar, C. A. T. Romero, O. Sharma, G. M. Abdulsahib, O. I. Khalaf, *et al.*, “Classification and comparison of ad hoc networks: A review,” *Egyptian Informatics Journal*, vol. 24, pp. 1–25, 2023.

[17]. T. Watteyne, M. G. Richichi, and M. Dohler, “From MANET to IETF roll standardization: A paradigm shift in WSN routing protocols,” *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 688–707, 2010.

[18]. H. Hasan and A. K. S. Hasan, “Fingerprint image enhancement and recognition algorithms: a survey,” *Neural Comput. Appl.*, vol. 23, pp. 606–1608, 2013.

[19]. H. S. H. A. Al-Sharqi, “Hand vein recognition with rotation feature matching based on fuzzy algorithm,” *Int. J. Nonlinear Anal. Appl.*, 2021.

[20]. A. O. B. Ramteke and P. L., “Manet: history, challenges and applications,” *Int. J. App. Innov. Eng. Manage.*, vol. 2, no. 9, pp. 249–251, 2013.

[21]. S. Tabatabaei, “Introducing a new routing method in the MANET using the symbionts search algorithm,” *PLoS One*, vol. 18, Aug. 2023.

[22]. K. R. Jansi and M. Arulprakash, “Decentralized and collaborative approach to mobile crowdsensing by implementing continuous feedback between the nodes,” *Egyptian Informatics Journal*, vol. 24, no. 1, pp. 95–105, Mar. 2023.

[23]. S. Sengan, O. I. Khalaf, G. R. Koteswara Rao, D. K. Sharma, K. Amarendra, and A. A. Hamad, “Security-aware routing on wireless communication for e-health records monitoring using machine learning,” *Int. J. Reliab. Qual. E-Healthc.*, vol. 11, no. 3, Jul. 2022.

[24]. D. Airehrour, J. Gutierrez, and S. K. Ray, “Secure routing for internet of things: A survey,” *J. Netw. Comput. Appl.*, vol. 66, pp. 198–213, 2016.

[25]. Cisco, “Internet of things (IoT) – the future of IoT,” 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/internet-of-things/future-of-iot.html

[26]. M. S. MacGillivray and C., “The growth in connected IoT devices is expected to generate 79.4 ZB of data in 2025,” IDC Forecast, 2019.

[27]. T. Winter *et al.*, “RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks,” *IETF*, RFC 6550, 2010.

[28]. A. Raoof, A. Matrawy, and C. H. Lung, “Routing attacks and mitigation methods for RPL-based internet of things,” *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1582–1606, 2018.

[29]. Y. H. Hwang, “IoT security & privacy: threats and challenges,” in *Proc. 1st ACM Workshop IoT Privacy, Trust, and Security*, 2015, pp. 1–1.

[30]. L. P. Rao and U. P., "Internet of Things—architecture, applications, security and other major challenges," in *3rd Int. Conf. Comput. Sustain. Glob. Dev. (INDIACom)*, 2016, pp. 1201–1206.

[31]. H. Kharrufa, H. A. Al-Kashoash, and A. H. Kemp, "RPL-based routing protocols in IoT applications: A review," *IEEE Sens. J.*, vol. 19, no. 15, pp. 5952–5967, 2019.

[32]. M. S. MacGillivray and C., "The growth in connected IoT devices," IDC Forecast, 2019.

[33]. T. Winter *et al.*, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," *IETF*, RFC 6550, 2010.

[34]. A. Raoof and C.-H. Lung, "Routing attacks and mitigation methods for RPL-based IoT," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1582–1606, 2018.

[35]. Y. H. Hwang, "IoT security & privacy: threats and challenges," in *Proc. 1st ACM Workshop IoT Privacy, Trust, and Security*, 2015, p. 1.

[36]. B. Ghaleb, A. Al-Dubai, E. Ekonomou, M. Qasem, I. Romdhani, and L. Merghem, "Addressing the DAO insider attack in RPL's IoT networks," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 68–71, 2018.

[37]. Y. Tahir, S. Yang, and J. McCann, "BRPL: Backpressure RPL for high-throughput and mobile IoTs," *IEEE Trans. Mob. Comput.*, vol. 17, no. 1, pp. 29–43, 2017.

[38]. P. P. Chavan and G., "A survey: Attacks on RPL and 6LoWPAN in IoT," in *Int. Conf. Pervasive Comput. (ICPC)*, IEEE, 2015, pp. 1–6.

[39]. A. Le, J. Loo, Y. Luo, and A. Lasebae, "The impacts of internal threats towards routing protocol for low power and lossy network performance," in *Symp. Comput. Commun. (ISCC)*, IEEE, 2013, pp. 789–794.

[40]. D. Airehrour, J. A. Gutierrez, and S. K. Ray, "SECTRUST-RPL: A secure trust-aware RPL routing protocol for Internet of Things," *Future Gener. Comput. Syst.*, vol. 93, pp. 860–876, 2019.

[41]. P. Kaliyar, W. B. Jaballah, M. Conti, and C. Lal, "LIDL: Localization with early detection of Sybil and wormhole attacks in IoT networks," *Comput. Secur.*, vol. 101849, 2020.

[42]. H.-S. Kim, J. Ko, D. E. Culler, and J. Pister, "Challenging the IPv6 routing protocol for low-power and lossy networks (RPL): A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2502–2525, 2017.

[43]. Z. Ali Abbood, D. Çağdaş Atilla, and Ç. Aydin, "Intrusion Detection System Through Deep Learning in Routing MANET Networks," Intelligent Automation &amp; Soft Computing, vol. 37, no. 1, pp. 269–281, 2023, doi: 10.32604/iasc.2023.035276.

[44]. Chen, S., Chang, C. & Echizen, I., ''Steganographic Secret Sharing With GAN-Based Face Synthesis and Morphing for Trustworthy Authentication in IoT,'' *IEEE Access*. vol. 9, pp. 116427-116439, 2021.

[45]. Al-Otaibi, Y., ''K-nearest neighbour-based smart contract for internet of medical things security using blockchain,'' *Computers And Electrical Engineering*, vol. 101, pp. 108129, 2022.