Received 29 August 2022; revised 26 September 2022; accepted 14 October 2022. Date of publication 19 October 2022; date of current version 27 October 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3215731

AI-Enabled Reliable QoS in Multi-RAT Wireless IoT Networks: Prospects, Challenges, and Future Directions

KAMRAN ZIA [1], ALESSANDRO CHIUMENTO [1], AND PAUL J. M. HAVINGA [1,2]

[1] Pervasive Systems Research Unit, University of Twente, 7500 AE Enschede, The Netherlands

[2] ICT, TNO, 2509 JE The Hague, The Netherlands

CORRESPONDING AUTHOR: K. ZIA (e-mail: k.zia@utwente.nl)

This work was supported by the European Project InSecTT under Grant 876038.

ABSTRACT Wireless IoT networks have seen an unprecedented rise in number of devices, heterogeneity and emerging use cases which led to diverse throughput, reliability and latency (Quality of Service) requirements. Fulfilling these diverse requirements in a rapidly changing and dynamic wireless environment is a complex and challenging task. On top of including new technologies and wireless standards, one solution is to deploy cross-layer Design (CLD) and multiple Radio Access Technologies (Multi-RAT) to develop scalable QoS-aware IoT networks. However, the complexity of such solutions is high as it involves complex inter-layer interactions and dependencies and inter-application dependencies in multi-RAT networks. Moreover, the wireless environment is very dynamic, so having an optimal constellation of parameters is a challenging task. In this paper, we address the possibilities of using Artificial Intelligence (AI) and Machine Learning (ML) to address these high dimensional and dynamic problems. Based on our findings, we have proposed a distributed network management framework employing AI & ML for studying inter-layer dependencies and developing cross-layer design, traffic classification and traffic prediction at the edge devices for reliable QoS in multi-RAT IoT networks. A thorough discussion on future directions and emerging challenges related to our proposed framework has also been provided for further research in this field.

INDEX TERMS QoS in IoT networks, AI & ML for cross-layer design, cross-layer optimization, reliable QoS, multi-RAT networks, edge intelligence.

I. INTRODUCTION

THE WORLD is seeing a massive expansion of IoT networks with billions of devices requiring network access. According to the Ericsson mobility report of June 2021, cellular IoT devices would surpass 5 billion devices by 2026 [1]. Similarly, Cisco's Annual Internet Report predicts an increase in networked devices to 29.3 billion with Machine to Machine (M2M) connections reaching 14.7 billion by the end of 2023 [2]. This would lead to a massive increase in wireless network traffic, requiring resilient networks with high user density. The increased traffic from billions of IoT devices needs new studies aiming at expanding network capacity and providing reliable connectivity in such dense IoT networks. Various research bodies and organizations, including 3GPP, 5GPPP, NGMN, IEEE and ITU etc. became aware of the requirements of increased capacity in wireless networks. As such, they have started to employ higher frequency bands (mmWave), network virtualization and spatial multiplexing to meet the growing requirements [3]. Development of 5th generation network (5G) is an important step towards increasing network capacities and supporting massive connectivity. Similarly, IEEE introduced the 802.11ax standard to support high user density in WiFi based IoT networks. These new technologies brought new complexities to provide reliable Quality of Service (QoS) to IoT devices.

Expansion of IoT networks have led to emerging new use cases like smart homes, smart cities, Industry 4.0, Smart Healthcare and autonomous vehicles etc. These use

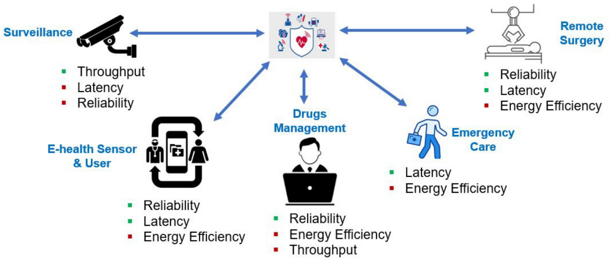

FIGURE 1. QoS Requirements in Healthcare IoT Network.

cases have not only increased the number of devices requiring network connectivity, but also introduced diverse QoS requirements like throughput, reliability, latency and energy efficiency. Present IoT networks have a range of QoS requirements that require efficient management and control of the network. For example, IoT networks in healthcare have many use cases such as remote patient monitoring, eHealth sensors, remote surgeries, emergency care and drugs management. Each use case has different QoS requirements as shown in Figure 1. These requirements keep changing depending on use cases thus requiring sophisticated monitoring and forecasting to assist decision making related to resources prioritisation. Moreover, the success of IoT use cases depends on end user/application's QoS provisioning and it's Quality of Experience (QoE), rather than improving network technical parameters (throughput, packet loss and latency). Meeting these requirements in large networks with a high number of devices and limited resources is a cumbersome task and, therefore, introduces challenges related to QoS provisioning and capacity enhancement in IoT networks. High user density also demands solutions that are scalable to support reliable connectivity to billions of IoT devices.

Among the diverse QoS requirements, many IoT use cases (e.g., self driving cars, Remote Surgery etc.) have stringent latency constraints thus requiring highly responsive networks. This necessitates developing edge intelligent solutions where edge devices are taking their resource management decisions themselves to meet their QoS requirements. Many QoS management solutions proposed in literature involve centralised control mechanism employing Software Defined Networking (SDN) approaches [4], [5]. This forces information to be pushed to the central controller for decision making thus inducing latency. However, low latency IoT use cases demand distributed network management and edge intelligence to improve responsiveness in wireless networks [6]. This distributed network management would also improve the network scalability by offloading computation and decision taking load to the edge devices, thus supporting high density of IoT devices in the network.

Present IoT networks are facing challenges related to user density, limited network capacity, diverse QoS requirements and scalability. To overcome these challenges, AI & ML algorithms can help develop edge intelligence in the network where APs, base stations and IoT devices themselves can take optimal decisions for meeting their QoS requirements and improving their QoE. Since networks operate with a layered structure following the Open System Interconnection (OSI) model, they rely on layer protocols and parameters to meet QoS requirements. Traditional optimization approaches target layer protocols and parameters independently to meet QoS requirements and therefore provide guarantees of QoS for each layer. They do not consider the overall QoS of IoT users/applications and hence cross-layer approaches involving inter-layer dependencies and interactions among layer parameters are employed to provide end-to-end QoS in IoT networks [7]. However, research efforts to develop cross-layer design remain limited due the complex inter-layer dependencies and interactions. Therefore, studies aiming at AI & ML based cross-layer design and optimization are required. To support large user density and increased network capacity, multi-RAT IoT networks employing CLD can be developed, however, they would involve high-dimensional complex decisions related to RAT selection, routing dynamics, interference among different RATs and medium access. Advances in AI & ML have underpinned stronger tools to handle such complexities to enable high capacity, reliable IoT networks with diverse QoS requirements.

In this paper, we have surveyed the cross-layer design and optimization approaches proposed in literature to highlight their results, determine their shortcomings and point out complexities that require AI & ML based algorithms designed to meet diverse QoS requirements. We have discussed the role of AI & ML in studying inter-layer dependencies and joint cross-layer parameter optimizations in single and multi-RAT networks to develop high capacity QoS aware IoT networks. To address the scalability concerns, we have proposed a distributed network management framework that employs CLD at the edge APs in a multi-RAT IoT network to provide reliable QoS to the IoT users/applications. To understand and follow the rest of the paper, the diagrammatic view of this paper's organization is given in Figure 2. The different abbreviations used in the paper are given in Table 1.

II. MOTIVATION

Networks operate in a layered structure following the OSI model. Although, layered structures have been successful in wired networks, they are not well suited in wireless networks due to unreliable link bandwidths and dynamic wireless environment [8]. The diverse QoS requirements in a wireless IoT network can be efficiently met through management and control decisions at different layers of the OSI stack. The choice of different parameters at transport, network, link and physical layers can significantly improve user QoS [9]. Moreover, the sharing of information between adjacent layers as well as beyond, which is also known as cross-layer Design (CLD), can help in insightful decisions that can drive network towards better performance. Especially, the

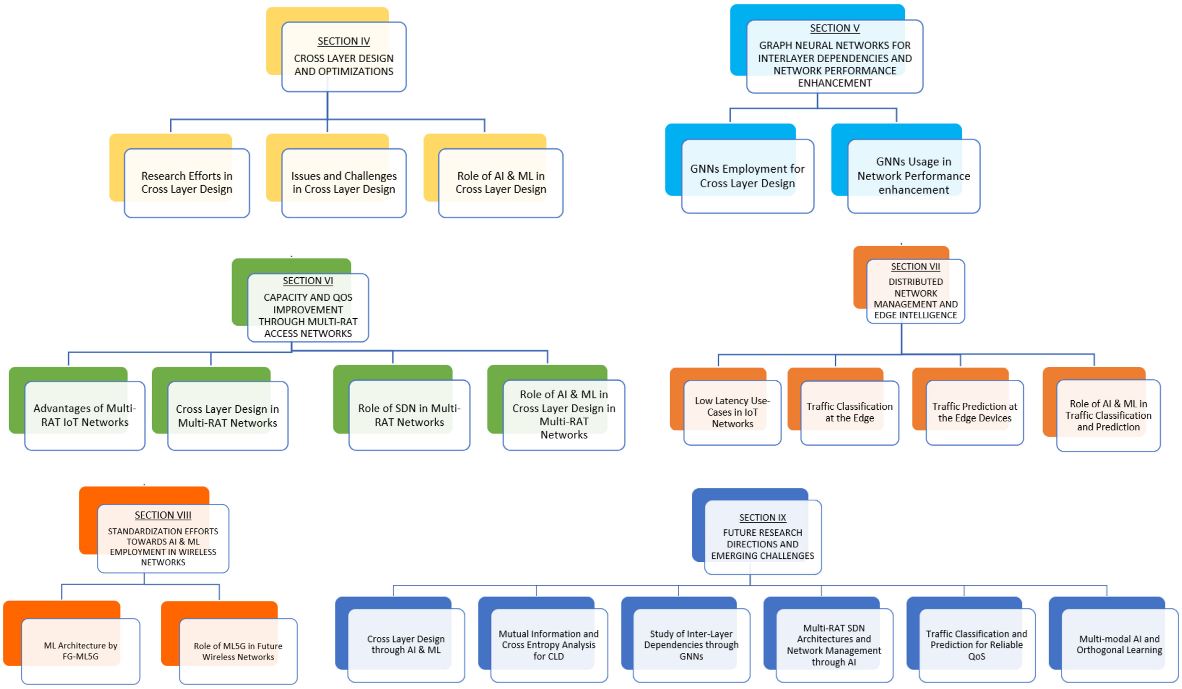
SECTION IV
CROSS LAYER DESIGN AND OPTIMIZATIONS
Research Efforts in Cross Layer Design
Issues and Challenges in Cross Layer Design
Role of AI & ML in Cross Layer Design
SECTION V
GRAPH NEURAL NETWORKS FOR INTERLAYER DEPENDENCIES AND NETWORK PERFORMANCE ENHANCEMENT
GNNs Employment for Cross Layer Design
GNNs Usage in Network Performance enhancement
SECTION VI
CAPACITY AND QOS IMPROVEMENT THROUGH MULTI-RAT ACCESS NETWORKS
Advantages of Multi-RAT IoT Networks
Cross Layer Design in Multi-RAT Networks
Role of SDN in Multi-RAT Networks
Role of AI & ML in Cross Layer Design in Multi-RAT Networks
SECTION VII
DISTRIBUTED NETWORK MANAGEMENT AND EDGE INTELLIGENCE
Low Latency Use-Cases in IoT Networks
Traffic Classification at the Edge
Traffic Prediction at the Edge Devices
Role of AI & ML in Traffic Classification and Prediction
SECTION VIII
STANDARDIZATION EFFORTS TOWARDS AI & ML EMPLOYMENT IN WIRELESS NETWORKS
ML Architecture by FG-ML5G
Role of ML5G in Future Wireless Networks
SECTION IX
FUTURE RESEARCH DIRECTIONS AND EMERGING CHALLENGES
Cross Layer Design through AI & ML
Mutual Information and Cross Entropy Analysis for CLD
Study of Inter-Layer Dependencies through GNNs
Multi-RAT SDN Architectures and Network Management through AI
Traffic Classification and Prediction for Reliable QoS
Multi-modal AI and Orthogonal Learning

FIGURE 2. Diagrammatic View of Paper Organization.

interaction of Physical (PHY) and Medium Access Control (MAC) layers with other layers of OSI stack have a considerably large effect on network performance [10]. Various schemes targeting PHY and MAC layers have been proposed in literature [11], to improve network throughput. Authors in [12], [13] have targeted PHY and MAC layers for guaranteed QoS delivery while optimizing spectrum efficiency in a wireless network. Authors in [14] focused on transport and MAC layers by dynamic estimation of channel for multimedia traffic delivery. TCP performance improvement through cross-layer interaction is a well known example of CLD in transport and MAC layers [10], [15]. Joint optimizations combining power control, scheduling and routing have also been done with limited problem complexities [16], [17]. Authors in [18] proposed an Adaptive Access parameter Tuning (ADAPT) algorithm that focuses on the application, network, MAC and physical layers to improve the energy efficiency in Wireless Sensor Networks (WSN). A similar work in [19] employed cross-layer design targeting transport, network, MAC and physical layers to improve data delivery in WSN. Many researchers have employed CLD for different performance enhancement objectives in wireless networks (rate maximization, routing decisions, energy efficiency, multimedia traffic delivery etc.) however, they have studied CLD with the objective of improving overall network throughput, latency or reliability without considering the user's QoE. As a result, research lacks the study of CLD with the end objective of meeting IoT user's/application's QoS requirements to improve their QoE. Moreover, researchers have limited their efforts to the study of only a few parameters simultaneously for cross-layer optimizations (CLO) [15] due to increased problem complexity that happens with the addition of each new parameter to the optimization framework.

IoT devices and sensors employ different ranges of access technologies depending upon their datarate and energy efficiency requirements. IEEE 802.11, Bluetooth Low Energy (BLE), Long Range Wide Area Network (LoRaWAN), Narrow Band IoT (NB-IoT) and ZigBee are widely used access technologies in IoT networks. As a result, IoT requires network infrastructures that support multiple RATs operating in harmony. Moreover, many IoT use-cases require stringent latency requirements and failure to do so can lead to catastrophic effects. Self driving cars, emergency patient care and remote surgery are few of those important use-cases requiring low latency. To support different IoT devices and their stringent QoS requirements, future IoT networks have to employ multiple Radio Access Technologies (RAT) and distributed network management frameworks. Distributed network management would not only increase responsiveness to satisfy QoS requirements of low latency use cases in industrial and healthcare IoT networks [20], [21], it would also improve network scalability by offloading decision making to edge devices. As a result, multi-RAT networks employing distributed network management can support large user density and can cope up with the rising number of IoT devices in the network.

Although CLD/CLO, multi-RAT IoT networks and distributed network management frameworks can solve the network density, diverse QoS requirements and scalability

TABLE 1. List of abbreviations.

Acronym	Definition
RAN	Radio Access Network
RAT	Radio Access Technology
RRM	Radio Resource Management
3GPP	3rd Generation Partnership Program
NGMN	Next Generation Mobile Networks
IMT	International Mobile Telecommunications
MTC	Machine type Communication
BLE	Bluetooth Low Energy
SDN	Software Defined Network
SBA	Service Based Architectures
GNN	Graph Neural Networks
CNN	Convolutional Neural Networks
LTE	Long Term Evolution
CLD	cross-layer Design
MAB	Multi Arm Bandits
NFV	Network Function Virtualization
EDGE	Enhanced Data for GSM Evolution
MIMO	Multiple Input Multiple Output
CSI	Channel State Information
UE	User Equipment
FBMC	Filter Bank Multi Carrier
GFDM	Generalized Frequency Division Multiplexing
NOMA	Non Orthogonal Multiple Access
mMTC	Massive Machine Type Communications
TDD	Time Division Duplexing
OFDM	Orthogonal Frequency Division Duplexing
SBA	Service Based Architectures
SVR	Support Vector Regression
TTI	Transmission Time Interval
eMBB	Enhanced Mobile Broadband
MAC	Medium Access Control
URLLC	Ultra-Reliable Low Latency Communication
QoS	Quality of Service
QoE	Quality of Experience
SON	Self-Organizing Network
EDCA	Enhanced Distributed Coordinated Access
5GPPP	5G Infrastructure Public Private Partnership
VOIP	Voice over IP
CW	Contention Window
AIFSN	Arbitrary Inter frame Space Number
TXOP	Transmission Opportunity
MU-MIMO	Multi-user MIMO
ITU	International Telecommunication Union
FG-ML5G	Focus Group Machine Learning for 5G Networks
VAE	Variational Auto Encoders
MPTCP	Multi-path Transmission Control Protocol
GRUs	Gated Recurrent Units
ACA	Automatic Channel Assignment
MSE	Mean Square Error
WMMSE	Weighted Minimum MSE
GCN	Graph Convolutional Network
ARIMA	Auto Regressive Integrated Moving Average
HARQ	Hybrid Automatic Repeat Request
PSNR	Peak Signal to Noise Ratio
QPS	Quality Prioritised Selection
PRR	Packet Received Ration

issues, there are numerous challenges that arise with these solution approaches. These challenges are:

- *Cross-layer Optimization (CLO):* cross-layer optimization involves large parameters to be jointly optimized for improved QoS performances. However, targeting more parameters simultaneously makes the problem non-convex leading to high mathematical complexity [22]. This increased complexity renders conventional optimization algorithms insufficient to find optimal solutions in realistic timescales (tens to hundreds of milliseconds). Moreover, optimizations need to be carried out keeping in view the user's and application's QoS requirements.
- *Cross-layer Design (CLD):* Information sharing among OSI layers requires additional overhead that may lead to non-optimal network performance thus requiring the determination of optimal time instances for such information sharing. Moreover, what information needs to be shared for a given QoS requirement should be determined under dynamic channel conditions, traffic variations and network load.
- *Multi-RAT Network Management:* Managing multiple access technologies requires continuous monitoring and feedback of network dynamics, user QoS requirements and channel conditions for optimal and QoS aware decision making. Moreover, routing dynamics, interference among RATs and shared medium access would make this decision making even more complex compared to single RAT networks. Cross-layer design and optimization involving lower layers of different RATs would also be different for each access technology.
- *Distributed Network Management:* Distributed network management for QoS provisioning requires edge devices to learn IoT users/applications QoS requirements through traffic flow classification. This traffic classification would help determine precise QoS requirements of IoT users/applications which can then be used for cross-layer optimization at the edge. Edge devices can also learn to predict IoT users/applications QoS requirements to develop proactive control in the network however, it requires understanding of traffic patterns as well as inter-application dependencies.

To address these challenges, deep learning algorithms, especially, Graph Neural Networks (GNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Deep Reinforcement Learning (DRL) possess strong properties to solve such high dimensional and non-convex problems [23], [24], [25]. Many AI & ML algorithms (LSTMs, VAE, GRUs) can help determine the optimal instances for cross-layer interactions, perform multi-RAT decisions (GNNs, CNN, DRL), classify traffic flows (SVM, Decision Trees, KNN, PCA etc.) and predict future IoT traffic (LSTM, SVR) to take proactive, QoS aware and optimal decisions. On top of it, they have the ability to adapt to changing network dynamics, traffic loads, channel conditions and QoS requirements [26]. GNNs possess strong generalization capabilities that can be exploited to extend ML based optimization over varying ranges of network topologies [27]. They can also capture OSI inter-layer parameter dependencies which possess a graphical structure between them [28] to enable a stable cross-layer design.

III. RELATED SURVEYS AND CONTRIBUTIONS

The awareness of importance of AI & ML and their suitability to address complex challenges in wireless networks already exists in literature. Due to their inherent strengths and strong adaptive abilities, network researchers have employed different variants of AI & ML algorithms in their work to optimize network performance [29], [30], [31], [32], [33]. Different network problems like channel access [34], link configurations [35], frame aggregation [36], traffic and channel predictions [37], adaptive beamforming [38] etc. have been addressed through AI & ML algorithms. From the network management perspective, user mobility prediction, handovers management [39], user associations [40] and network deployment problems have also been tackled efficiently through AI & ML. However, research efforts to employ AI & ML to understand complex relationships and dependencies between OSI layer parameters, cross-layer optimization and developing distributed intelligence in network edge devices to improve network scalability and IoT user's/application's QoS/QoE is somewhat unexplored.

The Networking research community has reviewed and surveyed many works employing AI & ML in wireless networks research. There are various surveys done on application of AI & ML in wireless networks [41], [42], [43] and WiFi networks [36], [44]. These surveys are focusing on AI & ML employment in resource allocation, user association, mobility management, network security and anomaly detection etc. and they did not discuss AI & ML role in QoS provisioning in wireless IoT networks. Some researchers have surveyed works on cross-layer design and optimization in wireless networks [7], [8], [9], [15] however, literature lacks a review of AI & ML applications and associated challenges in CLD/CLO. Similarly, many works have surveyed distributed network management [45], [46], [47] for improving traditional network parameters and did not address QoS provisioning at the edge through distributed learning. Different from the other works, we have focused on AI & ML applications in cross-layer design and optimization in multi-RAT IoT networks and distributed network management for the end objective of meeting IoT users/applications QoS and QoE. Our contributions to the body of research are as follows:

- We have briefly surveyed traditional cross-layer design and optimization approaches to highlight their limitations and surveyed recent advances in AI & ML employment in CLD. We have highlighted key issues and challenges in AI & ML employment for CLD/CLO and presented ways to address these challenges.
- We have presented a potentially novel approach to study cross-layer dependencies using GNNs and how they can be employed in cross-layer design.
- We have highlighted key challenges in employment of multi-RAT for network capacity enhancement and meeting QoS requirements in dense IoT networks and highlighted advances in multi-RAT network management along with new challenges/issues in them from cross-layer design perspective.
- We have presented a complete systematic approach to employ cross-layer design at the edge devices in a distributed framework. We have highlighted key elements required in the edge devices to enable reliable QoS for end users/applications in dense IoT network.
- We have presented research efforts to standardize AI & ML architecture for wireless networks and pointed out open key research challenges and future directions in the context of enabling reliable QoS in multi-RAT IoT networks using AI & ML algorithms.

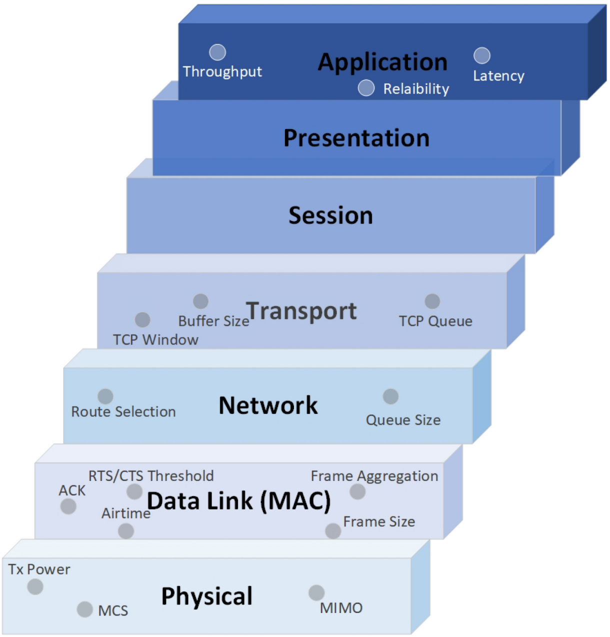

FIGURE 3. Performance affecting Parameters in OSI Layers.

We hope that this article would help researchers working in the field of wireless networks to understand the requirements and design flow of QoS provisioning using CLD and distributed AI & ML approach.

IV. CROSS-LAYER DESIGN AND OPTIMIZATION

Wireless IoT networks operate in layered structures with each layer performing its specified functions. All networks follow the Open System Interconnection (OSI) model for communication between devices. The OSI model is a seven layered structure with application, presentation, session, transport, network, data link and physical layers as shown in Figure 3. These layers make a modular structure which offers benefits in terms of standardisation and easy implementation, however, they fail to exploit inter-layer interactions to improve user QoS [15]. It has been studied in literature that interactions between OSI layers and parameters can exploit hidden relationships between them and can enhance network performance in terms of security, QoS, Mobility and Link Adaptation [9]. There are several parameters in different layers of the OSI model that directly influence the overall performance of a wireless network. These parameters include TCP window at transport layer for TCP traffic or segment size in case of UDP traffic, routing and queue management in network layer, airtime, channel assignment, scheduling, frame length and aggregation, contention window, transmit opportunity (TXOP), ACK (for TCP traffic) and RTS/CTS threshold on MAC layer and transmit power, MCS selection and MIMO configuration on PHY layer of the OSI model. A pictorial representation of the OSI model with performance affecting parameters at different layers is shown in Figure 3. These parameters have complex relationships and inter dependencies among them and proper configuration/reconfiguration of these parameters have great potential to improve QoS and even increase network capacity by reducing airtime wastage in wireless networks.

Additionally, the sharing of information between adjacent layers and across non-adjacent layers can provide valuable information for quick and efficient decisions towards reliable QoS. Informing the transport layer about non-congestion related failure in case of packet loss is a well-known example of information sharing between non-adjacent layers in the TCP/IP protocol [48]. Since, there exist diverse QoS requirements in IoT networks for smart cities, smart industry and smart healthcare, they can be met in number of ways through a cross-layer design approach and cross-layer optimization. The layer parameter selections and optimization can be done differently for different QoS requirements and can thus provide the desired QoS. For example, a reliable packet delivery under network congestion in an industrial use case may require momentary increase in buffer size rather than dropping packets out of queue to increase packet delivery probability. The same requirement can also be met by increasing transmission power at the physical layer under such congestion to ensure usage of higher MCS and faster transmission rates which in turn clears the buffer and reduces packet drop probability. However, such decision making requires understanding of application QoS needs, sharing of information among layers and cross-layer parameter optimization. To understand application QoS requirements, edge nodes can employ traffic classification algorithms to segregate traffic into different QoS categories which can then be used for cross-layer design and optimizations. This cross-layer approach can be employed in base stations/access points as well as in end devices to improve performance in both downlink and uplink of the network.

A. RESEARCH EFFORTS IN CROSS-LAYER DESIGN AND OPTIMIZATION

CLD has been studied extensively by the telecommunication research community in the past [9], [15]. Application-specific throughput and latency requirements were met through cross-layer optimization for video streaming application in [49]. Application, data-link and physical layers are jointly optimized using application oriented objective function to maximise user satisfaction. A cross-layer optimizer is used to find the optimal parameters for three layers (application, data link and physical) using parameter abstractions, and optimal parameters selection at these layers is done to meet application requirements. Video source rate, time slot and modulation schemes are used as parameters and are jointly optimized using Peak Signal to Noise (PSNR) in the optimization function [49]. This overhead can pay off with improved performance in a number of scenarios in IoT networks. A similar work is done in [50] in which MPEG-4 video transmission is optimized by CLD involving the data-link and physical layers. Graph Signal Processing has been used in [51] to optimise energy in WSN by considering application requirements and physical layer connectivity. The authors have shown that CLD can significantly improve performance, however, it comes at a cost of communication and computation overhead.

A well known example of CLD is the TCP window optimization using information from the link layer. TCP employs a congestion control mechanism whereby it readjusts its transmission rates and window size based on transmission errors. Whenever an error occurs, the TCP sender reduces its transmission rate considering network congestion has occurred, however, the transmission error can be caused by several other factors related to the uncertainty of wireless medium. The CLD here employs explicit notification of transmission error through Explicit Loss Notification (ELN) to inform TCP sender about the loss other than network congestion and therefore, avoids reducing the transmission rate [48]. A snoop agent keeps track of the acknowledgements (ACK) and sets an ELN bit in case of missing acknowledgement. This ELN bit is either included in the TCP header or communicated using Internet Control Message Protocol (ICMP) messages. Similar to TCP window optimization, network throughput is significantly increased through Automatic Modulation and Coding (AMC) at the PHY layer and Hybrid Automatic Repeat Request (HARQ) employment at the Link Layer (LL) [52]. HARQ addresses the fading problems and enables the MAC to select the best modulation schemes and also arranges re-transmission of lost packets upon information from receiver. This improves spectral efficiency and reduces the latency of the system as ACK packets do not have to go through the entire stack back to the transport layer. In mobility scenarios, the handovers causes significant delays in user traffic as users shift from one base station / access point to another. These handovers are handled by the network layer (L3) which has topology information and by the link layer (L2) which is controlling link management. The intercommunication between L2 and L3 can significantly improve network performance by initiating handover procedure on L3 prior to its completion on L2 [53], [54].

B. ISSUES AND CHALLENGES IN CROSS-LAYER DESIGN AND OPTIMIZATION

Despite the advantages of CLD, it has not seen much support by the research community. It is believed that layered and modular structure offers longevity, proliferation and parallel development of multiple technologies at OSI layers [55]. Moreover, it defines standard interfaces and intercommunication protocols that can be followed during development in parallel with the assurance that system will keep running. On the other hand, CLD involves feedback loops across the layers and which might cause instability in controlling the network [55]. CLD then requires extreme care. Since there are several performance affecting parameters and they are interrelated to each other in numbers of ways, the optimization of such parameters would be done in multiple loops at different timescales and would cause instability in the system [56]. Moreover, one interaction between two layers can initiate multiple unintended interactions with other layers that can cause catastrophic effects on network performance. This necessitates construction of dependency graphs and robust cross-layer designing that can ensure performance improvement while providing system stability.

In order to avoid the unintended consequences of cross-layer design, research must be focused on studying the inter-layer dependencies. As layer parameters have effects on the throughput, latency and packet loss for different QoS requirements, it is possible some parameters have the same effect on end KPIs. A useful tool to identify such parameters is to calculate the mutual information between layer parameters. If we have two parameters X and Y with probability distributions of their range of values as P(X) and P(Y), then the mutual information between them can be calculated as follows:

$$I = \sum_{y \in Y} \sum_{x \in X} P_{X,Y}(x, y) log\big(P_{X,Y}(x, y)/P_X(x)P_Y(y)\big) \quad (1)$$

Mutual Information (MI) employs concepts of probability and information theory (entropy). MI-based analysis would identify level of information that co-exist between multiple layer parameters. As such, their effects on network performance can be studied keeping in view the mutual information they carry together. This analysis can also help in identifying parameters that can cause potential instability in the system. These parameters can then be handled by carefully designing feedback and optimization loops in algorithms targeting cross-layer design [57]. Moreover, similar parameters carrying identical information or producing similar effect on network performance can be grouped together for dimensionality reduction [58] and easing the optimization complexity. This would significantly improve the optimization time and can target large range of time-constraint decisions in wireless IoT networks to satisfy latency sensitive services. Considering healthcare use cases from Figure 2, robotic surgery and emergency care use cases can be served with much lower latency in a healthcare IoT network with such optimizations.

C. THE ROLES OF AI & ML IN CROSS-LAYER DESIGN AND OPTIMIZATION

Interest in cross-layer design has decreased in the past due to the high complexity, however, with the advancement in technology, researchers can now exploit strengths of AI & ML algorithms for better and efficient cross-layer designs as well as cross-layer optimization. Since QoS requirements and traffic patterns keep on changing in an IoT network, AI & ML algorithms can learn to map QoS requirements to the optimal cross-layer parameters for meeting their changing requirements. The benefits of such approach are two-fold 1.) it can help improve the network level throughput, latency and packet loss and 2.) it can provide improved QoS in the network for large number of traffic flows. CNNs have seen great progress in the field of computer vision and completely revolutionized it as they have the ability to extract complex features from underlying data for a given objective. Similarly, DRL, LSTM and GNN algorithms have seen an upward trend to solve complex optimization problems in wireless networks [29], [41], [44], [59], [60]. A brief review of AI & ML algorithms employment in CLD/CLO in wireless networks towards meeting QoS requirements is given in Table 2. DRL, LSTM, GNN and CNNs possess great potential to study cross-layer interactions and handle complex cross-layer optimization which remained a bottleneck in the past CLD/CLO studies.

The strengths of AI & ML have been exploited in the literature for cross-layer design to some extent. A work in [61] employed cross-layer design to target network and MAC layers of an IEEE 802.15.4 based network and achieved a higher packet reception ratio and energy efficiency with less overhead. The authors in [62] used DRL for power efficiency and route latency reduction for better energy efficiency and reduced latency. A similar work [63] used DQN in a Cognitive Radio Network (CRN) to optimize physical and network layers parameters (SINR and Routing) for improved QoE in an interactive video use case. Multiple research works [63], [64] have employed AI & ML algorithms for cross-layer optimization, however, they have targeted few parameters (two or three) from only two layers at a time for joint optimization. However, advanced machine learning algorithms (CNN, GNN, LSTM, DRL) have the strength to optimize much more complex problems. Moreover, mutual information analysis combined with AI & ML algorithms for CLD can work together really well with mutual information performing the dimensionality reduction and AI algorithms performing the joint optimization of reduced parameter space. It is to be noted here that with each parameter addition in ML problem, its complexity increases but at the same time it can deliver more optimized results. Therefore, research efforts in this domain can prove very fruitful. Due to the multiple inter-dependencies of cross-layer parameters, systems can face potential instability with an ML approach therefore, caution is required while devising such solutions [55].

In the real world, human beings gather information from multiple sources such as eyes, nose, ears, touch etc. The human mind processes information from these sources jointly to take decisions. The information from each source can be encoded and represented as a single modality and it can also be combined with information from multiple sources to make it multi-modal information. Each information source has its own statistical properties and it is important to determine the

TABLE 2. Machine learning applications for CLD/CLO in wireless networks.

Classification	AI/ML Algorithm Used	Problems Addressed	Results	Year	Ref
cross-layer optimization	Distributed Actor-Critic DRL	QoE Enhancement for 360 degree Video	High Reward in 50K iterations over QPS	2020	[64]
	Graph Neural Networks	Multipath TCP optimization	14.87% TH Improvement over mesh protocol	2020	[28]
	Prioritized Memory DRL	Power Efficiency and Route Latency Reduction	25% & 28.5% improvement in EE and Lat over DQN	2019	[62]
	Deep Q-Network	QoE through SINR and Routing	Video QoE improvement over PHY layer optimization	2019	[63]
	Deep Q-Learning	Packet Reception Ratio and Energy optimization	11% EE and 20% improvement in PRR	2020	[61]
Parameter optimization	Deep Reinforcement Learning	Energy and Throughput optimization	14.1% EE and 2.9% TH improvement over CPLEX	2021	[25]
	Decision Trees	Throughput Improvement by optimising AIFSN	20% Video/Voice QoS improvement over IEEE 802.11e	2015	[30]
	Random Forest & M5P Regression	Frame Length optimization for Throughput	55% Network TH improvement over IEEE 802.11	2020	[36]
	Deep Neural Network	TCP Window optimization	9.02% TH Improvement	2019	[65]
	Graph Neural Networks	Throughput optimization for Multimedia Traffic	Prediction Accuracy 90% 29% Latency Reduction	2019	[66]
	Convolutional Neural Networks	Power control for Sum-rate maximization	95.13% Sum-rate & 2.45% Time Improvement over MLP	2020	[23]
	Graph Neural Networks	Power Control	Sum-rate improvement over WMMSE with varying topologies	2017	[67]
Multi-modal Learning	Bipartite Graph Learning	Channel Usage Improvement in Wireless Network	20.2% TH Improved & 23.9% reduction compared to ACA	2020	[68]
	Multi-modal Split Learning	mMWave Received Power Prediction	16x Lower Comm Latency & 2.8% Less Privacy Leakage	2020	[69]
	Deep Multi-modal Learning	MIMO Channel Predictions	Normalized MSE of 0.005 to 0.008 with different Models	2020	[70]

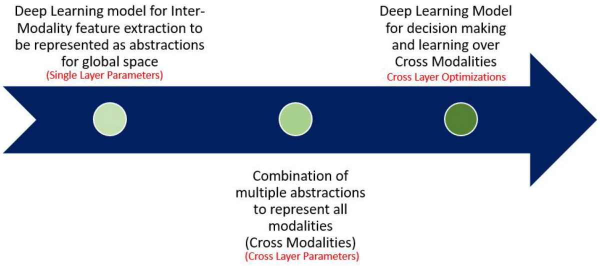
Deep Learning model for Inter-Modality feature extraction to be represented as abstractions for global space
(Single Layer Parameters)
Deep Learning Model for decision making and learning over Cross Modalities
Cross Layer Optimizations
Combination of multiple abstractions to represent all modalities (Cross Modalities)
(Cross Layer Parameters)

FIGURE 4. Multi-Modal Learning Flow.

relationship between statistical properties of these modalities for better and improved decisions. Analogous to human beings, information for the decision making in wireless networks come from multiple sources. For a central network controller, information comes from multiple switches, access points and gateways in the network to take policy decisions. Similarly, an edge node employing cross-layer design gets information from multiple OSI layers for optimal parameter selections to meet application QoS requirements. This information from multiple layers can be represented as multi-modal information and hence Multi-Modal Learning can be employed.

Multi-Modal Learning [71] aims at studying the relationship between different modalities to improve decision making by using AI algorithms. The inputs from a single layer can be called as intra-layer modalities and can be represented with a single abstraction. Similarly, inputs from other layers can be represented with abstractions which can be combined to extract cross modalities as it is done in multi-modal learning [72]. The working of multi-modal learning and its analogy with cross-layer optimization is shown in Figure 4. Multi-modal learning has proved very beneficial in different multi-modal classification tasks like video classification [73], sentiment analysis [74] and visual question answering [75]. Like other fields, multi-modal learning can be employed in cross-layer design for improving QoS in wireless networks. Different researchers have employed multi-modal learning in wireless networks for example, to optimize channel usage [68], mmWave received power prediction [69] and MIMO channel predictions [70]. Since QoS provisioning in wireless network through cross-layer design involves inputs from multiple layers, they can be represented with cross modalities and multi-modal information fusion and learning can be employed to learn the optimal parameter configurations for wireless network optimizations.

Besides cross-layer optimization, information sharing between OSI layers can significantly improve QoS in the network. Sharing of information between application and lower layers can help tailor parameters in lower layers to deliver information as per application needs. Other than increasing buffer size or transmit power upon information sharing about congestion, low latency traffic can be better handled if the application layer shares its QoS requirements through inclusion of traffic tags in packets. Another way of serving low latency traffic through cross-layer design is to share information about hardware transmission queues to the MAC layer to increase the quantum of high priority traffic thus reducing its delay. However, this additional information leads to overhead and can reduce network performance. Therefore, enabling this communication at important time instances only, determined through AI & ML algorithms, can significantly improve efficient QoS delivery in IoT networks.

Cross-layer design and optimization have great potential to improve QoS in IoT networks but they require understanding

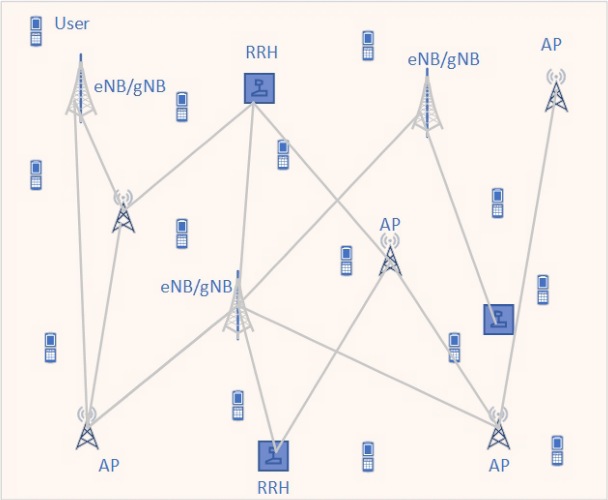

FIGURE 5. Network Nodes as an irregular Graph.

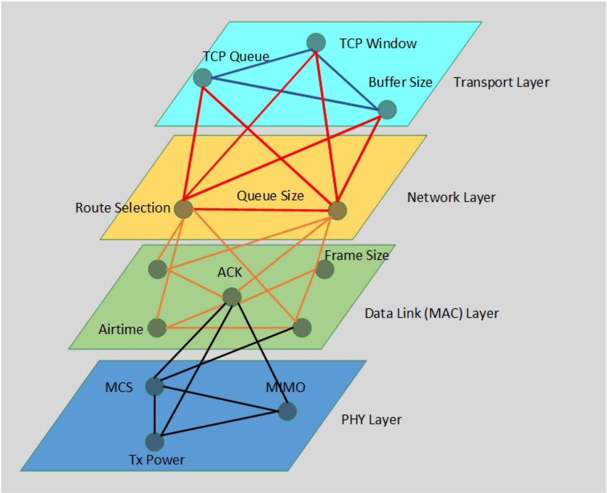

FIGURE 6. Graphical Representation of relationships between parameters at different OSI layers.

of inter layer dependencies and layer parameter dependencies for a stable system. Failure to do so can lead to much worse performance and failure of IoT use cases. However, AI & ML algorithms, especially GNNs possess strong properties to study inter layer dependencies that can generalize well over different QoS requirements. Multi-modal learning, CNNs and DRL can also help in cross-layer optimizations by targeting larger number of layer parameters in optimization problem.

V. GNN FOR INTER-LAYER DEPENDENCIES AND NETWORK PERFORMANCE ENHANCEMENTS

Cross-layer design and optimizations rely on study of inter-layer dependencies and their exploiting for different QoS requirements of applications and devices in an IoT network. Due to dependencies and relationships among layer parameters, they are better studied in a relational way and Graph Neural Networks (GNNs) are a promising tool to undertake such a study. The layer parameters and their relationships can be represented as graphs and nodes to employ GNNs, with the QoS requirements included in the overall optimization framework. Similarly, wireless networks include base stations (eNB, gNB), relays, APs and Remote Radio Heads (RRH) that are distributed by network operators in a planned and organised way with the end objective of serving more users and enhance coverage. The users in the wireless network are mobile and are randomly distributed under network coverage. This creates a non-structured network with irregularly positioned nodes as shown in Figure 5. Therefore, GNNs become an optimal choice to optimize overall network performance as well. GNNs possess strong relational inductive bias [27] that enables them to generalize solutions over changing network topologies as well as over changing QoS requirements.

A. GNNS EMPLOYMENT FOR CROSS-LAYER OPTIMIZATION

As we have seen in the previous section, network performance can be significantly increased through cross-layer optimization however, it requires the study of interdependencies between performance affecting parameters at different layers of the OSI model to create a stable design. Information theory is one way to study those dependencies and relationships, however, GNNs posses strong properties to learn inter-parameter and inter-layer relationships. Since, layer parameters have different dynamics and affects network performance (latency, throughput and packet loss) in different ways, they posses non-structured data and resultantly GNNs are more suitable for studying inter-layer dependencies as they can efficiently handle non-structured data. CNNs and LSTMs, on the other hand, require structured data like images represented with matrices or word embeddings in case of language data for LSTMs. The layer parameters from OSI layers are inter-related to each other in a number of ways that it creates a complex graphical structure. It is possible to represent these relationships with node and edges as shown in Figure 6. The relationships exist between parameters on the same layer as well as across the layer parameters. Each parameter from an OSI layer affects QoS metrics in a different way than the other, therefore it would make an extremely heterogeneous graphical structure and, thus, present huge challenge to study these relationships. Moreover, as QoS requirements change, the relationship between cross-layer parameters also change and they need to exploited in a different way to meet changing QoS requirements. GNNs have some inherent properties to address theses complexities and can be employed to study inter-layer dependencies. The changing QoS requirements in networks would also change the graphical structures between layer parameters, however, GNNs can generalize solutions well over large range of graphical structures [27]. Since GNNs have been employed in heterogeneous networks for example, for multi path routing [28], they possess the capability to capture dependencies in inter-layer parameters as well. Although very complex, the creation of dependency graphs through GNNs would aid in stable cross-layer design and can avoid unintended consequences as discussed in [55].

B. GNNS USAGE IN NETWORK PERFORMANCE ENHANCEMENT

Historically, graph theory has been greatly employed in wireless network problems like channel assignment and interference mapping [76], [77]. Weighted graph colouring has been used in [76] for channel reuse in a WLAN network. Similarly, [77] used graph colouring for channel allocation in a cellular network. Augmenting graph theory, GNNs have opened new doors to address network problems with increased complexities. Many research works have employed GNNs in their work for channel and power allocation, link prediction, route optimization, intrusion detection and traffic prediction problems. For example, RouteNet [78] is a novel solution based on GNN that takes into account the network topology, routing and input traffic to reduce delay, jitter and loss. A custom built packet level simulator with queues from OMNET++ is used for training and testing of models. 260,000 training samples were used with 100 different routing configurations and wide variety of traffic. The system is able to achieve on average 30% improvement in terms of delay and jitter compared to Shortest Path (SP) routing policy and utilisation based optimizer.

Taking RouteNet [78] as reference, authors in [79] proposed PLNet for an IP transport network that is able to achieve same performance as RouteNet but with better inference speed. For improving channel allocation in WLAN, Graph Convolutional Networks (GCN) are employed in [80] to capture the relationships between APs and DRL and are later employed to allocate channels to the APs in shorter time while maximizing reward. Multi-Path TCP, also known as MPTCP, in heterogeneous path network faces sub optimal performance therefore, GNNs are employed in [28] to perform throughput prediction over multiple paths and leverage learnt information to optimize multi path routing. The proposed approach is tested in a Software Defined Network (SDN) and is able to achieve lower Mean Square Error (MSE) in unseen environments (not seen during training) with higher network throughput compared to traditional multi-path routing algorithms. GNNs are employed in [67] for power control in Ad-hoc wireless networks where the weights of Weighted Minimum Mean Square Error (WMMSE) algorithm are parameterized using GNNs for power allocation decisions. The proposed technique is highly generalizable and is tested on variable network sizes and densities for validation. Virtual Network Functions (VNF) placement based on its past usage is done in [66] using GNNs to improve VOIP and other IP multimedia subsystems. Due to the generalisation capability and learning ability of GNNs for topological dependencies, GNNs have recently seen massive research in wireless networks [81]. Looking at the literature of GNN employment in wireless networks, it can be clearly seen that they have been used to optimize network holistically and targeted relational network parameters that impact neighboring users like transmit power, channel assignments and routing. This strengthens the fact that their employment to study inter-layer dependencies in a relational way can prove very beneficial for reliable QoS in wireless IoT networks.

Besides channel assignment, power control and routing optimization, GNNs can also be employed at network level to understand QoS requirements of different IoT devices in a relational way. They can predict link establishments between IoT nodes to proactively handle QoS requirements [82], [83]. Similarly, they can predict network level queue dynamics, routing bottlenecks and application similarities to handle network resources efficiently and improve QoS provisioning. GNNs have also been employed in combination with other machine learning algorithms to improve network wide performance [84]. Specifically, combination of GNNs and DRL has been seen in various networks problems [80], [84]. Although, GNNs have been used in combination with DRL, their combination with other learning algorithms like LSTM, GRU and VAE can be further explored for QoS improvement in wireless IoT networks.

Due to non-structured and continually changing topologies of wireless networks, GNNs have seen a lot of success in developing generalizable solutions for different problems in wireless network. As they can study systems in a relational way, they can help understand inter-layer dependencies for a stable cross-layer design and optimization. GNNs, in combination with DRL and LSTM etc, can optimize large parameter space while looking at the system from a global perspective thus enabling improved network management and control. However, more research efforts are required in this domain.

VI. CAPACITY AND QOS IMPROVEMENT THROUGH MULTI-RAT ACCESS NETWORKS

Network management and control play an important role in optimising network resources and improving overall network capacity. Different conventional and AI based algorithms have been proposed for efficient network management and control of network parameters to enhance network capacity. For example, Wi-Balance [85] provides better channel-aware user association to increase throughput. Similarly, [86] used Graph theory for user association and interference management in WLAN network for capacity enhancement. AI & ML algorithms (DRL, SVM, Decision trees) have also been employed to optimize Contention Window (CW) [87] and AIFSN [30] for improving QoS. Similarly, authors in [88] proposed a distributed‘Reinforcement Learning (RL) user specific cell association scheme with back haul capacity constraints to improve QoS. For efficient spectrum usage in wireless network, a DRL based spectrum allocation algorithm is proposed in [89].

Although AI & ML algorithms can achieve promising results in wireless network management and parameters optimization, they are still limited by the maximum bit

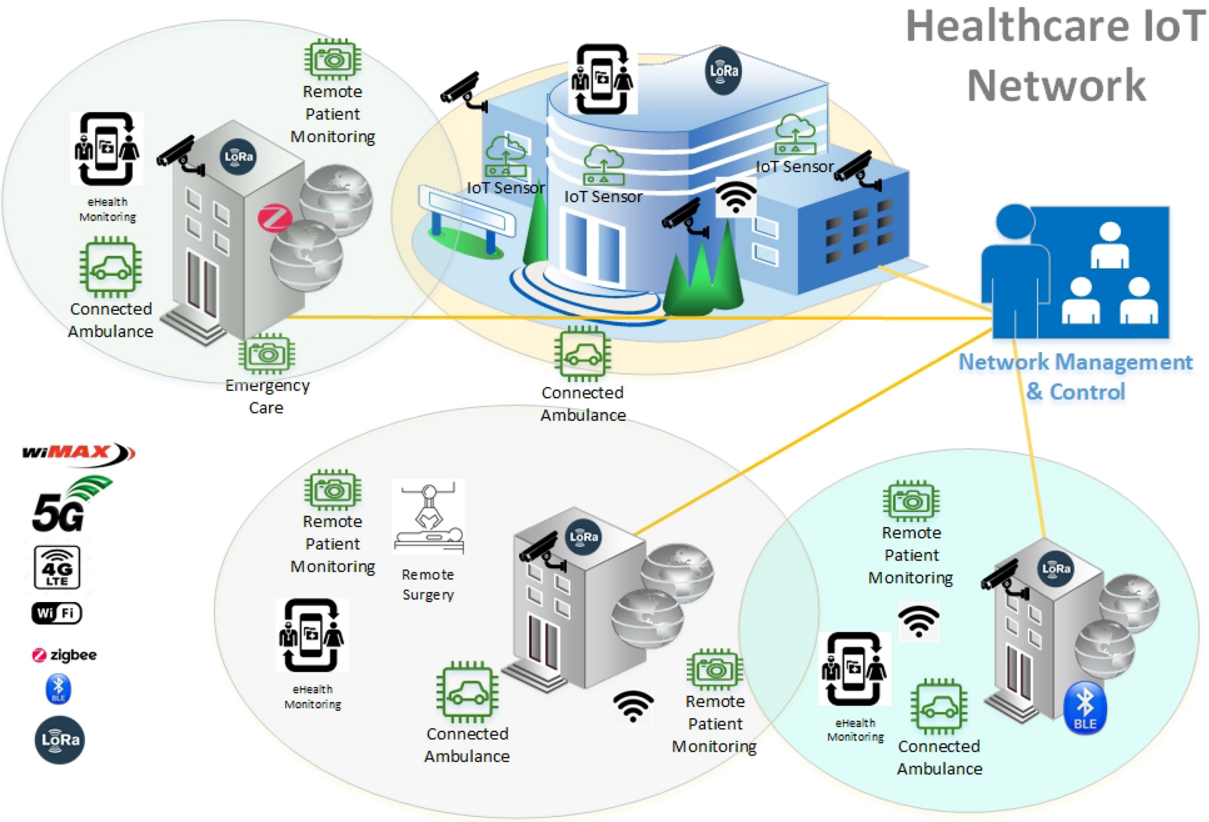

FIGURE 7. Multi-RAT Healthcare Wireless IoT Network.

rate possible at the physical layer technology. After passing through upper layers (Transport, Network, MAC), traffic packets due for transmission reach the PHY layer where they are put in the hardware transmission queue before getting into the air. This hardware queue is cleared at a rate that depends on the PHY layer technology. The access technology always supports transmission of a certain number of bits per second depending on the physical layer technology and channel conditions. For example, WiFi-6 (IEEE 802.11ax) can achieved a maximum bit rate of 9.6 Gbps with MIMO and maximum carrier bandwidth (160 MHz). Similarly, Bluetooth (BLE 5.0) has a theoretical maximum bit rate of 3 Mbps. However, having multiple radio access technologies at PHY/MAC layers can enable clearance of transmission queues at a much faster pace compared to single access technology, thus increasing network capacity. In addition to it, having multi-RAT APs in IoT networks might have additional benefits as IoT networks can serve a diverse range of devices and sensors that employ different access technologies ranging from Bluetooth, ZigBee, LoRa to WiFi and cellular NB-IoT as shown in an healthcare IoT network in Figure 7. Therefore, a single AP can provide the desired wireless connectivity to a wide range of IoT sensors and devices. Moreover, many IoT devices are battery operated and require very high energy efficient communications and a multi-RAT network can offer energy efficient communications to such devices while providing high data rate communication to bandwidth hungry devices using different access technologies.

A. ADVANTAGES OF MULTI-RAT IOT NETWORK

Multi-RAT IoT networks offer a wide range of advantages to provide reliable QoS. They can improve network capacity through use of multiple access technologies at the PHY layer to offload traffic [90]. For example, 5G networks employ WLAN technology to increase their capacity and support high user density [91]. They can also reduce the latency in networks by offloading wireless traffic over different access technologies. Authors in [92] have studied latency reduction in multi-RAT IoT network with RAT selection employed at edge devices and observed significant improvements in latency sensitive applications (see Figure 8). Multi-RAT also offer benefits in terms of energy efficiency and data rate. Authors in [93] report an increase in energy efficiency from 11% to 42% with Multi-RAT network while data rates increase up to 39% compared to single RAT network. A similar work in [94] reported both energy efficiency and latency gains in multi-RAT IoT network with LoRa and NB-IoT access technologies with dynamic QoS requirements and variable payload sizes. Another potential gain of multi-RAT network is availability of redundant connections in case of a single access technology failure. Under such scenario, another technology can hop in to keep connectivity alive in the network [94]. In this context, authors in [95] have

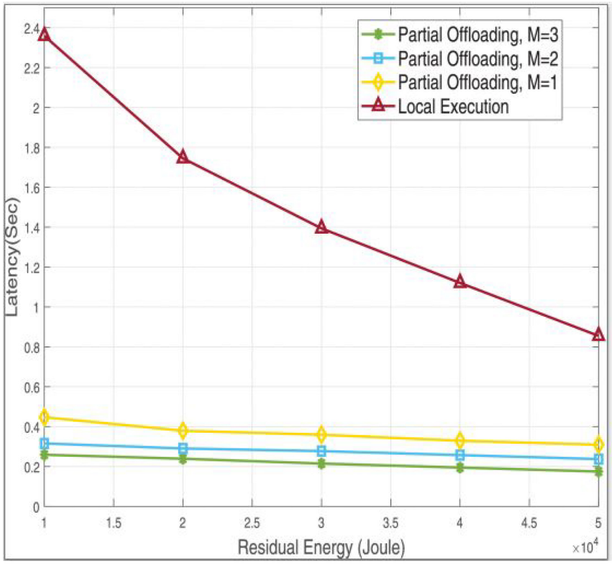

FIGURE 8. Latency Improvement with Multi-RAT Traffic Offloading [92].

improved reliability through Multiple Connectivity (MC) and sending data through multiple RATs. They have employed distributed RL to learn policies for each device for efficient MAC configuration and RAT selection.

Although multi-RAT networks have several benefits, they also face complex management and control challenges that must be handled efficiently. These challenges include dynamic RAT selection with traffic aware decisions, handovers of users over different RAT, interference management among unlicensed bands (WLAN, BLE, ZigBee), energy management of low-power and energy harvesting nodes, and multi-path routing with multiple RATs on the data path. There are various works in literature [93], [96], [97], [98] which have addressed these challenges using conventional as well as AI & ML techniques, however, our focus in this article is on cross-layer design in multi-RAT IoT networks.

B. CROSS-LAYER DESIGN IN MULTI-RAT NETWORKS

Just like with single RAT networks, cross-layer optimization in multiple RATs can provide promising results towards enhancing network capacity and meeting QoS requirements. However, it would require independent efforts for CLD for each access technology in the network. Cellular networks (LTE and 5G) have different layer structure and involve PHY, MAC, RLC (Radio Link Control) and PDCP (Packet Data Convergence Protocol) layer. Moreover, the operations of the PHY and MAC layers are also different from the same OSI layers in IEEE 802.11 networks. Similarly, the layered structure and operation are also different for IEEE 802.15 networks. These differences require independent cross-layer solutions for each access technology and would requires inter-layer interactions after the RAT selection decision. SDN architectures would effectively aid in CLD for multi-RAT networks as it separates the control plane from the underlying data plane. Applications requirements from devices in the network, wireless environment and channel statistics from the data plane can be communicated to the control plane to take RAT selection decisions. Afterwards, cross-layer optimization can be employed on the data plane for each access technology independently to maximise network performance.

C. ROLE OF SDN IN MULTI-RAT NETWORKS

Software Defined Networking (SDN) technology is a promising enabler for the efficient management of multi-RAT networks. It can provide benefits in terms of seamless handovers, unified authenticity and security and increased flexibility [99]. SDN facilitates the creation of network slices from a centralised controller thus creating multiple logical networks over same infrastructure to support different QoS requirements. These slices can belong to different RATs available on PHY layer of the multi-RAT network and can be managed through a single SDN controller. AI & ML can effectively aid in creation, management and operation of these network slices based on user QoS requirements while controlling each slice as per underlying access technology constraints. This would also help in employing independent cross-layer designs for each RAT while involving global network knowledge like routing paths, adjacent channel usage, transmit power of neighboring devices etc. into the decision matrix.

SDN architectures and controllers that support multiple access technologies are very limited, therefore, research efforts towards development of practical multi-RAT SDN controllers are required. Though 5GEmpower [100] support LTE and WLAN access technologies, support for other access technologies is yet to be integrated. Like 5GEmpower, the SDN controller proposed in [99] employs RAT Abstraction Functions (RAF) to control different RAT technologies. The SDN controller resides in the Core Network (CN) and takes RAT selection decisions without worrying about lower layer aspects of access technology, however, this controller has not been tested in a physical testbed and is not IoT focused. Multi-RAT SDN architectures have been proposed for cellular networks [101] and WLANs [102] however, incorporation of other low-power or long-range technologies for power constrained IoT devices is still needed. SDN-WISE [101] is a good effort towards developing a SDN controller for 802.15.4 based WSN to reduce signalling overhead and includes programmability, however, no effort has been seen in literature that targeted development of SDN controller for IEEE 802.11 and IEEE 802.15 based IoT networks simultaneously. Several other SDN architectures for WSN and IoT networks are discussed in [103], however, they also lack multi-RAT support.

D. ROLE OF AI & ML IN CROSS-LAYER DESIGN IN MULTI-RAT NETWORKS

Cross-layer design in Multi-RAT network is more complex and challenging task as access technologies differ in their operations specifically in MAC and PHY layer operations. Although many works have targeted CLD in different access technologies like IEEE 802.15 [104], [105], IEEE 802.11 [106], LoRa [107] and cellular networks [108] and also employed AI & ML [61], [62], [63], there is no universal cross-layer design addressing all technologies. Some researchers have employed Multi-agent DRL (MARL) for intelligent RAT selection and resource allocation at the edge however, they did not consider cross-layer information from multiple layers for maximizing user QoS satisfactions [109], [110]. Due to significant differences in the MAC and PHY layers operations of RAT, getting cross-layer information and employing AI algorithms become increasingly complex. However, SDN architectures provide a framework where different network applications can be deployed in SDN controller that can gather information from different layers of access technologies. This information can then be fused to employ AI & ML algorithms for CLD (see Section IV) and CLO for different access technologies together. Such a framework resembles manager based CLD discussed in [15] but offers flexible control over multi-RAT networks. Authors in [111], [112] have proposed SDN based cross-layer approaches to improve data plane functionality, however, AI & ML employment to perform CLD in multi-RAT network is yet to be explored. One drawback of the SDN approach is the centralized management of network which introduces significant delays and can fail many low latency IoT use-cases (see Section VII). This can also lead to single point of failure and jeopardize the whole network in case of controller failure. However, different distributed SDN controllers have also been proposed in literature to overcome this drawback [113].

Multi-RAT IoT networks employing CLD can provide reliable QoS to the wide range of IoT users/applications by providing multiple access technologies to offload their traffic under congested and dynamic environments, however, they present complex challenges that needs to be addressed. Despite additional challenges to perform cross-layer design in multi-RAT IoT networks, AI & ML algorithms in SDN controlled network can efficiently overcome these challenges. Since IoT networks are expanding everyday, more scalable solutions are required to handle large network densities. AI & ML algorithms can also help in this context to develop intelligence in edge devices thus enabling distributed network management. This would also help improve performance of low latency use cases in many IoT networks. Nevertheless, research efforts focusing on edge intelligence for network decisions and distributed SDN architectures are required.

VII. DISTRIBUTED NETWORK MANAGEMENT AND EDGE INTELLIGENCE

IoT is improving our lives in a number of ways. More and more smart devices and sensors are being manufactured that are being connected to Internet and improving our quality of lives. In smart industries, various sensors, cameras and robots are deployed to increase factory productions [20]. In smart healthcare, various kinds of patients data is being collected to monitor health of patients [21]. This data is continuously analyzed through intelligent algorithms to detect health concern and automatically initiate corrective actions. Inferences and proactive actions are now being taken through predictions and data analysis at the edge rather than in central server or cloud. AI & ML algorithms are playing a vital role in enabling edge intelligence into the end devices [114]. This edge intelligence can also be employed to take network resource management, optimal communication parameters selection and cross-layer design decisions to enable reliable QoS in the network. This would enable distributed network management to efficiently support low latency and highly responsive use cases in wireless IoT networks.

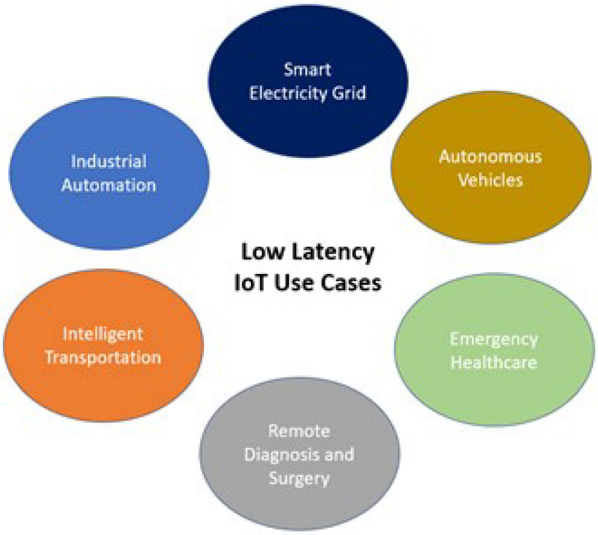

FIGURE 9. Low Latency IoT Use Cases.

A. LOW LATENCY USE-CASES IN IOT NETWORKS

A large number of uses cases in IoT networks require low latency and high reliability communication. To name a few, these use cases include industrial automation, autonomous vehicles, remote diagnosis and surgery, intelligent transportation, emergency healthcare and smart electricity grid as shown in Figure 9. Augmented Reality (AR) assisted robotic surgeries are being performed by expert surgeons while sitting in any part of the world. They require high throughput and ultra low latency communications to transport video and haptic feedbacks. Similarly, vehicle predictive maintenance requires high reliability communications with various sensors data like temperature and vibration to reduce maintenance costs and improve downtime. Low latency and reliable communications are also required to ensure protection of electricity grids and harbour automation. To enable these low latency use cases, the data processing and inferencing has already been shifted to the edge to increase responsiveness and reduce latency [115]. Moreover, AI & ML algorithms on heterogeneous data from IoT devices have enabled edge devices to initiate corrective actions without having to send the data to central server [116].

Low Latency IoT use-cases require information to be delivered reliably with minimal latency thus requiring highly responsive communication and network infrastructure. The true benefits of IoT in smart high tech environment can only be realized if their required communication needs are met in a satisfactory way [117]. Current communication requirements of an IoT network are met through multiple technologies including cellular, WiFi, Bluetooth, ZigBee and LoRa which involve centralized as well as user-centric (distributed) control of wireless resources. They employ standard channel access mechanisms without understanding the application/IoT sensors QoS requirements which induces increased latency to critical traffic flows. New medium access protocols have also been proposed to improve channel access and reduce latency for machine type communication in industrial IoT networks [118], [119]. Moreover, increasing user density in the network causes congestion thus affecting QoS in whole network [120]. Understanding QoS requirements and managing network resources as per application requirements at the edge possess great potential to enable high quality, low latency communication. AI & ML algorithms in the edge can autonomously understand users/applications QoS requirements and can efficiently map them to available network resources for reliable QoS. However, AI & ML at the edge is still in its infancy and researchers are trying to employ multiple learning algorithms for distributed network management [121]. Enabling the edge (Base Stations, Access Points, Devices themselves) to take decisions regarding channel assignments, routing, traffic scheduling, traffic shaping and channel access can not only reduce latency, it would also be beneficial in shedding off computational overload on the central network controller. Despite advantages of distributed network management and control, there are certain decisions that require global network knowledge. For this purpose, cooperative decision making and federated learning architectures, in which edge node share their learnt information with the central node which regulates the decision of edge nodes, have been proposed in the literature [122], [123], [124], [125]. Various works have tried to address decentralised and distributed network resource management, however, most of these works have focused on routing [125], power allocation [123], channel assignment [126] and scheduling [118], [119]. One of the fundamental requirements to meet QoS requirements at the edge is to understand user's/application's QoS requirements. Base stations/Access points are the edge devices handing user traffic and network resources, therefore, they need to employ AI & ML algorithms to classify IoT traffic to understand QoS requirements. Moreover, they would be able to predict traffic using LSTMs and regression techniques to develop proactive control of network resources, thus enabling reliable QoS in the network.

TABLE 3. DSCP tags in wired networks.

DSCP Value	Class	Standard Use-Case
0	BE	Best Effort
8	CS1	Scavenger
16	CS2	Network Control
24	CS3	Broadcasting
32	CS4	Streaming
40	CS5	High Priority
48	CS6	Network Management
56	CS7	Network Management
10	AF11	High Throughput Data
12	AF12	High Throughput Data
14	AF13	High Throughput Data
18	AF21	Low Latency Data
20	AF22	Low Latency Data
22	AF23	Low Latency Data
26	AF31	Multimedia Broadcasting
28	AF32	Multimedia Broadcasting
30	AF33	Multimedia Broadcasting
34	AF41	Multimedia Conferencing
36	AF42	Multimedia Conferencing
38	AF43	Multimedia Conferencing
44	Voice Admit (VA)	Voice Calls
46	EF	Real-Time Interaction

B. TRAFFIC CLASSIFICATION AT THE EDGE

IoT devices have a diverse range of QoS requirements depending on the underlying use cases. In order to take communication and networking decisions locally, intelligence needs to be embedded into the edge to classify IoT traffic into QoS classes according to the desired levels of throughput, latency and reliability [127]. Depending on traffic classes, underlying network resources can be managed to provide the desired QoS. In wired networks, QoS is provided through Differentiated Services (DS) by including QoS information in the form of Differentiated Services Code Point (DSCP) in the IP header. The different DSCP tags used in wired networks are given in Table 3. Depending on DSCP tags, the traffic is scheduled to meet the QoS requirements of different traffic flows present in the network. Wired networks have considerably large bandwidths and channels are reliable, therefore, meeting QoS requirements is easier compared to wireless networks. On the other hand, traffic traversing from wired to wireless networks involves the translation of DSCP to Quality Class Indicators (QCI) in cellular networks and IEEE Access Categories (AC) in 802.11 (WiFi) networks. The scheduling policies at MAC layer then handle traffic to meet QoS requirements of different traffic flows. In principle, traffic classification in wireless networks is done through the DSCP tags in the IP header. However, the IEEE 802.11e standard [128] defines only four ACs, namely voice, video, best effort and background traffic and, therefore, traffic traversing from wired to wireless interface loses QoS information embedded in the traffic flows. The WiFi ACs, their priorities and DSCP translations are given in Table 4. Moreover, most of the network traffic is tagged as Best Effort (BE) by the IoT devices despite having different QoS requirements [129]. This necessitates development of traffic classification algorithms at the edge devices to improve the QoS delivery in wireless networks. The classification cannot be universal

TABLE 4. IEEE 802.11 access categories and DSCP translations.

Priority	User Priority	802.11e Access Category	Description	DSCP to AC Translation
Lowest	1	AC_BK	Background Traffic	CS1
	2	AC_BK	Background Traffic	CS1
	0	AC_BE	Best Effort Traffic	AF11, AF12, AF13, CS2
	3	AC_BE	Best Effort Traffic	AF21, AF22, AF23
	4	AC_VI	Video Traffic	AF31, AF32, AF33, AF41, AF42, AF43, CS4, CS3
	5	AC_VI	Video Traffic	CS5
	6	AC_VO	Voice Traffic	VA, EF
Highest	7	AC_VO	Voice and Management Traffic	CS6, CS7

and depends on the various IoT applications in the network, therefore, understanding traffic classes is essentially required to initiate subsequent resource management for the desired QoS. AI & ML techniques can be employed to determine statistical characteristics of IoT traffic and classify them into various QoS classes [130]. For crude classification, traffic coming of various ports or IP addresses can be tagged at the processing nodes, however, it requires knowledge about VLANs and IP assignment to various kind of IoT devices. Another way is defining standard traffic tags based on QoS requirements just like multimedia traffic [129]. IoT sensors and applications can include these tags in IP packets to enable priority based resource management.

C. TRAFFIC PREDICTION AT THE EDGE

Proactive network management, where base station/access points take proactive decisions based on their previously encountered traffic patterns, decisions taken and their future predicted traffic patterns, can significantly enhance the QoS in the network. Besides traffic classification, real-time traffic prediction can also help in efficient mapping of network resources to user demands [131]. These user demands keep changing at different times of day and the traffic variations also occur due to user mobility which shifts traffic load from one access point to another. Therefore, continuous forecasting of traffic is required. AI & ML techniques such as Long Short Term Memory (LSTM) [132] and regression models (Decision Tree Regression, Gradient Boosted Regression Tree, K-Nearest Neighbour Regression, Support Vector Regression etc.) [133] are promising tools to understand time dependencies and accurately forecast/estimate user requirements. LSTMs were able to achieve the MSE and MAE of 0.05 and 0.3 respectively while regression models achieved MSE and MAE of 0.004 and 0.002 respectively for different traffic predictions tasks indicating their potential to perform this task accurately [132], [133].

D. ROLE OF AI & ML IN DISTRIBUTED NETWORK MANAGEMENT AND EDGE INTELLIGENCE

Distributed learning algorithms are increasingly being researched in control problems, edge inferencing and network decisions [121]. Federated learning and Actor-Critic classes of DRL are being employed to improve latency performance of IoT traffic and multimedia traffic delivery [64], [115]. Similarly, contention window optimization is now done in a distributed manner through federated learning to improve system throughput [31]. Cooperative DRL algorithms are also being employed to develop distributed control in the network for improving energy efficiency and reducing delay of IoT sensors [125]. As discussed earlier, distributed learning for reliable QoS requires first to understand QoS requirements autonomously. This requires traffic classification to be performed at the edge devices and subsequently employing either AI & ML aided CLD/CLO or RAT selection. A brief review of distributed learning, network management, traffic classification and prediction has been given in Table 5.

Traffic classification in wired as well as wireless networks has been researched well over the past two decades. Both machine learning [134], [135] and conventional traffic classification techniques [136], [137] have been studied and promising results have been achieved. Recently, traffic classification studies have focused on IoT networks due to different characteristics of IoT traffic and diverse QoS requirements [138], [139]. The existing traffic classification techniques are either port based, payload based, behaviour based or statistics based with different accuracies. Port based classification have low accuracy as applications keep changing port usage and payload based classification requires deep packet inspections, thus incurring delays. Therefore, statistics based traffic classification are well suited at edge devices considering their limited computational powers as well. AI & ML algorithms have shown promising results in classifying IoT traffic with higher accuracy (up to 83.3% [127] with Decision Trees and up to 94% [135] with CNNs) and their employment at the edge devices can build highly reliable IoT networks with diverse QoS needs. Readers are referred to [127], [134], [135] for detailed surveys of AI & ML techniques for traffic classification.

Like traffic classification, AI & ML has been vastly employed in traffic prediction tasks in wireless IoT networks. Different approaches and learning algorithms have been proposed in literature. Authors in [131] included prior knowledge in information fusion to train neural networks and achieved a 10% improvement over statistical traffic predictions. Similarly, authors in [132] and [133] were able to achieve MAE of 0.3 and 0.002% using LSTMs and regression models respectively. Authors in [140] were able to achieve a Root Mean Square Error (RMSE) of 0.0298 in their prediction accuracy with 500 units in their LSTM model. To predict traffic in an online network, authors in [141] employed Monte Carlo based DRL algorithm to predict IoT traffic and they were able to achieve performance

TABLE 5. Machine learning applications in wireless networks management and distributed control.

Classification	AI/ML Algorithm Used	Problems Addressed	Results	Year	Ref
Distributed Learning	Distributed Cooperative DRL	Energy Aware QoS (Delay and Reliability)	Improved QoS performance over Distributed Cooperative Routing	2020	[125]
	Distributed Actor-Critic DRL	QoE Enhancement for 360 degree Video	High Reward in 50K iterations over QPS	2020	[64]
	Federated DRL	Edge caching for offloading backhual traffic	27% Avg Delay and 15 % Backhaul traffic Improvement over FIFO	2020	[115]
	Federated Reinforcement Learning	Contention Window optimization for Throughput	Faster Convergence (200 iterations) vs RL (700 iterations) and high TH	2021	[31]
	Multiple AI & ML Algorithms	Survey on Distributed Learning	Survey	2022	[121]
Network Management	Support Vector Machines (SVM)	Handover optimization	Reduced Service Interruptions	2017	[65]
	Deep Reinforcement Learning	Backhaul aware User Association	Better QoS and Convergence with only 20 Episodes	2015	[88]
	Deep Reinforcement Learning	Interference Management through spectrum allocation	Improved D2D sum-rate and outage probability over DQN	2016	[89]
	DRL and GCN	Channel Assignments	17.5% Improvement in system TH over DRL without GCN	2020	[80]
	Multi Arm Bandits	Channel Allocation and User Association	MAB supremacy over Dynamic AP & Dynamic Channel for Fairness	2020	[40]
Classification and Prediction	Random Forest, Decision Tress & KNN	Traffic Classification in SDN-IoT	83.3% Accuracy with Six Classes	2020	[127]
	Multiple AI & ML Algorithms	ML based Traffic Classification for QoS	Survey	2014	[130]
	Linear Discriminant Analysis and K Means	Traffic Classification beyond Diff Serv	Classification of Diff Serv classes to 23 sub-classes	2020	[129]
	Long Short Term Memory	IoT Traffic Prediction	1.9% Reduction in MAE, 2.1% increase in R^2 compared to SVM	2022	[131]
	Long Short Term Memory	Traffic Generation Prediction by IoT Devices	LSTM RMSE:0.04,ARIMA RMSE:0.18 Feedforward-NN RMSE:0.08	2021	[132]

ration of 73.4% and 63.18% compared to PCA and LSTM respectively. There are many more research works [142] that employ AI & ML techniques to predict network traffic, however, most of them are focused on IP networks and cellular networks data sets. Few works have been focused on IoT networks but they fail to consider the dynamic characteristics of IoT sensors/devices, different packet structures compared to IP traffic and IoT traffic burstiness. Moreover, these prediction algorithms need to be deployed at the edge base stations/access points, therefore, studies considering computational requirements of proposed algorithms and their suitability to be deployed at edge must also be studied.

Distributed network management and edge intelligence brings several benefits in terms of increased responsiveness, distributed computational load and adaptability. This would significantly increased network scalability and would support large density of IoT sensors, devices and users in the network. However, there are certain network decision that affect neighboring users and require global knowledge for better decision making, such as routing, power control and channel selection decisions. These decisions, although taken at the edge, need to be regulated from a central controller that has the global picture of the network.

Based on the discussion so far, a complete edge intelligent solution for reliable oS in wireless IoT network would have edge APs/base stations employing AI & ML enabled CLD/CLO with their decisions being regulated from the central controller as shown in Figure 10. GNNs along with federated and multi-agent deep reinforcement learning are promising techniques that can be employed in such distributed and edge intelligent network management framework. This makes it a Hybrid network control architecture where intelligence in the edge nodes for traffic classification, traffic prediction and local resource management through CLO/CLD is regulated by the central controller to provide the reliable QoS for a large number of emerging IoT use cases.

VIII. STANDARDIZATION EFFORTS TOWARDS AI & ML EMPLOYMENT IN IOT NETWORKS

Enabling a reliable QoS in wireless IoT networks would require employment of AI & ML algorithms to handle the complexities and emerging challenges. Moreover, AI & ML are being considered as a key requirement for 5G and beyond networks [143]. Realising this importance of AI & ML, the International telecommunication Union (ITU) has formed a focus group on Machine Learning for 5G networks known as FG-ML5G to assist the research community in the application of ML in wireless networks and keep all efforts aligned with common standards. FG-ML5G was tasked to provide technical specifications of ML architectures, protocols, data structures, ML interfaces and algorithms for ML based network optimization. It has released ML architectural requirements for future networks as a guideline to implement AI & ML based solutions to a wide range of network problems and use cases [144]. In this section, we will discuss the essential elements proposed in the FG-ML5G architecture.

A. ML ARCHITECTURE BY FG-ML5G

A standard architecture always proves useful in converging various research efforts towards a common goal. For the same purpose, FG-ML5G has defined the architectural components and ML pipelines along with interfaces that form up the ML architecture (see Figure 11). These

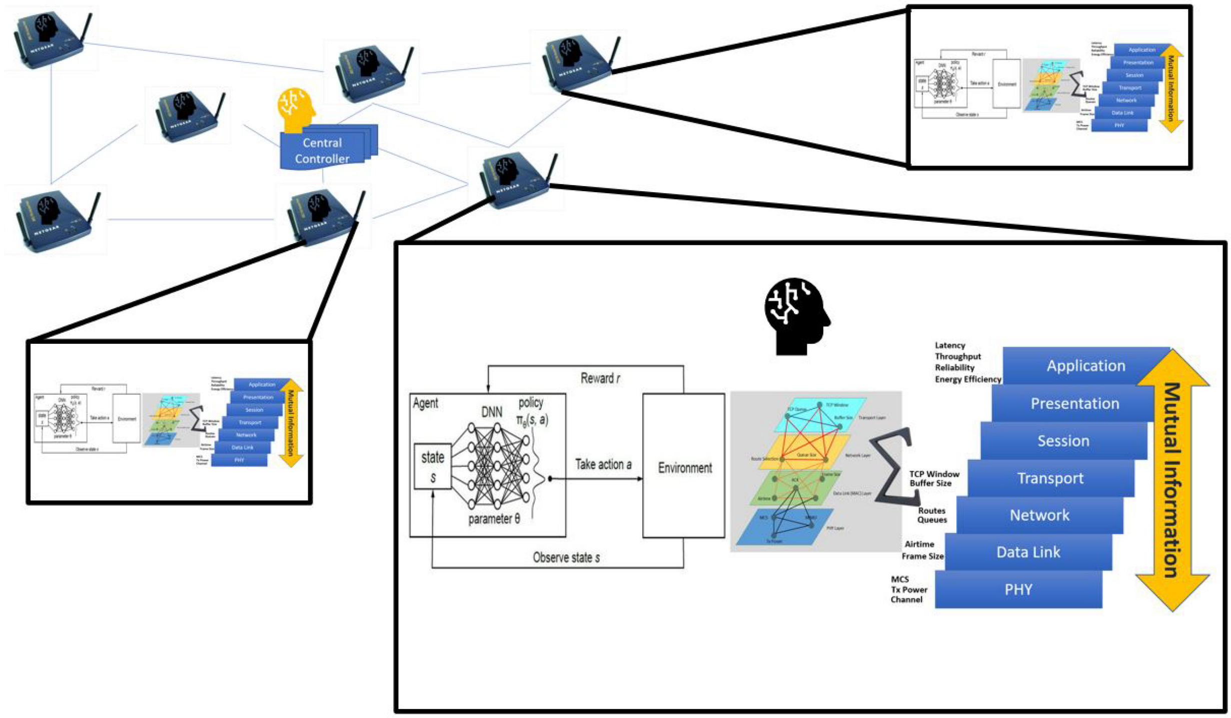

FIGURE 10. Distributed Network Management with Edge APs employing AI based Cross Layer Design.

components include the source (nodes), collector (that's collects data from multiple nodes), Pre-Producer (that prepares data for ML algorithm), Model (the ML Model), Policy (the learnt policy), Distributor (that distributes the ML policy to sinks) and Sink (the nodes or APs). Due to the lack of interpretability of ML models [29], the FG-ML5G architecture includes an ML Sandbox that provides a framework to train ML models over simulated data for testing them before their deployment in production network. An ML Function Orchestrator (MLFO) is deployed to manage all ML pipeline functionalities and selection of ML algorithms for different problems. FG-ML5G architecture guidelines include high-level architecture requirements like enablers for correlation of data coming in from heterogeneous sources with distributed instantiation of ML functionalities. It also provide enablers defining interactions points with the network that are independent of ML functionalities, support for their flexible placement and addition of new ML data sources while the system is running [144].

For interactions, FG-ML5G has defined interface requirements for transfer learning as well as APIs/protocols that match with the ones used in the underlying network technology. The selection of models at startup, and independence of network performance from training and model updates is also among the requirements of an ML architecture. FG-ML5G's standardisation will prove very useful in the long run as it will keep research directions focused and interoperable for seamless integration of various algorithms.

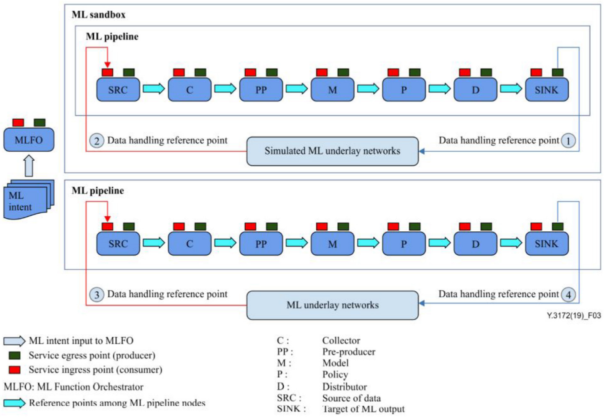
ML sandbox
ML pipeline
SRC
C
PP
M
P
D
SINK
2 Data handling reference point
Simulated ML underlay networks
Data handling reference point 1
MLFO
ML intent
ML pipeline
3 Data handling reference point
ML underlay networks
Data handling reference point 4
Y.3172(19)_F03
ML intent input to MLFO
Service egress point (producer)
Service ingress point (consumer)
MLFO: ML Function Orchestrator
Reference points among ML pipeline nodes
C : Collector
PP : Pre-producer
M : Model
P : Policy
D : Distributor
SRC : Source of data
SINK : Target of ML output

FIGURE 11. ML Architecture by FG-ML5G [144].

B. ROLE OF ML5G ARCHITECTURE IN FUTURE RESEARCH

There is a huge amount of research being done to employ AI & ML in wireless communications and networking domains. Many different ML architectures are being proposed to address different problems [36], [65], however, following ML5G architecture is recommended as it would keep research efforts interoperable with each other. FG-ML5G has defined various ML use cases [145] along with guidelines to follow ML5G architecture while developing solutions. The uses cases range from network-management specific uses cases, to application oriented use cases to facilitate researchers in defining their data collection, storage and processing requirements inline with the ML5G architecture.

IX. FUTURE RESEARCH DIRECTIONS AND EMERGING CHALLENGES

Wireless Networks have seen a massive increase in employment of AI & ML for performance enhancement in almost all domains. A number of comprehensive surveys [29], [41], [44], [59], [60] have been published to cover the AI & ML application in wireless networks. Various ML data related challenges are already identified and require research efforts to cater heterogeneity in wireless network data. These challenges include the labelling of data, the interpretability of ML algorithms, ease of use of ML models, training of ML models in reasonable time and support for ML toolboxes in network simulators [29]. The network research community is trying to solve dynamic network problems through AI & ML that evolve with changing network conditions, ranging from resource allocation, interference coordination, user association, scheduling, rate adaptability, user authentication/security and application aware QoS management. However, there are few research directions that still need exploration and exploitation of AI & ML algorithms especially towards QoS satisfaction in dense IoT networks. We will discuss these research areas and challenges in subsequent paragraphs.

A. CROSS-LAYER DESIGNING AND OPTIMIZATION THROUGH AI & ML

CLD is proven to provide significant benefits in wireless networks by enabling inter-layer information sharing and interactions. However, it is not encouraged and supported historically to preserve the modular structure of the OSI model, which has its own benefits. Previous CLD efforts threatened the modular structure of network communications, however, cross-layer design can be done by introducing new messages following the same modular structure. These messages (packets) would be generated to communicate cross-layer information at crucial instances, determined through AI & ML, to drive the network towards the desired performance which can vary from lower latency at one time instant to high reliability and throughput at another. Learning these dynamic user requirements would help change parameters at various layer through intercommunication and can improve overall QoS in the network.

Similarly, modelling and capturing the combined effects of large number of cross-layer parameters on different user QoS requirements was not possible due to the problem's complexity. Few efforts have been made in the past to employ AI & ML in CLO, however, they targeted a few parameters for joint optimizations. Looking at Figure 3, we can find multiple parameters at different layers that can be jointly optimized. Recently, advances in AI & ML algorithms have introduced new and powerful tools that can handle large number of parameters in CLO. Nowadays, GNNs, CNNs, LSTMs and DRL can be used to develop ML aware cross-layer optimizations to improve overall QoS in wireless network. These optimizations would also improve network capacities to support more IoT sensors/devices in the network.

Though AI & ML can capture complex relationships in cross-layer parameters, model driven AI (CNNs) require a lot of labelled network data for their training. A comprehensive data pre-processing would be required to represent cross-layer parameters and their effect on network KPIs. Many previous research works have already highlighted challenges related to data heterogeneity and labelling in wireless networks, CLD/CLO would add new dimensions to those challenges in data labelling. Moreover, information from OSI layers have different timing constraints (millisecond at MAC layer and microsecond at PHY layer) therefore, handling these different timing constraints is also a complex challenge. Luckily, Network simulators like NS3 and Omnet++ can prove very useful to generate IoT networking data, handle timing constraints of OSI layers and train/test AI & ML algorithms before moving on to real world test beds. Unlike supervised learning, DRL can provide model free learning solution over real time network data through interactions with wireless environment and exploitation of learnt policies over time.

B. MUTUAL INFORMATION AND CROSS ENTROPY ANALYSIS FOR CLO

Although AI & ML algorithms can help in cross-layer optimizations with large number of parameters, they do face a drawback of longer convergence speed. This would not affect non real-time decisions in the core network (associations, security authentications, handovers), however, various performance affecting decisions in the access network require near real-time decision making (for example the selection of radio resources, transmit power, MCS and contention window, etc.). To improve convergence speed of AI & ML algorithms, mutual information and cross entropy analysis of various parameters at different layers can identify similar information carrying parameters that would have the same effect on desired QoS metrics. This would significantly reduce the number of parameters to be optimized, thus causing dimensionality reduction of the underlying AI & ML problem. Such a study has not been done for a stable CLO solution. Mutual information and cross entropy analysis would involve probability theory and its results can be embedded in the AI frameworks to speed up the convergence of CLO.

C. STUDY OF INTER-LAYER DEPENDENCIES THROUGH GNNS

Current research on CLD and CLO lacks a formal study of relationships between layer parameters. Moreover, CLD suffers from instability problems if not designed correctly. The instability occurs due to the usage of dependent variables (parameters) in more than one loop (by optimization algorithms). This requires the creation of dependency graphs prior to employing AI & ML algorithms in CLD [55]. The cross-layer parameters can have multiple relationships among themselves that may go beyond adjacent layers. These relationships posses typical graphical structure as shown in Figure 6. Therefore, Graph Neural networks (GNNs) can be employed to study inter-layer dependencies and capture relationships between layer parameters. Recently, GNNs have seen rise in their application to wireless networks and are used in combination with other ML algorithms for channel allocation, routing and interference mapping. GNNs employment with DRL and RNN have been seen in literature, however, their usage with other AI & ML algorithms can be studied for a wide range of wireless network problems.

D. MULTI-RAT SDN ARCHITECTURES AND NETWORK MANAGEMENT THROUGH AI

Future networks are becoming increasingly heterogeneous with multiple radio access technologies being employed to increase network capacity. With multiple RATs, handovers across multiple access technologies, medium access and optimal routes with multiple RATs on path add additional dimensions to the task complexity. Since different access technologies (e.g., WLAN, BLE, ZigBee) use unlicensed band (2.4 GHz) for communication, they require complex interference management and interference-aware RAT selection. Current research on multi-RAT networks lack multi-dimensional decision making that takes into account RAT specific congestion, routing latencies and channel interference, specifically for QoS satisfaction of IoT users/applications.

Software Defined Networking (SDN) possess all the necessary traits to address the above mentioned challenges in multi-RAT networks and even employ CLD. An SDN controller can collect information from distributed network devices and OSI layers by deploying various network apps. This information can be fused using AI & ML algorithms to undertake multi-dimensional decisions in multi-RAT networks. However, research efforts are required to develop information collection and information fusion applications in multi-RAT SDN controllers. AI & ML can really help in efficient information fusion for CLO and CLD in multi-RAT networks, however, very few SDN architectures have been proposed in literature that support multiple RATs and have been tested in real world environments. Most of the SDNs proposed in literature support 4G, 5G and WLAN access technologies while neglecting LoRa, BLE and ZigBee technologies which are used by many IoT sensors and devices. Therefore, research efforts to develop SDN controllers supporting LoRa, BLE and ZigBee along with WLAN and Cellular technologies are required. Moreover, the different timing constraints in CLD/CLO of OSI layers would remain a complex challenge in multi-RAT networks as well.

E. TRAFFIC CLASSIFICATION AND PREDICTION FOR RELIABLE QOS

In order to meet QoS requirements in IoT networks, it is necessary to understand the QoS needs of the applications and IoT devices. Unlike multimedia traffic, IoT traffic is not categorised into multiple classes depending on QoS needs. As such, traffic classification algorithms are required to classify IoT traffic into multiple classes representing their throughput, latency, reliability and energy efficiency requirements. This classification is a complex and cumbersome task as all packets appear similar with variable packet sizes, packet arrival rates, source ports and temporal characteristics. Although there is a plethora of research done on IP traffic classification, IoT network traffic classification on the edge devices remains an open research. Machine learning, especially, deep learning algorithms can be employed to analyze multiple features of IoT traffic and classify them into various QoS classes. This can help create more QoS classes and would also improve QoS granularity in the network, offering more control over user QoE.

Besides traffic classification, another way to classify IoT traffic is to standardise IoT traffic classes and include them in the IoT device design. IoT devices can then tag packets with their QoS requirements while sending traffic. This would enable base stations / access points to treat traffic as per their QoS requirements and manage network resources accordingly, however, changes in present radio access technologies to incorporate those IoT traffic tags in their traffic handling procedures and QoS frameworks would be required.

Meeting diverse QoS requirements is not only the matter of using powerful and recent hardware, rather, it can also be achieved through efficient mapping of user needs with available resources [146] therefore, traffic prediction and forecasting can also benefit in improving QoS in IoT networks. Knowing traffic requirements in advance allows efficient resource sharing and utilisation to avoid violations of agreed Service Level Agreements (SLA) in the network. Machine learning algorithms like LSTMs and GRUs can be employed to study complex time series traffic data for accurate forecasting however, they require long training. Classical techniques like Auto Regressive Integrated Moving Average (ARIMA) and Exponential Smoothing are fast to train however, their prediction accuracy are lower. Therefore, high accuracy prediction algorithms with training time constraints are required to forecast network traffic in milli-seconds timescale for rapidly changing QoS requirements in IoT networks.

F. MULTI-MODAL AI AND ORTHOGONAL LEARNING

AI & ML have been employed in every layer of network protocol stack from the application layer to the physical layer [44], [41], [60]. Network researchers have mostly employed AI & ML to one or two problems at a time like channel-aware user association, load balancing, congestion control and spectrum allocation etc. It is known that the addition of more parameters increases problem complexity, however, many learning algorithms can run simultaneously in a wireless network for performance enhancement. For example GNNs can be used to optimize deployment of network nodes and assigning channels (frequencies) while multiple DRL algorithms can be run to manage handovers, associations, transmit powers, scheduling, etc at the same time. Similarly, decision trees, LSTMs and SVMs can be run to optimize frame lengths, airtime and TCP window size to improve link performances while other algorithms are addressing different problems. This would create an AI pipeline where multiple AI algorithms would run towards optimising their target parameter with the end objective of user QoS satisfaction. This can definitely produce improved results, however, there are multiple parameters that have overlapping effects due to non-orthogonal processes (e.g., decisions related to airtime would affect decisions related to TCP window size or transmit power). Therefore, a mechanism to develop inter-communication between different AI algorithms would be required to keep local AI algorithm decisions aligned with the overall network performance objectives.

Moreover, decision making in wireless networks based on inputs from multiple OSI layers and different range of devices in the network (Switches, Gateways, Access Points) can be represented as different modalities and, resultantly, multi-modal learning algorithms and approaches can be employed. As the data from these layers and devices would have different structures, veracity and, timescales, therefore, learning upon such data is a complex task. Multi-modal learning requires data from all layers of the OSI model which would require a good network monitoring framework. This adds to overhead in network operations, therefore, multi-model learning models with limited or insufficient data are also required for creating a balance between performance improvement and overhead addition in the network.

X. CONCLUSION

The number of devices requiring connectivity is growing at a fast pace with IoT networks becoming ever more dense. At the same time, IoT devices are becoming increasingly heterogeneous in terms of QoS requirements. Research efforts are ongoing to improve network capacities and QoS provisioning in dense IoT networks and new technologies and techniques are being proposed to meet emerging requirements. Among other technologies, Cross-layer Design, Cross-layer Optimization and Distributed Network Management of multi-RAT IoT networks are promising methods to meet diverse QoS requirements in dense IoT networks. Previous CLD/CLO research targets only a few OSI layer parameters while neglecting the true potential of all layers parameters optimizations which was not possible in past due to problem complexity. Moreover, they had targeted improvements in technical network parameters while overseeing user QoS requirements and QoE. Similarly, CLD and CLO is not considered in multi-RAT IoT networks to enable high capacity QoS aware networks. On top of it, QoS requirements in various IoT use cases have become far more stringent and require edge devices to learn QoS needs and take decisions of network resource management. Research efforts in these domains are disconnected from each other, therefore, unified efforts to develop edge intelligence and distributed network management where CLD and CLO would be employed in multi-RAT network on the edge are required.

To develop such complex solutions, advancements in AI & ML have underpinned new algorithms and methods for capturing inter-layer dependencies, performing cross-layer optimization and taking distributed decisions. These AI & ML based solutions should be developed while following FG-ML5G architectural guidelines to keep research efforts aligned with a common standard. However, there are numerous research challenges that needs to be solved by the networking research community.

REFERENCES

[1] P. Karwel et al., "Ericsson mobility report," Technol. Emerg. Bus., Ericsson AB, Stockholm, Sweden, Rep. EAB-21, 2021.

[2] "Cisco annual Internet report 2018-2023," Cisco, San Jose, CA, USA, White Paper, Mar. 2020.

[3] "3GPP technical specification groups for RAN, SA and CT, release 15," 3GPP, Sophia Antipolis, France, Rep. TR 21.915, 2017.

[4] M. Condoluci and T. Mahmoodi, "Softwarization and virtualization in 5G mobile networks: Benefits, trends and challenges," *Comput. Netw.*, vol. 146, pp. 65–84, Dec. 2018.

[5] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless network virtualization with SDN and C-RAN for 5G networks: Requirements, opportunities, and challenges," *IEEE Access*, vol. 5, pp. 19099–19115, 2017.

[6] M. Pooyandeh and I. Sohn, "Edge network optimization based on AI techniques: A survey," *Electronics*, vol. 10, no. 22, p. 2830, 2021.

[7] I. Al-Anbagi, M. Erol-Kantarci, and H. T. Mouftah, "A survey on cross-layer quality-of-service approaches in WSNs for delay and reliability-aware applications," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 525–552, 1st Quart., 2016.

[8] V. Srivastava and M. Motani, "Cross-layer design: A survey and the road ahead," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 112–119, Dec. 2005.

[9] F. Foukalas, V. Gazis, and N. Alonistioti, "Cross-layer design proposals for wireless mobile networks: A survey and taxonomy," *IEEE Commun. Surveys Tuts.*, vol. 10, no. 1, pp. 70–85, 1st Quart., 2008.

[10] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 756–769, Dec. 1997.

[11] E. Charfi, L. Chaari, and L. Kamoun, "PHY/MAC enhancements and QoS mechanisms for very high throughput WLANs: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1714–1735, 4th Quart., 2013.

[12] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 5, pp. 1056–1066, May 2005.

[13] S. M. Abd El-atty, "Efficient packet scheduling with pre-defined QoS using cross-layer technique in wireless networks," in *Proc. 11th IEEE Symp. Comput. Commun. (ISCC)*, 2006, pp. 820–826.

[14] Q. Zhang, F. Yang, and W. Zhu, "Cross-layer QoS support for multimedia delivery over wireless Internet," *EURASIP J. Adv. Signal Process.*, vol. 2005, 2005, Art. no. 896852, doi: 10.1155/ASP.2005.207.

[15] B. Fu, Y. Xiao, H. Deng, and H. Zeng, "A survey of cross-layer designs in wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 110–126, 1st Quart., 2014.

[16] R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multihop wireless networks," in *Proc. 22nd Annu. Joint Conf. IEEE Comput. Commun. Soc.*, vol. 1, 2003, pp. 702–711.

[17] B. Zhang and H. T. Mouftah, "QoS routing for wireless ad hoc networks: Problems, algorithms, and protocols," *IEEE Commun. Mag.*, vol. 43, no. 10, pp. 110–117, Oct. 2005.

[18] M. Di Francesco, G. Anastasi, M. Conti, S. K. Das, and V. Neri, "Reliability and energy-efficiency in IEEE 802.15.4/ZigBee sensor networks: An adaptive and cross-layer approach," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1508–1524, Sep. 2011.

[19] L. Shu, M. Hauswirth, L. Wang, Y. Zhang, and J. H. Park, "Cross-layer optimized data gathering in wireless multimedia sensor networks," in *Proc. Int. Conf. Comput. Sci. Eng.*, vol. 2, 2009, pp. 961–966.

[20] S. K. Rao and R. Prasad, "Impact of 5G technologies on industry 4.0," *Wireless Pers. Commun.*, vol. 100, no. 1, pp. 145–159, 2018.

[21] B. Pradhan, S. Bhattacharyya, and K. Pal, "IoT-based applications in healthcare devices," *J. Healthcare Eng.*, vol. 2021, Mar. 2021, Art. no. 6632599.

[22] Y. Fu and Q. Zhu, "Joint optimization methods for nonconvex resource allocation problems of decode-and-forward relay-based OFDM networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 4993–5006, Jul. 2016.

[23] M. H. Rahman, M. M. Mowla, and S. Shanto, "Convolutional neural network based optimization approach for wireless resource management," in *Proc. 2nd Int. Conf. Adv. Inf. Commun. Technol. (ICAICT)*, 2020, pp. 280–285.

[24] Q. Meng, K. Wang, B. Liu, T. Miyazaki, and X. He, "QoE-based big data analysis with deep learning in pervasive edge environment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–6.

[25] K. Ryu and W. Kim, "Multi-objective optimization of energy saving and throughput in heterogeneous networks using deep reinforcement learning," *Sensors*, vol. 21, no. 23, p. 7925, 2021.

[26] C.-W. Huang, C.-T. Chiang, and Q. Li, "A study of deep learning networks on mobile traffic forecasting," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, 2017, pp. 1–6.

[27] P. W. Battaglia et al., "Relational inductive biases, deep learning, and graph networks," 2018, *arXiv:1806.01261*.

[28] T. Zhu, X. Chen, L. Chen, W. Wang, and G. Wei, "GCLR: GNN-based cross layer optimization for Multipath TCP by routing," *IEEE Access*, vol. 8, pp. 17060–17070, 2020.

[29] E. Coronado, S. Bayhan, A. Thomas, and R. Riggio, "AI-empowered software-defined WLANs," *IEEE Commun. Mag.*, vol. 59, no. 3, pp. 54–60, Mar. 2021.

[30] E. Coronado, J. Villalón, and A. Garrido, "Dynamic AIFSN tuning for improving the QoS over IEEE 802.11 WLANs," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, 2015, pp. 73–78.

[31] R. Ali, Y. B. Zikria, S. Garg, A. K. Bashir, M. S. Obaidat, and H. S. Kim, "A federated reinforcement learning framework for incumbent technologies in beyond 5G networks," *IEEE Netw.*, vol. 35, no. 4, pp. 152–159, Jul./Aug. 2021.

[32] J. D. Herath, A. Seetharam, and A. Ramesh, "A deep learning model for wireless channel quality prediction," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[33] C. Liu et al., "Progressive neural architecture search," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 19–34.

[34] W. Wydmański and S. Szott, "Contention window optimization in IEEE 802.11 ax networks with deep reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2021, pp. 1–6.

[35] R. Karmakar, S. Chattopadhyay, and S. Chakraborty, "A deep probabilistic control machinery for auto-configuration of WiFi link parameters," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8330–8340, Dec. 2020.

[36] E. Coronado, A. Thomas, S. Bayhan, and R. Riggio, "AiOS: An intelligence layer for SD-WLANs," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp.*, 2020, pp. 1–9.

[37] H. Feng, Y. Shu, S. Wang, and M. Ma, "SVM-based models for predicting WLAN traffic," in *Proc. IEEE Int. Conf. Commun.*, vol. 2, 2006, pp. 597–602.

[38] T. Nishio et al., "Proactive received power prediction using machine learning and depth images for mmWave networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2413–2427, Nov. 2019.

[39] E. Zeljković, N. Slamnik-Kriještorac, S. Latré, and J. M. Marquez-Barja, "ABRAHAM: Machine learning backed proactive handover algorithm using SDN," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 4, pp. 1522–1536, Dec. 2019.

[40] Á. López-Raventós and B. Bellalta, "Concurrent decentralized channel allocation and access point selection using multi-armed bandits in multi BSS WLANs," *Comput. Netw.*, vol. 180, Oct. 2020, Art. no. 107381.

[41] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.

[42] R. Li et al., "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 175–183, Oct. 2017.

[43] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, 4th Quart., 2019.

[44] S. Szott et al., "Wi-Fi meets ML: A survey on improving IEEE 802.11 performance with machine learning," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1843–1893, 3rd Quart., 2022.

[45] L. Qian et al., "Distributed learning for wireless communications: Methods, applications and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 326–342, Apr. 2022.

[46] J. Park et al., "Communication-efficient and distributed learning over wireless networks: Principles and applications," *Proc. IEEE*, vol. 109, no. 5, pp. 796–819, May 2021.

[47] E. Muscinelli, S. S. Shinde, and D. Tarchi, "Overview of distributed machine learning techniques for 6G networks," *Algorithms*, vol. 15, no. 6, p. 210, 2022.

[48] H. Balakrishnan and R. H. Katz, "Explicit loss notification and wireless Web performance," in *Proc. IEEE Globecom*, vol. 98, 1998, pp. 1–5.

[49] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, "Application-driven cross-layer optimization for video streaming over wireless networks," *IEEE Commun. Mag.*, vol. 44, no. 1, pp. 122–130, Jan. 2006.

[50] M. Miyoshi, M. Sugano, and M. Murata, "Improving TCP performance for wireless cellular networks by adaptive FEC combined with explicit loss notification," *IEICE Trans. Commun.*, vol. 85, no. 10, pp. 2208–2213, 2002.

[51] A. Chiumento, N. Marchetti, and I. Macaluso, "Energy efficient WSN: A cross-layer graph signal processing solution to information redundancy," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2019, pp. 645–650.

[52] D. Wu and S. Ci, "Cross-layer design for combining adaptive modulation and coding with hybrid ARQ," in *Proc. Int. Conf. Wireless Commun. Mobile Comput.*, 2006, pp. 147–152.

[53] J. C.-S. Wu, C.-W. Cheng, N.-F. Huang, and G.-K. Ma, "Intelligent handoff for mobile wireless Internet," *Mobile Netw. Appl.*, vol. 6, no. 1, pp. 67–79, 2001.

[54] C.-C. Tseng, L.-H. Yen, H.-H. Chang, and K.-C. Hsu, "Topology-aided cross-layer fast handoff designs for IEEE 802.11/mobile IP environments," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 156–163, Dec. 2005.

[55] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," *IEEE Wireless Commun.*, vol. 12, no. 1, pp. 3–11, Feb. 2005.

[56] K. J. Åström and B. Wittenmark, *Adaptive Control*. Reading, MA, USA: Addison-Wesley, 1995.

[57] A. Chiumento, B. Reynders, Y. Murillo, and S. Pollin, "Building a connected BLE mesh: A network inference study," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, 2018, pp. 296–301.

[58] I. K. Fodor, "A survey of dimension reduction techniques," U.S. Dept. Energy, Center Appl. Sci. Comput., Lawrence Livermore Nat. Lab., Livermore, CA, USA, Rep. UCRL-ID-148494, 2002.

[59] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[60] M. Kulin, T. Kazaz, E. De Poorter, and I. Moerman, "A survey on machine learning-based performance improvement of wireless networks: PHY, MAC and network layer," *Electronics*, vol. 10, no. 3, p. 318, 2021.

[61] A. Musaddiq, Z. Nain, Y. Ahmad Qadri, R. Ali, and S. W. Kim, "Reinforcement learning-enabled cross-layer optimization for low-power and lossy networks under heterogeneous traffic patterns," *Sensors*, vol. 20, no. 15, p. 4158, 2020.

[62] Y. Du, Y. Xu, L. Xue, L. Wang, and F. Zhang, "An energy-efficient cross-layer routing protocol for cognitive radio networks using apprenticeship deep reinforcement learning," *Energies*, vol. 12, no. 14, p. 2829, 2019.

[63] S. Chitnavis and A. Kwasinski, "Cross layer routing in cognitive radio networks using deep reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–6.

[64] M. Krouka, A. Elgabli, M. S. Elbamby, C. Perfecto, M. Bennis, and V. Aggarwal, "Cross layer optimization and distributed reinforcement learning approach for tile-based 360 degree wireless video streaming," 2020, *arXiv:2011.06356*.

[65] I. Chih-Lin, Q. Sun, Z. Liu, S. Zhang, and S. Han, "The big-data-driven intelligent wireless network: Architecture, use cases, solutions, and future trends," *IEEE Veh. Technol. Mag.*, vol. 12, no. 4, pp. 20–29, Dec. 2017.

[66] R. Mijumbi, S. Hasija, S. Davy, A. Davy, B. Jennings, and R. Boutaba, "A connectionist approach to dynamic resource management for virtualised network functions," in *Proc. 12th Int. Conf. Netw. Service Manage. (CNSM)*, 2016, pp. 1–9.

[67] A. Chowdhury, G. Verma, C. Rao, A. Swami, and S. Segarra, "Unfolding WMMSE using graph neural networks for efficient power allocation," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6004–6017, Sep. 2021.

[68] A. Alarifi, A. A. AlZubi, and A. Alwadain, "Multimodal learning optimization for enhancing channel usage in wireless communication," *Circuits, Syst., Signal Process.*, vol. 39, no. 2, pp. 1146–1162, 2020.

[69] Y. Koda et al., "Communication-efficient multimodal split learning for mmWave received power prediction," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1284–1288, Jun. 2020.

[70] Y. Yang, F. Gao, C. Xing, J. An, and A. Alkhateeb, "Deep multimodal learning: Merging sensory data for massive MIMO channel prediction," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1885–1898, Jul. 2021.

[71] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.

[72] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, May 2020. [Online]. Available: https://doi.org/10.1162/neco_a_01273

[73] Y. Liu, X. Feng, and Z. Zhou, "Multimodal video classification with stacked contractive autoencoders," *Signal Process.*, vol. 120, pp. 761–766, Mar. 2016.

[74] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, *arXiv:1707.07250*.

[75] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, *arXiv:1606.01847*.

[76] A. Mishra, S. Banerjee, and W. Arbaugh, "Weighted coloring based channel assignment for WLANs," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 9, no. 3, pp. 19–31, 2005.

[77] B. Yao, J. Yin, H. Li, H. Zhou, and W. Wu, "Channel resource allocation based on graph theory and coloring principle in cellular networks," in *Proc. IEEE 3rd Int. Conf. Cloud Comput. Big Data Anal. (ICCCBDA)*, 2018, pp. 439–445.

[78] K. Rusek, J. Suárez-Varela, A. Mestres, P. Barlet-Ros, and A. Cabellos-Aparicio, "Unveiling the potential of graph neural networks for network modeling and optimization in SDN," in *Proc. ACM Symp. SDN Res.*, 2019, pp. 140–151.

[79] Y. Kong, D. Petrov, V. Räisänen, and A. Ilin, "Path-link graph neural network for IP network performance prediction," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, 2021, pp. 170–177.

[80] K. Nakashima, S. Kamiya, K. Ohtsu, K. Yamamoto, T. Nishio, and M. Morikura, "Deep reinforcement learning-based channel allocation for wireless LANs with graph convolutional networks," *IEEE Access*, vol. 8, pp. 31823–31834, 2020.

[81] W. Jiang, "Graph-based deep learning for communication networks: A survey," 2021, *arXiv:2106.02533*.

[82] X. Li, N. Du, H. Li, K. Li, J. Gao, and A. Zhang, "A deep learning approach to link prediction in dynamic networks," in *Proc. SIAM Int. Conf. Data Min.*, 2014, pp. 289–297.

[83] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[84] O. Orhan, V. N. Swamy, T. Tetzlaff, M. Nassar, H. Nikopour, and S. Talwar, "Connection management xAPP for O-RAN RIC: A graph neural network and reinforcement learning approach," 2021, *arXiv:2110.07525*.

[85] E. Coronado, R. Riggio, J. Villalón, and A. Garrido, "Wi-balance: Channel-aware user association in software-defined Wi-Fi networks," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp.*, 2018, pp. 1–9.

[86] T. Lei, X. Wen, Z. Lu, and Y. Li, "A semi-matching based load balancing scheme for dense IEEE 802.11 WLANs," *IEEE Access*, vol. 5, pp. 15332–15339, 2017.

[87] W. Wydmański and S. Szott, "Contention window optimization in IEEE 802.11ax networks with deep reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2021, pp. 1–6.

[88] P. V. Klaine, M. Jaber, R. D. Souza, and M. A. Imran, "Backhaul aware user-specific cell association using Q-learning," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3528–3541, Jul. 2019.

[89] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1828–1840, Feb. 2020.

[90] X. Wang, S. Mao, and M. X. Gong, "A survey of LTE Wi-Fi coexistence in unlicensed bands," *Mobile Comput. Commun.*, vol. 20, no. 3, pp. 17–23, 2017.

[91] X. Wang, S. Mao, and M. X. Gong, "A survey of LTE Wi-Fi coexistence in unlicensed bands," *GetMobile Mobile Comput. Commun.*, vol. 20, no. 3, pp. 17–23, 2017.

[92] M. Qin et al., "Service-oriented energy-latency tradeoff for IoT task partial offloading in MEC-enhanced multi-RAT networks," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1896–1907, Feb. 2021.

[93] D. Więcek, I. Michalski, K. Rzeźniczak, and D. Wypiór, "Multi-RAT orchestration method for heterogeneous wireless networks," *Appl. Sci.*, vol. 11, no. 18, p. 8281, 2021.

[94] G. Leenders, G. Callebaut, G. Ottoy, L. Van der Perre, and L. De Strycker, "Multi-RAT for IoT: The potential in combining LoRaWAN and NB-IoT," *IEEE Commun. Mag.*, vol. 59, no. 12, pp. 98–104, Dec. 2021.

[95] H. Lee, S. Vahid, and K. Moessner, "Machine learning based RATs selection supporting multi-connectivity for reliability," in *Proc. Int. Conf. Cogn. Radio Oriented Wireless Netw.*, 2019, pp. 31–41.

[96] A. Tolli and P. Hakalin, "Adaptive load balancing between multiple cell layers," in *Proc. IEEE 56th Veh. Technol. Conf.*, vol. 3, 2002, pp. 1691–1695.

[97] G. Panwar, R. Tourani, T. Mick, A. Mtibaa, and S. Misra, "DICE: Dynamic multi-RAT selection in the ICN-enabled wireless edge," in *Proc. Workshop Mobility Evolv. Internet Archit.*, 2017, pp. 31–36.

[98] J. S. Perez, S. K. Jayaweera, and S. Lane, "Machine learning aided cognitive RAT selection for 5G heterogeneous networks," in *Proc. IEEE Int. Black Sea Conf. Commun. Netw. (BlackSeaCom)*, 2017, pp. 1–5.

[99] A. N. Manjeshwar, A. Roy, P. Jha, and A. Karandikar, "Control and management of multiple RATs in wireless networks: An SDN approach," in *Proc. IEEE 2nd 5G World Forum (5GWF)*, 2019, pp. 596–601.

[100] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: A software-defined networking platform for 5G radio access networks," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 2, pp. 715–728, Jun. 2019.

[101] L. Galluccio, S. Milardo, G. Morabito, and S. Palazzo, "SDN-WISE: Design, prototyping and experimentation of a stateful SDN solution for wireless sensor networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2015, pp. 513–521.

[102] B. Dezfouli, V. Esmaeelzadeh, J. Sheth, and M. Radi, "A review of software-defined WLANs: Architectures and central control mechanisms," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 431–463, 1st Quart., 2019.

[103] N. Bizanis and F. A. Kuipers, "SDN and virtualization solutions for the Internet of Things: A survey," *IEEE Access*, vol. 4, pp. 5591–5606, 2016.

[104] S. W. Kim and B.-S. Kim, "Adaptive cross-layer packet scheduling method for multimedia services in wireless personal area networks," *J. Commun. Netw.*, vol. 8, no. 3, pp. 297–305, 2006.

[105] G. A. Shah, W. Liang, and O. B. Akan, "Cross-layer framework for QoS support in wireless multimedia sensor networks," *IEEE Trans. Multimedia*, vol. 14, no. 5, pp. 1442–1455, Oct. 2012.

[106] L. Song, Y. Liao, K. Bian, L. Song, and Z. Han, "Cross-layer protocol design for CSMA/CD in full-duplex WiFi networks," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 792–795, Apr. 2016.

[107] J. J. L. Escobar, F. Gil-Castiñeira, and R. P. D. Redondo, "JMAC protocol: A cross-layer multi-hop protocol for LoRa," *Sensors*, vol. 20, no. 23, p. 6893, 2020.

[108] H. Baligh et al., "Cross-layer provision of future cellular networks: A WMMSE-based approach," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 56–68, Nov. 2014.

[109] M. S. Allahham, A. A. Abdellatif, N. Mhaisen, A. Mohamed, A. Erbad, and M. Guizani, "Multi-agent reinforcement learning for network selection and resource allocation in heterogeneous multi-RAT networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 1287–1300, Jun. 2022.

[110] M. Yan, G. Feng, J. Zhou, and S. Qin, "Smart multi-RAT access based on multiagent reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4539–4551, May 2018.

[111] A. Arefin, R. Rivas, R. Tabassum, and K. Nahrstedt, "OpenSession: SDN-based cross-layer multi-stream management protocol for 3D teleimmersion," in *Proc. 21st IEEE Int. Conf. Netw. Protocols (ICNP)*, 2013, pp. 1–10.

[112] G. Carella et al., "Cross-layer service to network orchestration," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 6829–6835.

[113] F. Bannour, S. Souihi, and A. Mellouk, "Distributed SDN control: Survey, taxonomy, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 333–354, 1st Quart., 2018.

[114] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for Internet of things data analysis: A survey," *Digit. Commun. Netw.*, vol. 4, no. 3, pp. 161–175, 2018.

[115] X. Wang, C. Wang, X. Li, V. C. Leung, and T. Taleb, "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9441–9455, Oct. 2020.

[116] J. Pan and J. McElhannon, "Future edge cloud and edge computing for Internet of Things applications," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 439–449, Feb. 2018.

[117] G. Tanganelli, C. Vallati, and E. Mingozzi, "Ensuring quality of service in the Internet of Things," in *New Advances in the Internet of Things*. Cham, Switzerland: Springer, 2018, pp. 139–163.

[118] J. Gao, W. Zhuang, M. Li, X. Shen, and X. Li, "MAC for machine-type communications in industrial IoT—Part I: Protocol design and analysis," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9945–9957, Jun. 2021.

[119] J. Gao, M. Li, W. Zhuang, X. Shen, and X. Li, "MAC for machine-type communications in industrial IoT—Part II: Scheduling and numerical results," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9958–9969, Jun. 2021.

[120] C. Pei et al., "Latency-based WiFi congestion control in the air for dense WiFi networks," in *Proc. IEEE/ACM 25th Int. Symp. Qual. Service (IWQoS)*, 2017, pp. 1–10.

[121] O. Nassef, W. Sun, H. Purmehdi, M. Tatipamula, and T. Mahmoodi, "A survey: Distributed machine learning for 5G and beyond," *Comput. Netw.*, vol. 207, Apr. 2022, Art. no. 108820.

[122] O. Yazdani and G. Mirjalily, "A survey of distributed resource allocation for device-to-device communication in cellular networks," in *Proc. Artif. Intell. Signal Process. Conf. (AISP)*, 2017, pp. 236–239.

[123] W. Dai, Y. Shen, and M. Z. Win, "Distributed power allocation for cooperative wireless network localization," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 28–40, Jan. 2015.

[124] J. Zhang, D. Zhang, K. Xie, S. He, H. Qiao, and B. Zeng, "A cooperative routing algorithm for maximizing network lifetime," in *Proc. China Conf. Wireless Sens. Netw.*, 2012, pp. 665–675.

[125] D. Wang, J. Liu, and D. Yao, "An energy-efficient distributed adaptive cooperative routing based on reinforcement learning in wireless multimedia sensor networks," *Comput. Netw.*, vol. 178, Sep. 2020, Art. no. 107313.

[126] G. R. R. Dewa, A. S. Alfathani, C. Park, and I. Sohn, "Distributed channel assignment for ultra-dense wireless networks using belief propagation," *IEEE Access*, vol. 9, pp. 117040–117051, 2021.

[127] A. I. Owusu and A. Nayak, "An intelligent traffic classification in SDN-IoT: A machine learning approach," in *Proc. IEEE Int. Black Sea Conf. Commun. Netw. (BlackSeaCom)*, 2020, pp. 1–6.

[128] *IEEE Standard for Information Technology—Local and Metropolitan Area Networks–Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications—Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, IEEE Standard 802.11e-2005, 2005.

[129] D. Aureli, A. Cianfrani, A. Diamanti, J. M. S. Vilchez, and S. Secci, "Going beyond diffserv in IP traffic classification," in *Proc. IEEE/IFIP Netw. Operations Manage. Symp.*, 2020, pp. 1–6.

[130] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, "Towards the deployment of machine learning solutions in network traffic classification: A systematic survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1988–2014, 2nd Quart., 2019.

[131] C. Pan, Y. Wang, H. Shi, J. Shi, and R. Cai, "Network traffic prediction incorporating prior knowledge for an intelligent network," *Sensors*, vol. 22, no. 7, p. 2674, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2674

[132] S. P. Khedkar, R. A. Canessane, and M. L. Najafi, "Prediction of traffic generated by IoT devices using statistical learning time series algorithms," *Wireless Commun. Mobile Comput.*, vol. 2021, Aug. 2021, Art. no. 5366222.

[133] A. Hameed, J. Violos, N. Santi, A. Leivadeas, and N. Mitton, "A machine learning regression approach for throughput estimation in an IoT environment," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Physical Social Comput. (CPSCom) IEEE Smart Data (SmartData) IEEE Congr. Cybermatics (Cybermatics)*, 2021, pp. 29–36.

[134] T. T. Nguyen and G. Armitage, "A survey of techniques for Internet traffic classification using machine learning," *IEEE Commun. Surveys Tuts.*, vol. 10, no. 4, pp. 56–76, 4th Quart., 2008.

[135] M. Camelo, P. Soto, and S. Latré, "A general approach for traffic classification in wireless networks using deep learning," *IEEE Trans. Netw. Service Manag.*, early access, Nov. 24, 2021, doi: 10.1109/TNSM.2021.3130382.

[136] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *Int. J. Netw. Manage.*, vol. 25, no. 5, pp. 355–374, 2015.

[137] A. Callado, C. Kamienski, S. Fernandes, D. Sadok, and G. Szabó, "A survey on Internet traffic identification and classification," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 3, pp. 37–52, 3rd Quart., 2009.

[138] H. Tahaei, F. Afifi, A. Asemi, F. Zaki, and N. B. Anuar, "The rise of traffic classification in IoT networks: A survey," *J. Netw. Comput. Appl.*, vol. 154, Mar. 2020, Art. no. 102538.

[139] R. M. AlZoman and M. J. Alenazi, "A comparative study of traffic classification techniques for smart city networks," *Sensors*, vol. 21, no. 14, p. 4677, 2021.

[140] A. R. Abdellah and A. Koucheryavy, "Deep learning with long short-term memory for IoT traffic prediction," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Cham, Switzerland: Springer, 2020, pp. 267–280.

[141] L. Nie et al., "A reinforcement learning-based network traffic prediction mechanism in Intelligent Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2169–2180, Mar. 2021.

[142] A. Chen, J. Law, and M. Aibin, "A survey on traffic prediction techniques using artificial intelligence for communication networks," *Telecom*, vol. 2, no. 4, pp. 518–535, 2021.

[143] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[144] *ITU-ML5G, Architectural Framework for Machine Learning in Future Networks Including IMT-2020*, Rec. ITU-T Y.3172, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[145] *ITU-ML5G, Machine Learning in Future Networks Including IMT-2020: Use Cases*, Rec. ITU-Y.3170, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[146] *AI and ML—Enablers for Beyond 5G Networks, Version 1.0*, 5G PPP Technol. Board, Heidelberg, Germany, 2021.

KAMRAN ZIA received the bachelor's (with Distinction) and master's (President's Gold Medal) degrees in avionics engineering from the College of Aeronautical Engineering, National University of Sciences and Technology, Pakistan, in 2011 and 2018, respectively. He is currently pursuing the Ph.D. degree with the University of Twente, The Netherlands. His research interests include machine learning and artificial intelligence in wireless networks, distributed and edge learning, quality of service in wireless networks, multi-RAT networks, cross-layer design and optimizations, resource management, and software-defined radios and networks.

ALESSANDRO CHIUMENTO is an Assistant Professor with the Pervasive Systems Group, University of Twente. His research interests include distributed intelligence over networked systems, mobile computing, autonomous multiagent systems, and socially responsible engineering. He is particularly interested in building extremely flexible, self-configurable, and adaptable, wireless networks for communication systems that are able to adapt to stringent and conflicting objectives in terms of user satisfaction, configurability, and durability.

PAUL J. M. HAVINGA received the Ph.D. degree on the thesis titled "Mobile Multimedia Systems" in 2000. He is a Full Professor with the University of Twente, The Netherlands, and a Principal Scientist with TNO. His research themes have focused on wireless sensor networks, Internet of Things, sensor data analytics, and energy-efficient wireless communication. This research has resulted in over 700 scientific publications in journals and conferences, and eight patents. He has a significant experience as a Project Manager in international research projects. In 2001, he initiated the first European project on wireless sensor networks. Many research projects on sensor networks followed, all addressing different aspects of wireless sensor networks and IoT. He is the Co-Founder of several high-tech companies resulting from his research.