Federated Learning Empowered Ultra-Dense Next-Generation Wireless Networks

Mohammad Mahmoud Younes Al-Quraan

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

James Watt School of Engineering
College of Science and Engineering
University of Glasgow

February 2024

Abstract

The evolution of wireless networks, from first-generation (1G) to fifth-generation (5G), has facilitated real-time services and intelligent applications powered by artificial intelligence (AI) and machine learning (ML). Nevertheless, prospective applications like autonomous driving and haptic communications necessitate the exploration of beyond fifth-generation (B5G) and sixth-generation (6G) networks, leveraging millimeter-wave (mmWave) and terahertz (THz) technologies. However, these high-frequency bands experience significant atmospheric attenuation, resulting in high signal propagation loss, which necessitates a fundamental reconfiguration of network architectures and paves the way for the emergence of ultra-dense networks (UDNs). Equipped with massive multiple-input multiple-output (mMIMO) and beamforming technologies, UDNs mitigate propagation losses by utilising narrow line-of-sight (LoS) beams to direct radio waves toward specific receiving points, thereby enhancing signal quality. Despite these advancements, UDNs face critical challenges, which include worsened mobility issues in dynamic UDNs due to the susceptibility of LoS links to blockages, data privacy concerns at the network edge when implementing centralised ML training, and power consumption challenges stemming from the deployment of dense small base stations (SBSs) and the integration of cutting-edge techniques like edge learning. In this context, this thesis begins by investigating the prevailing issue of beam blockage in UDNs and introduces novel frameworks to address this emerging challenge. The main theme of the first three contributions is to tackle beam blockages and frequent handovers (HOs) through innovative sensing-aided wireless communications. This approach seeks to enhance the situational awareness of UDNs regarding their surroundings by using a variety of sensors commonly found in urban areas, such as vision and radar sensors. While all these contributions share the common goal of proposing sensing-aided proactive HO (PHO) frameworks that intelligently predict blockage events in advance and performs PHO, each of them presents distinctive framework features, contributing significantly to the improvement of UDN operations. To provide further details, the first contribution adhered to conventional centralised model training, while the other contributions employed federated learning (FL), a decentralised collaborative training approach primarily designed to safeguard data privacy. The utilisation of FL technology offers several advantages, including enhanced

data privacy, scalability, and adaptability. Simulation results from all these frameworks have demonstrated the remarkable performance of the proposed latency-aware frameworks in improving UDNs' reliability, maintaining user connectivity, and delivering high levels of quality of experience (QoE) and throughput when compared to existing reactive HO procedures lacking proactive blockage prediction. The fourth contribution is centred on optimising energy management in UDNs and introduces FedraTrees, a lightweight algorithm that integrates decision tree (DT)-based models into the FL setup. FedraTrees challenges the conventional belief that FL is exclusively suited for Neural Network (NN) models by enabling the incorporation of DT models within the FL context. While FedraTrees offers versatility across various applications, this thesis specifically applies it to energy forecasting tasks with the aim of achieving the energy efficiency requirement of UDNs. Simulation results demonstrate that FedraTrees performs remarkably in predicting short-term energy patterns and surpasses the state-of-the-art long short-term memory (LSTM)-based federated averaging (FedAvg) algorithm in terms of reducing computational and communication resources demands.

Acknowledgements

I want to begin by extending my profound gratitude to the Almighty God for preserving my health and guiding me through the challenges of my PhD journey. I am forever thankful for the inspiration, knowledge, strength, and His grace that allowed me to attain this scholarship.

I want to express my heartfelt gratitude to my primary supervisor, Dr. Lina Mohjazi, who accepted me as her Ph.D. student and provided invaluable mentorship, encouragement, and support throughout my doctoral journey. Her unwavering passion for academic research excellence has profoundly influenced both my academic development and personal growth. Additionally, I extend my immense appreciation to my co-supervisor, Dr. Ahmed Zoha, who has consistently offered strong support for my research endeavor. I am deeply thankful for his enthusiasm, patience, and the valuable insights that have sparked numerous intriguing research ideas and broadened my scholarly horizons.

I would also like to thank Dr. Anthony Centeno and Prof. Muhammad Ali Imran, who have supported me in every way necessary to complete my Ph.D.

I appreciate the University of Glasgow for awarding me the UESTC scholarship, which enabled me to undertake this Ph.D. research. I am also grateful to the Jordan University of Science and Technology, Irbid, Jordan, for granting me a study leave from work to pursue my PhD.

I would like to extend my deepest gratitude to my dear parents, *Mahmoud* and *Amal*, as well as my siblings. Their unwavering love and sacrifices have empowered me to pursue my dream. Their constant support and encouragement have been a continual source of inspiration and motivation, without which I would not have reached this significant milestone in my life.

My deepest appreciation goes to my wife, *Najwa*, and my children, *Alma* and *Ayham*, for their patience, understanding, and support throughout this journey. Their love has been my anchor, providing comfort and motivation in the face of challenges.

I also want to thank my friends and colleagues who provided intellectual stimulation and emotional support during this challenging process.

Thank you to everyone who played a role, no matter how small, in bringing this research to fruition.

To my beloved children, Alma and Ayham,
the lights of my life
and the pillars of my inspiration.

University of Glasgow

College of Science & Engineering

Statement of Originality

Name: Mohammad Mahmoud Younes Alquraan
Registration Number: XXXXXXXX

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

The copyright of this thesis rests with the author. No quotation from it is permitted without full acknowledgement.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Signature: _____

Date: _____

Contents

List of Tables

List of Figures

List of Abbreviations

1G	First Generation
2D	Two Dimension
3D	Three Dimension
3G	Third Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
6G	Sixth Generation
Adam	Adaptive Moment Estimation
ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
AirComp	Over-the-Air Computation
AMI	Advanced Metering Infrastructure
AoA	Angle of Arrival
AoI	Area of Interest
AP	Access Point
AR	Autoregressive
AWGN	Additive White Gaussian Noise
B5G	Beyond Fifth Generation
BS	Base Station
BW	Bandwidth
C-RAN	Cloud Radio Access Network
CDF	Cumulative Distribution Function
CNN	Convolutional Neural Network
CS	Charging Station
CSIT	Channel State Information at the Transmitter
CSI	Channel State Information
CV	Computer Vision
DC	Dual Connectivity
DL	Deep Learning

DNN	Deep Neural Network
DoS	Denial of Service
DPT	Dynamic Positioning Table
DT	Decision Tree
EFB	Exclusive Feature Bundling
eMBB	Enhanced Mobile Broadband
EV	Electric Vehicle
F-RAN	Fog Radio Access Network
F-RGB	Flat Red-Green-Blue
FANET	Flying Ad-Hoc Network
FedAvg	Federated Averaging
FedBoost	Federated Boosting
FedDist	Federated Distance
FedProx	Federated Proximal
FedraTrees	Federated Trees
FedSGD	Federated Stochastic Gradient Descent
FedSplit	Federated Splitting
FedVoting	Federated Voting
FEEL	Federated Edge Learning
FFT	Fast Fourier Transform
FL	Federated Learning
FMCW	Frequency-Modulated Continuous Wave
FoV	Field of View
FTL	Federated Transfer Learning
GBDT	Gradient-Boosted Decision Tree
GD	Gradient Descent
GOSS	Gradient-Based One-Side Sampling
GPS	Global Positioning System
HetNet	Heterogeneous Network
HFL	Hierarchical Federated Learning
HO	Handover
IF	Intermediate Frequency
IIoT	Industrial Internet of Things
IoT	Internet of Things
IoV	Internet of Vehicles
IMT	International Mobile Telecommunications
ITS	Intelligent Transportation System
ITU	International Telecommunication Union

LGBM	Light Gradient Boosting Model
LiDAR	Light Detection and Ranging
LoS	Line of Sight
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MBS	Macro Base Station
MCS	Mobile Crowd Sensing
MEC	Multi-Access Edge Computing
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MLR	Multiple Linear Regression
mMIMO	Massive Multiple-Input Multiple-Output
mMTC	Massive Machine-Type Communication
mmWave	Millimetre Wave
MOS	Mean Opinion Score
MSE	Mean Square Error
NLoS	Non Line of Sight
NN	Neural Network
Non-IID	Non-Independent and Identically Distributed
NP	Non Polynomial
NR	New Radio
ODL	Object Detection and Localisation
OFDM	Orthogonal Frequency Division Multiplexing
PHO	Proactive Handover
PSO	Particle Swarm Optimisation
QoE	Quality of Experience
QoS	Quality of Service
RaDaR	Radar-Aided Dynamic Blockages Recognition
RadMAC	Radar-Based Medium Access
RAN	Radio Access Network
ReLU	Rectified Linear Unit
ResNet	Residual Network
RF	Radio Frequency
RGB	Red-Green-Blue
RIS	Reconfigurable Intelligent Surface
RL	Reinforcement Learning

RNN	Recurrent Neural Network
RSS	Received Signal Strength
RSSI	Received Signal Strength Indicator
RTP	Real-Time Protocol
SBS	Small Base Station
SD	Standard Definition
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Over-Sampling Technique
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
THz	Terahertz
UAV	Unmanned Aerial Vehicle
UCA	Uniform Circular Array
UDN	Ultra-Dense Network
UE	User Equipment
ULA	Uniform Linear Array
URLLC	Ultra-Reliable Low-Latency Communication
VEC	Vehicular Edge Computing
ViWi	Vision Wireless
XGBoost	Extreme Gradient Boosting
XR	Extended Reality

List of Symbols

$(\cdot)[t]$	Function at time instant t
$(\cdot)^T$	Transpose of $(\cdot)$
$(\cdot)^H$	Hermitian transpose of $(\cdot)$
α_p	Channel gain associated with the arrival of path p
β	3D line scalar parameter
Υ	Object movement direction
Δd	Horizontal displacement of a moving user
$\ell(\cdot)$	Pulse-shaping filter
ϵ	Proportion of users who succesfully perfromed PHO
η	Learning rate parameter
$\hat{b}$	Predicted value of blockage
$\hat{b}_{u,o}$	Predicted blockage status for a data sample containing a user u and an object o
$\hat{s}$	Predicted label
$\hat{T}_b$	Predicted remaining time until a user reaches a blocked area
$\hat{T}_{b_{u,o}}$	Predicted remaining time until a user reaches a blocked area for a data sample containing a user u and an object o
$\hat{T}_{D_i}$	Predicted time delay for the sample i
ι	y-component of the direction vector
κ	A single client
Λ	Points for performing PHO
λ	Wavelength
λ_R	Radar wavelength
Λ_{opt}	Optimal point for performing PHO
$\log(\cdot)$	Base-10 logarithm of $(\cdot)$
$\mathbb{C}$	Set of complex numbers
$\mathbb{E}$	Expectation operator
$\mathbb{P}(\cdot)$	Probability of $(\cdot)$
$\mathbb{R}$	Set of real values

$\mathbb{R}^+$	Set of real positive values
$\mathbb{Z}^+$	Set of positive integers
$\mathbf{a}$	Receiver array steering vector
$\mathbf{f}^\star$	Optimal beamforming vector
$\mathbf{f}_i$	i-th beamforming vector
$\mathbf{h}$	Channel between an SBS and a UE
$\mathbf{h}^{\text{LoS}}$	LoS channel between an SBS and a UE
$\mathbf{h}^{\text{NLoS}}$	NLoS channel between an SBS and a UE
$\mathbf{R}$	Samples of radar measurements
$\mathcal{F}$	Beamforming codebook
$\mathcal{K}$	Number of participant clients
$\mathcal{L}$	Mobility/localisation vector
$\mathcal{L}_o$	Mobility/localisation vector of object o
$\mathcal{L}_u$	Mobility/localisation vector of user u
$\mathcal{O}(\cdot)$	Complexity of $(\cdot)$
$\mathcal{P}_{Shift}$	Percent shift parameter
$\mathcal{S}_{PHO}$	PHO success rate
$\max(\cdot)$	Maximum of a function
$\min(\cdot)$	Minimum of a function
μ	Slope of the chirp signal
∇	Gradient operator
ϕ_p	Elevation angle of arrival for path p
$\prod(\cdot)$	Products of elements
ψ	Beamforming steering angle
$\Psi_\Theta(\cdot)$	ML model
ρ	Object range from the radar
$\sim$	Approximately
$\sqrt{\mathcal{E}}$	Transmitter gain
$\sum(\cdot)$	Sum of elements
τ_c	Chirp signal duration
τ_P	Pause time between chirp signals
τ_p	Delay of arrival for path p
τ_{rt}	Round-trip delay of the radar signal
Θ	Set of ML model parameters
θ_o	Angle between positive x-axis and the line passing x_o, y_o and the origin

θ_u	Angle between positive x-axis and the line passing x_u, y_u and the origin
θ_p	Azimuth angle of arrival for path p
υ	Object's velocity
$\varkappa$	x-component of the direction vector
φ	Signal angle of arrival at the radar
ς	z-component of the direction vector
$\vec{d}$	Movement direction
ζ	Average PHO latency per user
A_r	Radar receiver gain
A_t	Radar transmitter gain
B	Number of beams in a beamforming codebook
b	LoS link blockage status
B_c	Chrip signal bandwidth
$b_{u,o}$	Blockage status of the data sample containing a user u and an object o
C	Set of frequency-modulated chirps
c	Speed of light
C_c	Clients count
C_l	Communication limit
D	All data samples
d	Distance between adjacent antenna elements
E_n	Number of epochs
f_c	Chirp signal frequency
f_i	i-th feature of user-object data sample
f_s	Sampling rate
G	Number of trees per batch
H	Height of the radar
h_o	Height of the object
j	Current communication round
K	Number of subcarriers in OFDM modulation scheme
k	k-th element from the subcarrier set K
$L(\cdot)$	Loss function of a global model
$l(\cdot)$	Loss function of a local model
M	Number of antenna elements in ULA
m	Number of data points
M_b	Number of mini-batches

M_r	Number of receive antennas in a radar
n	AWGN with zero mean and σ^2 variance
N_S	Number of samples with successful PHO
N_T	Total number of blocking samples
O	Set of objects
o	An object in a set of objects O
P	Number of paths a signal traverses before arriving at the UE
p	p-th path in the set of paths P
P_r	Received power
Q	Cyclic prefix length in OFDM modulation scheme
q	q-th element in the cyclic prefix length Q
r	Distance between radar and an object
r_0	Initial distance between the radar and an object
r_o	Distance of the object from the SBS
R_s	Random set of clients
S	Samples per chirp after performing the ADC
s_k	Transmitted data symbol at Kth subcarrier
$S_{u,o}$	Data sample containing the mobility/localisation vector of user u and object o
$s_{u,o}$	Label of the data sample $S_{u,o}$
t	An instant of time
T_b	Remaining time until a user reaches a blocked area
T_c	Object classification time
T_D	Delay time
T_F	Framework total execution time
T_m	Radar measurement duration
T_R	The time from radar activation to processing a single measurement
T_s	Sampling time
$T_{b_{u,o}}$	Remaining time until a user reaches a blocked area for a data sample containing a user u and an object o
$T_{D_i}^{max}$	Maximum actual time delay for the sample i
T_{DO}	Delay offset
T_{FFT}	Time associated with performing FFT
T_{Inf}	Model inference time
T_{ODL}	Time required to perform ODL on two successive RGB images
T_{PHO}	Time associated with performing PHO

U	Set of wireless users
u	A user connected to a wireless network
v	Speed of a user
W	Global NN model weights
w	Local NN model weights
W_m	Width of an RGB image in meter scale
W_p	Width of an RGB image in pixel scale
$x(t)$	Represents a variation in the object distance from the radar
x_A	x-coordinate of the antenna array
x_c	x-coordinate of the center point of a boundary box
x_I	x-coordinate of the intersection point
x_o	x-coordinate of the object o
x_u	x-coordinate of the user u
X_{chirp}	Chirp signal mathematical representation
x_{lr}	x-coordinate of the lower right corner of a boundary box
x_{ul}	x-coordinate of the upper left corner of a boundary box
y_A	y-coordinate of the antenna array
y_c	y-coordinate of the center point of a boundary box
y_I	y-coordinate of the intersection point
y_k	Received signal at the UE on the kth subcarrier
y_o	y-coordinate of the object o
y_u	y-coordinate of the user u
Y_{IF}	Mathematical representation of an IF signal
y_{lr}	y-coordinate of the lower right corner of a boundary box
y_{ul}	y-coordinate of the upper left corner of a boundary box
z_A	z-coordinate of the antenna array
z_I	z-coordinate of the intersection point
z_u	z-coordinate of the user u

Chapter 1

Introduction

1.1 Scope and Motivation

Wireless communication systems have undergone a progressive series of improvements, evolving from the first generation (1G) to the current fifth generation (5G) networks. As per the International Telecommunication Union (ITU) and the 3rd generation partnership project (3GPP), 5G wireless networks are designed to deliver enhanced quality of experience (QoE) by offering improved data rate, reliability, capacity, and energy efficiency. In light of this, 5G systems were mapped out based on three fundamental concepts, namely, enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC) [1]. These concepts pave the way for the emergence of novel technological trends as well as the evolution of connected intelligence paradigms, promoting massive scale connectivity [2]. Notably, the number of internet of things (IoT) devices per human was 1.84 in 2010, with a total of 12.5 billion devices, but by 2020, this figure had risen to 6.58 devices per human, reaching nearly 50 billion devices [3]. With the remarkable revolutionary advancements in the field of wireless communications, it is envisioned that these numbers will continue to rise exponentially. This surge is driven not only by technological advancements but also by the imperative to address societal challenges such as emergency response and public safety, interconnected transportation, and healthcare enhancements through remote patient monitoring. Moreover, the introduction of services like extended reality (XR) and massive IoT, and the expected future applications such as holographic communications and multi-sense experience, impose much more stringent requirements beyond what 5G networks can currently provide and highlight the necessity for the next level of network improvements. Consequently, research efforts will shift towards beyond fifth generation (B5G) and sixth generation (6G) communication networks.

Millimetre wave (mmWave) and terahertz (THz) technologies play a crucial role in enabling next-generation wireless networks. Embracing new high-frequency bands is

essential to achieve the global connectivity vision, offering significant improvements in terms of multi-Gbit/s throughput, supporting a massive number of devices, and delivering ultra-low latency and reliable connections [4]. This transition to higher frequency bands changes the paradigm of future wireless networks to small coverage cells, and thus, forming the concept of ultra-dense networks (UDNs) [5].

High-frequency wireless networks are indispensable for meeting the ever-increasing demands for broadband access and unlocking a wide range of revolutionary applications, including intelligent healthcare, holographic telepresence, and autonomous driving. Nevertheless, UDNs are accompanied by certain challenges that must be addressed before their widespread deployment. One of the challenges arises from the need to accommodate a massive number of connected devices, leading to an unprecedented surge in data traffic that requires efficient storage and processing, giving rise to the concept of big data [6]. Artificial intelligence (AI), particularly machine learning (ML), emerges as a cutting-edge technology to leverage big data for delivering pervasive smart services and applications [7]. Traditional ML algorithms perform model training centrally on cloud-based servers, where datasets are collected, transmitted, and stored in a single location, then processed to train ML models using one or multiple servers [8]. While this approach has proven effective with advantages in terms of model convergence, consistency, and transparency, its centralised nature comes with some limitations: (i) **lack of built-in privacy**, raising concerns about security and data privacy. (ii) **Network congestion and elevated energy consumption** due to increased communication overhead between end devices and the cloud. (iii) **Propagation delay** experienced in centralised ML techniques, limiting their scalability and suitability for real-time applications. These limitations necessitate exploring distributed and decentralised training mechanisms, leading to the introduction of federated learning (FL), a collaborative ML algorithm that leverages datasets from distributed entities for local model training without the need to exchange raw data with a central server. In the context of this thesis, FL serves as a pivotal investigative tool, examining its potential to help address specific challenges related to mobility management and energy optimisation within UDNs.

1.1.1 Mobility management

Mobility management presents a critical aspect that poses substantial challenges in UDNs. High-frequency bands operating such networks suffer significant challenges stemming from their electromagnetic properties, including atmospheric attenuation, propagation loss, and susceptibility to blockages. Therefore, multiple-input multiple-output (MIMO) and beamforming techniques are becoming indispensable enablers in next-generation wireless networks. MIMO can compensate for the attenuation and propagation losses using large antenna arrays to generate narrow line of sight (LoS) beams, which improve the quality of

the received signal. Nevertheless, LoS links experience rapid and temporary fluctuations in the received signal strength (RSS) when obstructed by obstacles, especially in dynamic environments. This may lead to frequent handovers (HOs), which negatively impact the network latency, reliability, and energy efficiency. HO is a fundamental mechanism in any wireless network that transfers the ongoing call or data session from one base station (BS) to another. The 3GPP organisation has introduced predefined measurement events; if one occurred, HO must be conducted [9]. Typically, a user equipment (UE)-assisted network for controlling HO receives a measurement report from the user with information about the RSS/quality of a specific downlink reference signal from the serving BS and other neighbouring BSs. Upon meeting specific event criteria, the network will trigger the HO process and negotiations eventuate between the serving BS and the target BS to complete handing the user to the new BS, thus ensuring user connectivity.

The structural downscaling of communication systems has considerably intensified the complexities of addressing mobility-related issues, including beam blockage and frequent handovers [10]. Unlike in third generation (3G) and fourth generation (4G) networks, mobility management in B5G networks is accompanied with negative impacts at both user and network levels. Specifically, frequent HOs lead to increased data transmission delays and throughput degradation at the UE level, as well as higher power consumption and poorer network quality of service (QoS), especially when the target BS rejects the request due to full resource occupation. This problem is exacerbated in smart cities due to the highly dynamic environment and the existence of blocking objects that can shade the serving beam. For instance, a link budget undergoes a 20 dB or more power loss when the connection is blocked by obstacles, such as human bodies or vehicles [11, 12]. Such a sudden drop in the received power affects the signal quality and degrades the data rate of the communication link, making the network unreliable for time-sensitive applications. Thus, it is crucial to comprehensively address the challenge of beam blockage to fully exploit the potential of mmWave and THz bands and to develop reliable wireless communication systems.

1.1.2 Energy management

Energy management is another essential aspect considered in this thesis to aid fulfilling the stringent energy efficiency requirement of next generation wireless networks. The study in [13] indicates that UDNs exhibit higher energy consumption in comparison to their 4G network counterparts. This surge in energy utilisation is primarily attributed to the massive deployment of small base stations (SBSs), a configuration that contributes significantly to the network's overall energy consumption, accounting for as much as 80% of the total energy expenditure. Moreover, deploying edge learning across the extensive number of SBSs and exploring novel research directions involving sensors integration, such

as cameras and radars to aid UDN operations, significantly intensifies energy consumption. Consequently, managing energy usage effectively becomes a paramount challenge in these advanced networks.

1.2 Problem Statement and Objectives

For the efficient deployment of future UDNs and ensuring their readiness in delivering anticipated services to a large number of heterogeneous users, irrespective of time or location, it is essential to proactively tackle potential design intricacies and post-deployment challenges that could arise. For example, with the utilisation of high-frequency multiarray antennas that offer beamforming capabilities, UDNs are expected to enjoy beam-based communications. Nonetheless, the reliance on high-frequency LoS communications renders these systems more susceptible to the adverse impacts of blockage and penetration losses. Beam blockage is a complicated challenge facing high-frequency wireless networks that requires moving beyond mere network detection towards proactive prediction and prevention strategies to ensure uninterrupted functionality. Diverse sensor types, such as cameras, light detection and rangings (LiDARs), radars, ultrasonic sensors, microphones, and global positioning system (GPS), hold the potential to enhance the operational effectiveness of high-frequency communication systems. Hence, this thesis aims to tackle the beam blockage challenge through the convergence of the emerging paradigm of sensing-aided wireless communication and the application of AI techniques. Furthermore, the introduction of new components into network operations and the implementation of edge learning contribute to elevated energy consumption. This thesis also intends to address this aspect by proposing an innovative solution that aids in predicting and managing energy consumption in UDNs.

Towards achieving highly reliable, efficient, and seamlessly performing UDNs, this thesis adopts a systematic approach, focusing on the examination of one framework at a time. Specifically, the thesis begins with considering the beam blockage in vision-aided UDNs and investigates the feasibility of adopting centralised learning techniques in such problems. This deliberate starting point with a centralised approach establishes a foundational step and offers thorough insights for subsequent studies, particularly with more advanced approaches such as distributed learning in later contributions. Subsequently, the investigation transitions towards the application of FL to resolve the same problem, but within a more complex service environment. Next, the study explores the benefits derived from the integration of radar sensors as an alternative to vision sensors for the beam blockage problem, while also leveraging the FL mechanism. Lastly, the thesis delves into methods aimed at enhancing energy management strategies to optimise energy efficiency.

1.2.1 Problem Statement

Beam Blockage Prediction in Sensing-aided UDNs

The issue of beam blockage is examined across three distinct scenarios, each designed to reflect various practical environmental conditions.

Scenario 1) A single moving user with a stationary blocking object in vision-aided UDNs. High-frequency wireless networks, exemplified by UDNs, are developed to serve highly dynamic and ever-changing environments. Within such settings, LoS beams are susceptible to obstruction by diverse objects, ranging from immobile structures like buildings to mobile entities such as humans and vehicles. In the event of a blockage, the network reactively detects it, but its options are confined to attempting user reconnection via beam switching or HO. However, the reactive nature of blockage detection is unfeasible and imposes numerous drawbacks on the network and users, including resource consumption and poor QoE. As a result, there is a need to proactively predict blockages and transition from reactiveness to proactiveness when detecting blockages. Such proactive measures can be achieved using camera sensors capable of gathering environmental data, thereby augmenting network awareness. While the collected visual data holds great potential, careful processing is essential to extract the required information for training a specially designed neural network (NN) model. Moreover, this scenario delves into the feasibility of adopting a centralised learning methodology to execute the model training process.

Scenario 2) Multiple moving users and dynamic/stationary blocking objects in vision-aided UDNs. In this scenario, the beam blockage problem resembles that of the first scenario, but with consideration of more complex environment characterised by the presence of multiple mobile users and many stationary or moving blocking objects. Here, the processing of visual data is more complex, necessitating the classification of each object within the red-green-blue (RGB) images, differentiation between wireless users and blocking objects, and the identification of the target user amidst other wireless users depicted in the captured images. Moreover, due to the large size and volume of visual data, transmitting this data to a centralised server for processing and training would lead to significant depletion of network resources and raises privacy concerns. Consequently, there is a need to move from conventional centralised learning paradigms to distributed learning methodologies, specifically employing the FL approach.

Scenario 3) Stationary users and dynamic blocking objects in radar-aided UDNs. Similar to the first and second scenarios, the third scenario considers the beam blockage

problem in high-frequency wireless networks. However, the incorporation of vision sensors introduces limitations that may hinder achieving an optimal solution for blockage prediction and prevention. The deployment of vision sensors might not always be viable due to regulatory and privacy constraints. Additionally, image quality can be compromised under conditions of low-light and adverse weather. Hence, an exploration of alternative sensor modalities that can provide greater enhancements to the beam blockage framework becomes imperative. In light of their numerous advantages and inherent capacity to address the limitations of vision sensors, radar sensors are considered in this scenario to aid the network's decision-making process. Moreover, this scenario also examines the use of FL approach, offering noteworthy advantages. Alongside the core attribute of FL in maintaining data privacy, which holds relatively less significance in the context of radar data due to its inherently reduced privacy risk, FL yields three additional pivotal benefits: scalability, knowledge sharing, and resource efficiency.

Optimisisng Power Consumption in UDNs

UDNs have emerged as a prominent network paradigm to fulfil the stringent requirements of international mobile telecommunications (IMT)-2020 for 5G networks [14]. In qualitative terms, UDN signifies a network characterised by considerably enhanced radio resource density compared to prevailing networks. This, in turn, yields a notably more compact configuration for small cell networks, wherein inter-site distances between SBSs are reduced to a mere few meters [15]. The extensive deployment of SBSs and their operational activities are inherently tied to an increase in energy consumption, posing challenges in realising the envisioned energy efficiency objectives of future wireless networks. Moreover, the new research concept of sensing-aided wireless communications, coupled with the integration of diverse sensor types at the network edge, stands to amplify power consumption at the network level. In addition, the execution of FL mechanism at the network edge contributes to an elevated overall energy consumption due to the communication between FL server and clients, alongside the computational intensity inherent in the mathematical operations of ML models. As such, there is a need to optimise the network's power consumption by developing novel lightweight energy forecasting frameworks that can help predict future power needs and effectively manage power resources.

1.2.2 Objectives

In light of the previously mentioned technical problems, the aim of this thesis is outlined within the following objectives:

O1 : Conduct an in-depth exploration of the mobility and energy management approaches

within UDNs. The discussion will explore diverse strategies addressing the challenge of beam blockage, as well as frameworks predicting network energy consumption for optimised network performance. Additionally, conduct a comprehensive review of the FL concept, covering its architecture, categories, and operational principles. Furthermore, provide an elaboration into the implementation of FL in state-of-the-art wireless applications, highlighting its role in enhancing data privacy, network efficiency, and collaborative learning across distributed entities within the wireless domain.

O2 : Explore a new research direction known as vision-aided wireless communication empowered with AI to tackle beam blockage problems inherent in high-frequency wireless networks. Computer vision (CV) algorithms enable real-time environmental perception, anticipating and navigating potential blockages caused by obstacles. Moreover, the integration of AI facilitates adaptive learning for continuous improvement in predicting and responding to blockage scenarios. Simultaneously, investigate the feasibility of using both centralised and decentralised training methodologies to develop vision-aided proactive blockage prediction frameworks. Centralised models leverage cloud-based servers for unified processing, while decentralised methods distribute the training process across local entities, promoting adaptability to diverse environments.

O3 : Examine the advantages of integrating radar sensors instead of visual sensors to tackle the beam blockage problem within UDNs. Assess the distinct benefits of radar technology, such as enhanced accuracy in obstacle detection, operability in low-light conditions, and resilience to adverse weather. Additionally, investigate the potential of the FL approach in facilitating knowledge sharing among distributed entities, enhancing scalability, and optimising resource efficiency. The adaptive nature of FL aligns seamlessly with the dynamic conditions of UDNs, providing an intelligent and self-improving mechanism for proactive blockage prediction. This exploration aims to foster the development of radar-aided proactive blockage prediction frameworks.

O4 : Develop a new FL aggregation algorithm tailored to integrate decision tree (DT)-based models within the FL framework. Then, explore its effectiveness in smart energy application, assessing its potential to meet the energy-efficient requirements of UDNs. Additionally, conduct a comparative analysis against modern energy forecasting frameworks, evaluating both computational and communication efficiency. This research objective aims to advance energy management strategies within UDNs through the development and assessment of the proposed FL aggregation algorithm.

1.3 Contributions and Research Outcome

Based on the objectives outlined earlier, this research aims to propose innovative AI-driven frameworks tailored to optimise the mobility and energy management facets of next-generation wireless systems, thereby ensuring their seamless service provisioning. This section highlights the main contributions and research outcomes of this thesis.

1.3.1 Contributions

The major contributions of this thesis are summarised and itemised as follows:

- A comprehensive conceptual foundation of the FL algorithm's operational principles is presented, encompassing a thorough elucidation of its architecture, categories, operation, and optimisation strategies. Furthermore, an in-depth exploration of the critical factors motivating the deployment of FL in modern wireless applications is undertaken, considering relevant performance metrics and ongoing research efforts. The outcome of this contribution is published in [J2].

- An innovative approach to addressing beam blockage and frequent HOs in next-generation wireless networks is proposed through the integration of CV and NN algorithms. The incorporation of CV enhances network awareness of the surroundings, while the NN model predicts occurrences of blockages due to stationary obstacles. Moreover, a new HO event, termed *BLOCK*, is introduced as an extension to the standardised events defined by 3GPP. This event detects the presence of obstacles and users moving towards blocked areas, making it applicable in B5G and 6G networks. This contribution also includes an analysis to determine the optimal HO triggering point to maintain a high level of user QoE once the BLOCK event is detected. The outcome of this contribution is published in [J1].

- An advanced CV-aided blockage prediction for multi-user/objects UDNs is formulated. This involves the development of a comprehensive end-to-end latency-aware framework that leverages visual data to proactively predict blockages and perform proactive handover (PHO), aiming to maximise users' QoE. In addition, an FL training mechanism is adopted as a distributed learning approach rather than the conventional centralised learning method to train the model locally in each SBS where the visual information resides. This strategy secures data privacy, reduces communication overhead, and ensures secure data handling. The outcome of this contribution is published in [C4].

- A novel radar-aided dynamic blockages recognition (RaDaR) framework is developed to improve the reliability of federated UDNs by integrating radars for the anticipation

of LoS link blockages while considering latency and QoE metrics. RaDaR hinges upon the utilisation of FL algorithm to perform collaborative model training at each SBS by using information acquired from radar placed at the top of the SBS. The FL approach provides the framework with vital features, including scalability, knowledge sharing, and conserving network resources. The efficacy of the RaDaR framework is evaluated using a large-scale real-world dataset comprising mmWave channel information and radar data, collectively ensuring the framework's validation and effectiveness within real-world scenarios. The outcome of this contribution is currently under review for publication, as indicated in [J5].

- A novel lightweight aggregation algorithm, named federated trees (FedraTrees), is proposed with the purpose of optimising energy consumption in UDNs. FedraTrees leverages DT-based models within the FL setup to enhance performance metrics, particularly in energy forecasting scenarios, by reducing required communication rounds and computation time. The FedraTrees approach incorporates a monitoring mechanism that observes the progress of the FL training process and halts it when further improvements are no longer attainable. Comparative results against the commonly employed long short-term memory (LSTM)-based federated averaging (FedAvg) technique demonstrate a significant enhancement in overall performance through the implementation of the FedraTrees framework. The outcome of this contribution is published in [J4].

1.3.2 Research Outcome

The outcomes of this thesis, along with collaborations with fellow research students, have resulted in the following publications:

Journals

[J1] **M. Al-Quraan**, A. Khan, L. Mohjazi, A. Centeno, A. Zoha and M. A. Imran, 'Intelligent beam blockage prediction for seamless connectivity in vision-aided next-generation wireless networks', IEEE Transactions on Network and Service Management, vol. 20, no. 2, pp. 1937–1948, Oct. 2022. DOI: 10.1109/TNSM.2022.3216556.

[J2] **M. Al-Quraan**, L. Mohjazi, L. Bariah, A. Centeno, A. Zoha, K. Arshad, K. Assaleh, S. Muhaidat, M. Debbah and M. A. Imran, 'Edge-native intelligence for 6G communications driven by federated learning: A survey of trends and challenges', IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 7, no. 3, pp. 957–979, Mar. 2023. DOI: 10.1109/TETCI.2023.3251404.

[J3] S. Hafeez, A. Khan, **M. Al-Quraan**, L. Mohjazi, A. Zoha, M. A. Imran and Y. Sun, 'Blockchain-assisted UAV communication systems: A comprehensive survey', IEEE Open Journal of Vehicular Technology, vol. 4, pp. 558-580, July 2023. DOI: 10.1109/OJVT.2023.3295208.

[J4] **M. Al-Quraan**, A. Khan, A. Centeno, A. Zoha, M. A. Imran and L. Mohjazi, 'FedraTrees: A novel computation-communication efficient federated learning framework investigated in smart grids', Engineering Applications of Artificial Intelligence, vol. 124, p. 106654, Sept. 2023.

[J5] **M. Al-Quraan**, A. Zoha, A. Centeno, H. Bany Salameh, S. Muhaidat, M. A. Imran and L. Mohjazi, 'Enhancing reliability in federated mmWave networks: A practical and scalable solution using radar-aided dynamic blockage recognition', submitted to IEEE Transactions on Mobile Computing (Major revision).

[J6] A. Khan, A. Jabbar, I. Ahmad, **M. Al-Quraan**, L. Mohjazi, M. Rehman, S. Hussain, M. A. Imran, A. Zoha 'Proactive Blockage Prediction for UAV-assisted Handover in Future Wireless Network', submitted to Mobile Networks and Applications (Major revision).

[J7] Y. Chen, G. Shi, **M. Al-Quraan**, Y. Sambo, O. Onireti and M. A. Imran, 'LoRa Mesh-5G integrated network for trackside smart weather monitoring', submitted to IEEE Transactions on Vehicular Technology (Under review).

[J8] I. Mefgouda, H. Idoudi, **M. Al-Quraan**, O. Waqar, A. Zoha, M. A. Imran, and L. Mohjazi, 'A QoS-Aware Service-Driven Network Selection for HWNs based on MARCOS and Utility Functions', submitted to IEEE Open Journal of the Communications Society (Under review).

Conference Proceedings

[C1] **M. Al-Quraan**, A. Khan, L. Mohjazi, A. Centeno, A. Zoha and M. A. Imran, 'A hybrid data manipulation approach for energy and latency-fficient vision-aided UDNs', in proc. IEEE Eighth International Conference on Software Defined Systems. (SDS), Gandia, Spain, Dec. 2021, pp. 1–7. DOI: 10.1109/SDS54264.2021.9732115.

[C2] H. U. Manzoor, A. Khan, **M. Al-Quraan**, L. Mohjazi, A. Taha, H. Abbas, S. Hussain, M. A. Imran and A. Zoha, 'Energy management in an agile workspace using AI-driven forecasting and anomaly detection', in proc. 4th Global Power, Energy and Communications Conference (GPECOM), Nevsehir, Turkey, Mar. 2022, pp. 644-649.DOI: 10.1109/GPECOM55404.2022.9815599.

[C3] S. Hafeez, M. A. Shawky, **M. Al-Quraan**, L. Mohjazi, M. A. Imran and Y. Sun, 'BETA-UAV: Blockchain-based efficient and trusted authentication for UAV communication', in proc. IEEE 22nd International Conference on Communication Technology (ICCT), Nanjing, China, Nov. 2022, pp. 613–617. DOI: 10.1109/ICCT56141.2022.10072981.

[C4] **M. Al-Quraan**, A. Centeno, A. Zoha, M. A. Imran and L. Mohjazi, 'Federated learning for reliable mmWave systems: Vision-aided dynamic blockages prediction', in proc. IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, UK, Mar. 2023, pp. 1–6. DOI: 10.1109/WCNC55385.2023.10118675.

1.4 Thesis Outline

The rest of this thesis is structured as follows. Chapter 2 evaluates the methodologies discussed in the literature for mitigating mobility and energy management challenges within high-frequency wireless networks and offers an in-depth overview of the FL concept. Chapter 3 proposes a novel framework that employs vision-aided wireless communication and centralised ML approaches to effectively address the beam blockage problem in UDNs. Building upon this, Chapter 4 introduces a more advanced framework that addresses the beam blockage problem within UDNs by considering complex scenarios involving multiple users and blockages. This is accomplished through the application of collaborative FL training mechanisms. In Chapter 5, a pioneering approach that combines radar technology and the FL methodology is presented. This framework is designed to effectively address beam blockage challenges in real-world scenarios within UDNs. Chapter 6 is dedicated to proposing an energy-efficient framework that is tailored to fulfill the energy efficiency requirements of next-generation wireless networks. Finally, Chapter 7 concludes the dissertation and discusses potential future research extensions.

Chapter 2

Literature Review

In this chapter, the complexities of mobility management within high-frequency communications systems are introduced, with a specific spotlight on the beam blockage problem. Various adopted approaches aimed at mitigating this issue are presented. Afterwards, the focus shifts towards discussing the energy management issue, which holds the potential to impact the energy efficiency of next-generation wireless networks. It also delves into the techniques adopted in the literature for predicting energy consumption, thereby facilitating the formulation of enhanced energy management strategies. Finally, a brief exploration of the fundamental aspects of FL is presented. This includes an examination of its architectural underpinnings, categorical classifications, operational mechanisms, and strategies employed for model aggregation. Then, it delves into an extensive exposition of the diverse applications of FL within wireless communication systems. This comprehensive review aligns with the objectives outlined in *O1*.

2.1 Mobility Management

Despite the numerous benefits gained when shifting the operational frequencies of next-generation communication systems from lower bands (sub-6GHz) to the higher bands (mmWave and THz), reliance on high-frequency bands introduces critical mobility challenges, such as link blockage and frequent HOs. To this end, many attempts have been made to provide solutions to address the connectivity issue in high-frequency wireless networks. This section presents a comprehensive review of state-of-the-art studies that have proposed solutions to mitigate the issue of beam blockage in high-frequency communication systems. These works can be categorised into two main strategies in terms of data acquisition for training the ML models: (i) wireless information-based approaches and (ii) sensing information-based approaches. Table 2.1 briefly outlines the key techniques, including their advantages and limitations.

2.1.1 Wireless Information-based Approaches

These studies are further classified into three distinct groups: those designed for indoor environments [16–19], those designed for outdoor environments [20–22], and those that are suitable for use in both indoor and outdoor settings [23, 24]. In order to predict the quality of multi-directional links in mmWave systems, the paper [16] proposed utilising an LSTM network to capture the long-term signal-to-noise ratio (SNR) dependencies across mmWave links and predict the fluctuations in received power for one or more links in the next time instant. Accurate anticipation of link quality can facilitate the communication system in predicting link failure and choosing alternative robust channels to maintain system connectivity. Nevertheless, the study did not allocate adequate time for resource optimisation to address changes in link quality. The contribution in [17] was characterised by the development of a recurrent neural network (RNN) predictor trained according to the meta-learning concept and capable of predicting mmWave link blockages in indoor industrial environments. Shared initialisation parameters are optimised based on SNR observations from distinct deployments to decrease the necessary dataset size required for any new deployment. Although meta-learning has shown promise in training models to adapt to a variety of tasks, the risk of overfitting poses a significant challenge for generalisation to unseen data.

The study described in [18] explored an online learning-based method for predicting beam blockages in indoor hotspot scenarios using deep learning (DL). This approach includes an offline learning phase, where a deep neural network (DNN) model is trained on a fingerprinting database containing user positions, data traffic demands, and blockage statuses. During the online prediction phase, the model predicts blockage statuses based on the user's location and traffic requirements. However, this approach has limited scalability, as it is designed for a specific scenario. Additionally, it fails to proactively predict blockages, leading to service interruptions and frequent HO issues. The proposed work in [19] presented an online blockage detection and beam recovery algorithm for mmWave indoor networks in Industry 4.0. The method uses a DNN model trained solely on beam measurement reports to predict future beam states and the length of blockage events, enabling the system to switch to alternative beams in advance, thereby reducing handover delay and enabling a stable link quality with minimal reduction in data transmission rate. However, the proposed framework has not been evaluated for speeds greater than 2 m/s, which limits its applicability in real-world scenarios with faster-moving obstacles.

Moving to the outdoor methodologies, the contribution in [20] relied on controlling the mmWave propagation channel by adopting reconfigurable intelligent surface (RIS) technology to avoid blockages and facilitate beam management. By integrating environmental awareness and users' motion information into a DNN model, the network can effectively select the optimal RIS among those deployed around the base station,

maintaining link connectivity and optimising beam management while reducing network operation overhead. One challenge associated with this approach is the increased complexity in network design and operation, particularly with regard to coordinating multiple reflective surfaces within a single network. The authors in [21] utilised the uniform circular array (UCA) geometry for beamformers at both mobile and base stations to take advantage of the two-dimensional scanning capability required for the proposed beam recovery scheme for standalone mmWave networks. They propose the use of an LSTM model trained on previous optimal beamforming and combining vectors obtained from blockage observations to predict future blockages and identify the best beam directions for the next time frame. However, such techniques may be inefficient in practical scenarios as they assume a highly scattering environment and do not consider the link quality of alternative non line of sight (NLoS) links. The study in [22] categorised blockage types that affect mmWave links' stability into three classes: self, static, and dynamic blockages. Furthermore, the study proposed the use of convolutional neural network (CNN) with Kalman filters to estimate channel state information (CSI) and reduce the interference caused by the deployment of many base stations. The blockages are addressed through different approaches, including optimal handover, connecting the user to the nearest BS, and utilising drones as a BS while taking into account the CSI. However, implementing these techniques increases network complexity, and the study lacks a clear strategy for identifying blockages.

Based on dual band network operation, the authors in [23] exploited the sub-6GHz uplink channel information to improve the reliability of mmWave downlink channels. The decision is motivated by the spatial correlation between the two frequency bands. A DL model is trained to determine if the LoS link is blocked or not by using a tuple of sub-6GHz channel information and blocking status. However, this approach may not be practical in real-world scenarios as it only classifies the channel status as blocked or unblocked and cannot avoid mmWave link blockages. The contribution in [24] relied on utilising the diffraction properties of signal propagation to provide an early warning of potential obstructions in the hybrid sub-6 GHz and mmWave communication system. Diffraction causes a significant amount of jitter in the received signal power just before a blockage occurs. Furthermore, it has been observed that the lower the frequency, the earlier the jitter sample appears. As a result, this study employs the diffraction characteristics of the signals in sub-6GHz uplink channels to predict blockages by training an LSTM model. However, incorporating both frequency bands increase the complexity and cost of the communication system.

2.1.2 Sensing Information-based Approaches

Similarly, while some of the sensing-aided works presented here are tailored for indoor networks [25], the majority of them are designed for outdoor systems [26–30]. Notably, the emerging research direction of exploiting CV for developing wireless communications and tackling complicated problems in UDNs has gained much interest recently. For example, the studies presented in [25] and [26] both harnessed LiDAR sensory information to enhance the reliability of high-frequency networks. This was achieved by establishing an understanding of the communication environment and proactively predicting potential link blockages. In the methodology outlined by [25], two ML techniques were employed: a three dimensional (3D) convolutional network and a gradient-boosting regression tree, to capture the spatiotemporal features of the acquired time series point clouds. The experimental outcomes reveal that the proposed approach can anticipate substantial attenuation in link quality up to 1000 ms ahead. Conversely, the approach described in [26] involved the collection of LiDAR point cloud sequencies, followed by the removal of static clusters. The data was subsequently transformed into heat maps, which were then employed to train a CNN architecture for link quality prediction. However, while LiDAR sensors have showcased their effectiveness, certain limitations exist. Their coverage area is constrained, and real-world scenarios often involve the fact that not every object within the environment results in a blockage. The act of blocking a link is intricately tied to factors like the locations/heights of users and base stations, as well as the dimensions of the detected objects. Hence, the practical applicability of these studies may possess limitations.

Moving to camera-based approaches, the authors in [27] proposed a framework for supporting blockage prediction in mmWave networks using the concept of semantic communication. The framework extracts environment semantics from visual information and serves two purposes: first, it protects user privacy and reduces system overhead by retaining only the class and layout information of the data; and second, it aids in channel-related downstream tasks such as blockage prediction. The developed network architecture consists of an environment semantics extraction network, a feature selection technique, a task-oriented encoder, and a decision network to predict the blockage state. However, semantic communication relies on accurately understanding the visual scene and extracting relevant information. Complex scenes with multiple objects and dynamic changes can make scene understanding challenging, leading to errors in predicting beam blockages. In [28] and [29], camera imagery and DL were leveraged to tackle the beam blockage problem in mmWave systems. The proposed techniques predict the time series of the mmWave received power to several hundred milliseconds in advance based on the depth images of the served area, allowing for sufficient time to perform HO. However, predicting the received power in advance does not necessarily suit HO decision problems. Furthermore, it requires large quantities of training datasets and computational resources

to prepare a model for accurate prediction. The authors in [30] utilised the visual sensory information collected from the served area to train a DL model and predict beam blockages. RGB images captured by a camera installed on each BS are labelled with the beam blockage status and used to fine-tune the ResNet18 model to classify images based on blockage status. However, the proposed technique does not predict in advance, so service disconnection cannot be avoided.

The reviewed approaches in the domain of sensing-aided solutions for wireless communications in UDNs showcase both notable benefits and drawbacks. The integration of CV in addressing challenges within UDNs stands out as a promising avenue of research. For instance, studies leveraging LiDAR sensors ([25], [26]) demonstrate enhanced reliability by proactively predicting link blockages. Nevertheless, these approaches face limitations, such as restricted coverage area and the challenge of precisly determining the objects causing blockages. Meanwhile, camera-based methods ([27], [28], [29], [30]) offer an alternative, employing CV and semantic communication to protect user privacy and aiding in blockage prediction. However, challenges arise in accurately understanding complex visual scenes, leading to potential errors in blockage prediction. While some techniques predict mmWave received power in advance, ensuring time for HO, they require extensive datasets and computational resources. Notably, certain camera-based approaches do not predict in advance, posing challenges in avoiding service disconnection. This critical review highlights the advancements made by these approaches while identifying gaps related to coverage limitations, object-specific blockages, scene complexity, and the trade-offs between prediction accuracy and real-time applicability.

Table 2.1: Existing research about the challenge of beam blockage in high-frequency networks.

Ref.	Date	Algorithm	Indoor or Outdoor?	Advantages	Limitations
[28]	Aug. 2019	DL model trained on visual information.	Outdoor	CV is employed to foster an awareness of the surrounding environment.	The need for large quantities of training datasets and computational resources.
[30]	May 2020	DL model trained using visual and wireless information.	Outdoor	CV is employed to foster an awareness of the surrounding environment.	Latency-unaware and privacy concerns.
[16]	June 2020	Multi-directional links quality prediction using an LSTM model.	Indoor	Predcting the quality of several links within a single or multiple cells.	Insufficient time for optimising resources to mitigate potential link failures.
[23]	Sept. 2020	DL model trained using sub-6GHz and mmWave information.	Both	Reduce computational burden by leveraging transfer learning.	Increased system complexity, reactive HO.
[20]	Oct. 2020	DL-based blockage avoidance in IRS-assisted mmWave networks.	Outdoor	High efficient beam management with reduced frequent handover.	Increased network design and operation complexity.
[21]	June 2021	UCA beamformers to predict LoS link blockages.	Outdoor	Maintaining communication sessions while avoiding the need to reset beam scanning procedures.	Assuming a highly scattering environment, not considering the QoS of alternative NLoS links.
[29]	June 2021	CNN and RNN models trained using visual and wireless information.	Outdoor	CV is employed to foster an awareness of the surrounding environment.	Privacy concerns, No target user identification.
[22]	Sept. 2021	CNN and Kalman filter to estimate CSI.	Outdoor	Several techniques to handle blockages based on their respective types.	Increased network complexity.
[17]	Dec. 2021	Meta-learning based RNN for blockage prediction.	Indoor	Less data is required to train ML models.	Risk of overfitting, challenge for generalisation to unseen data.
[18]	Jan. 2022	DNN model trained using user location, data demands and blockage statuses.	Indoor	The model exhibits a prediction accuracy exceeding 80%.	Limited system scalability, inability to prevent service outages, frequent handovers.
[26]	Apr. 2022	CNN-based blockage prediction using LiDAR point cloud.	Outdoor	Predicting blockages occurring within 100 ms with a precision of 95%.	Limited to a very specific scenario, limited coverage area.
[24]	June 2022	LSTM-based blockage prediction using signal diffraction characterestics.	Both	Generalisation from outdoor to indoor scenarios.	An increase in both system complexity and cost.
[19]	June 2022	DNN model trained on beam measurement reports.	Indoor	Rely exclusively on the mmWave channel measurements.	Evaluations based on limited obstacles speed.
[27]	Jan. 2023	Channel-oriented semantic communications.	Outdoor	Enhanced mmWave system reliability.	No target user identification.
[25]	Jan. 2023	CNN and GBRT models trained using LiDAR information.	Indoor	Anticipate attenuation in link quality up to 1000 ms ahead.	Limited coverage area, not every object means a blockage.

2.2 Energy Management

Energy efficiency in UDNs has been explored in the literature through diverse network-related techniques, encompassing infrastructure optimisation [31], spectrum and resource management [32], and innovative approaches like network slicing [33]. Complementing these methodologies, this thesis aims to enhance energy efficiency by proposing a lightweight energy forecasting framework. By accurately predicting future power needs, the proposed framework enables proactive and strategic allocation of energy resources, contributing to a more efficient and sustainable operation of the network. Importantly, the existing literature lacks studies incorporating energy forecasting to achieve energy efficiency in UDNs. Consequently, and taking into account that energy forecasting requires load profiles regardless of the domain or area under investigation, this thesis examines energy forecasting approaches within the context of smart grids. Subsequently, the thesis examines local and central load forecasting methodologies before delving into cutting-edge research leveraging FL frameworks for load forecasting.

2.2.1 Local and Centralised Approaches

Load profiles store energy/power consumption information in time-series data, which can be predicted using several approaches, such as statistical and computational intelligence methods. Statistical methods have been used in the literature for the short and medium-term forecasting ranges, demonstrating favorable performance. For instance, the study in [34] used the multiple linear regression (MLR) technique to verify its reliability in forecasting energy demand, whereas the traditional autoregressive (AR) methodology was used to predict electrical energy in [35]. Moreover, the studies in [36–38] benefited from the combination of AR and moving average (MA) to improve the forecasting process. Later, the focus on ML methods became dominant owing to the advantages of AI in analysing substantial data volume. The contribution in [39] highlighted the use of the support vector machine (SVM) to forecast the monthly electrical load of Taiwan. On the other hand, numerous studies exploited NNs in the energy forecasting domain by virtue of their impressive performance in various domains [40, 41]. RNN is also widely used in smart energy, particularly LSTM and its variants [42, 43].

Apart from NNs, DTs have also been used in the energy forecasting task. While the DT approach is simple, it exhibits favorable performance when predicting future energy consumption [44, 45]. However, DTs alone have not seen widespread adoption due to several limitations, including instability, susceptibility to overfitting, and performing poorly with noisy and nonlinear data. Later, ensemble-based algorithms gained popularity and were explored in the energy research domain. In general, ensemble techniques possess attractive features that draw the research community's attention, such as simplicity,

usability, interpretability, and computational efficiency. For instance, the forecasting performance of random forest was compared with NN in [46], demonstrating the feasibility and effectiveness of both approaches in building energy applications. Gradient-boosted decision tree (GBDT) and extreme gradient boosting (XGBoost) algorithms were harnessed to predict future electricity loads and demonstrated their effecacy in [47] and [48], respectively.

The proposed studies and frameworks mentioned above are generally based on a centralised model training, where energy consumption information is transmitted across the network and combined in a central location. However, this training scheme raises privacy concerns. Load profiles hold sensitive information that can be used in various dimensions like inferring operation patterns of networks, unauthorised data use, and data selling, among others.

2.2.2 FL-based Load Forecasting

Few studies have begun to consider employing FL in the context of smart energy. For example, [49] demonstrated the potential of using federated settings to predict electrical load consumption patterns, assist load monitoring, and enable energy demand response. The FedAvg technique was applied to aggregate LSTM models parameters and produce a generalised global model. Similarly, FedAvg and LSTM were adopted in [50] to provide a generalised electrical load forecasting model. The authors incorporated complementary features related to calendar and weather conditions, in addition to sequences of previous electrical loads, to improve the forecast model.

The study described in [51] investigated the significance of employing smart meters in residential areas for short-term forecasting tasks using LSTM-based FL. Similarly, the contribution in [52] evaluated the performance of FL versus centralised and local training methods when using LSTM models for electrical load forecasting. The study concluded that local learning is better suited to predict individual energy consumption than FL. However, FL is necessary when a generalised forecasting model is required and access to aggregated data is impossible. Very recently, the research in [53] adopted the LSTM algorithm under the FL setting to forecast energy profiles. Two strategies are considered: federated stochastic gradient descent (FedSGD) and FedAvg to perform parameters aggregation. Experimental results demonstrated that FedAvg achieves better accuracy and requires fewer communication rounds.

The aforementioned FL-based energy forecasting studies rely on DL algorithms, specifically LSTM networks. Although LSTMs have been shown to achieve excellent prediction accuracy, they require intensive processing duties that yield a heavy computational burden.

2.3 FL Background

2.3.1 FL Fundamentals

The concept of FL has attracted significant attention in academia and industry [54]. The key principle of FL is to construct a generalised global model by performing distributed model training. The recent advancement in edge devices' communication and computation capabilities and the large amount of data generated and stored locally on the devices facilitate the spread of this emerging technology widely. This section presents the fundamentals, architecture, categories, operation principles, and aggregation schemes of FL algorithms.

FL Architecture

Based on the nature of the network, the architecture of FL can be categorised into classical and hierarchical federated learning (HFL). The classical FL approach consists of two main parts: the server and the participating clients [55], as illustrated in Fig. 2.1(a). At the beginning of the FL process, the server initiates the training procedure by sharing a new or pretrained model with the participating clients. After that, the clients will personalise the received model by training it based on their local data, and then share their local models with the server for aggregation and global model update. On the other hand, HFL framework [56], depicted in Fig. 2.1(b), optimally fits in heterogeneous networks that include different cell coverage. This architecture is introduced to alleviate the bandwidth (BW) overhead at the FL servers, resulting from the large number of model updates communicated from the clients. Furthermore, HFL can reduce the communication latency experienced between the clients and the server by reducing the link distance. The HFL framework consists of two stages; in the first one, the clients send and receive the model parameters by communicating with a server located at the SBS, i.e., the edge server, and the server performs local model aggregation. Meanwhile, in the second stage, the edge servers send the aggregated models to a central server that can be located at the macro base station (MBS) or in the cloud, in which the server performs edge model aggregation for global model update and sends it back to the edge servers.

It is noteworthy that the FL server must have certain specifications to orchestrate the FL process efficiently. These specifications are drawn from the considered ML technique and the number of clients. For instance, training a DL model through many clients requires a high server capacity, a huge computation capability, high-speed interfaces, and locating the server in close proximity to the clients. On the other hand, the specifications may be less stringent when considering simpler models of NNs and a few clients.

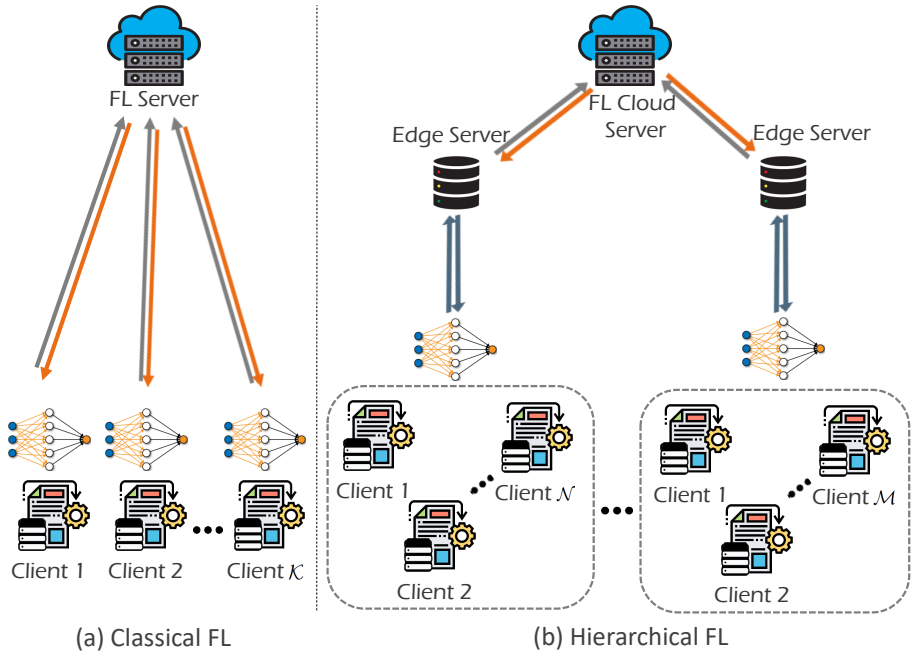


Figure 2.1: Types of FL architecture (a) Classical FL in client-server architecture (b) HFL in client-edge-server architecture.

FL Categories

Given the significant role of local datasets in realising efficient training and assuming the data is structured in a two dimension (2D) matrix format where rows correspond to data samples and columns represent features, FL systems can be classified based on the data distribution characteristics between different parties into horizontal, vertical, and federated transfer learning (FTL) [57].

1) *Horizontal* FL*:* This is the most common category of FL, also known as sample-based FL. The distinctive characteristic of this category lies in the fact that various parties' datasets share an identical feature space while diverging in the sample space. For example, consider two regional educational institutions with comparable interests in monitoring research outcomes, which constitutes the feature space. However, these institutions maintain distinct research groups, reflecting the sample space. This category facilitates the adoption of a unified ML model with the same architecture for all datasets. Therefore, the global model can be derived through the averaging of all local updates. FedAvg technique [58] is an example of this type of FL system.

2) *Vertical* FL*:* This category, referred to as feature-based FL, comes into play when multiple datasets share the same sample space, yet their feature spaces differ. For instance, consider two different parties within the same city: a healthcare institute and an e-commerce company that records customers' purchasing behaviors. Their user sets likely comprise local residents, creating a shared sample space. The objective here is to exploit the diverse features from these two parties to construct a model predicting residents'

future health statuses based on their buying patterns. When implementing vertical FL, the participating parties may be curious to know each other's data. In such cases, a trusted third-party coordinator can ensure data confidentiality throughout the training process. Alternatively, if a certain level of trust exists between the participating parties, the necessity for a third party can be obviated, allowing one of the parties to take on the role of coordinator.

3) FTL*:* When the dataset of different clients exhibits partial overlap in both feature and sample spaces, FTL, also known as hybrid learning, emerges as the optimal choice. FTL facilitates the transference of knowledge from one domain to another, enhancing the learning outcomes. More specifically, a locally trained model from one party is conveyed to another party, enabling the utilisation of insights gleaned from non-overlapping regions to enhance the model training at the other party. The most common example of transfer learning is evident in image classification. Various models are tailored for classifying specific datasets, and with minor tuning, they can be repurposed to classify other dataset types.

FL Operation

FL protocol consists of three main phases [59] detailed as the following:

1) *Clients selection*: While large-scale deployment stands as an appealing facet of FL, in contrast to classical ML, the number of participating clients in model training can swiftly escalate to thousands or even millions of devices. This enormous number of endpoints reflects the capacity enhancements envisaged within B5G (1 million/km^2) and 6G networks (100/m^3). As a result, end-device onboard capabilities and data distribution will vary considerably among the participants, rendering client selection a critical design aspect in FL. To address this challenge, various methods have been proposed, as exemplified by [60]. In this work, the authors propose a technique that improves the time-to-accuracy training performance by guiding FL developers in selecting participants even at the scale of millions of clients.

2) *Configuration*: In this phase, the selected participants receive the initial model parameters and proceed to train their local models using their respective datasets. Specifically, upon the selection of participating devices, $\mathcal{K}$ edge nodes are ready to begin the training process. Fig. 2.2 illustrates the FL's architecture and the operation steps. Each device, denoted as $\kappa \in 1, 2, ..., \mathcal{K}$, has a local dataset, $D_\kappa \in D_1, D_2, ..., D_\mathcal{K}$ consisting of input-output pairs of samples (x_i, y_i), where $x_i, y_i \in \mathbb{R}$. In step ①, the FL server initiatialises the global model designed for a specific task and shares it with the choosen participants. Next, at the t-th iteration, each participating node obtains the model weights W_{t-1} and begins the model training by exploiting data samples stored locally. The training objective is to minimise the loss function $L_\kappa(W_t^\kappa)$ across all data samples in the training

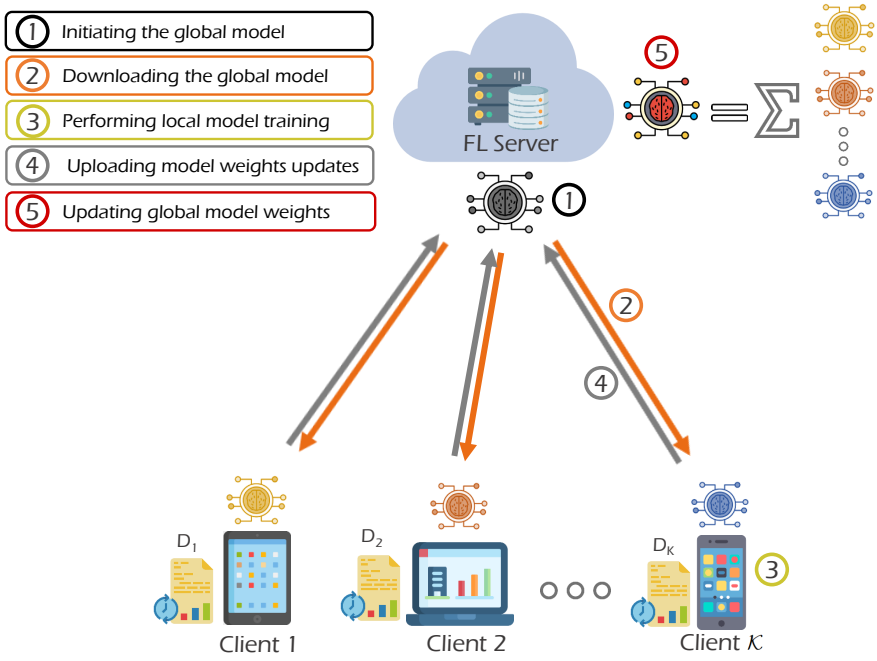


Figure 2.2: Sequential operation steps of FL involving $\mathcal{K}$ participants.

dataset. Mathematically, $L_\kappa(W_t^\kappa) = \frac{1}{D_\kappa}\sum_{i\in D_\kappa} l_i(W_t^\kappa)$, with the aim of determining the optimal model parameters W_t^κ that minimise the loss function in each round of training. This can be represented as, $\arg\min_{W_t^\kappa\in\mathbb{R}} L_\kappa(W_t^\kappa)$. Here $l_i(W_t^\kappa)$ signifies the loss associated with data sample i given the parametrisation W_t^κ, (steps ②, ③).

3) *Reporting*: At this point, participants share the local model updates with the central server, either synchronously or asynchronously [61]. The synchronous strategy yields a highly precise and rapidly converging model in the absence of stragglers, which are instances arising from inadequate hardware or network resources that can hinder the training process. Stragglers threaten the scalability of FL by slowing down the training process. Conversely, the asynchronous mechanism naturally handles stragglers by incorporating participants' updates as they become available. However, it comes with the risk of compromising model quality and insecure aggregation, resulting in a suboptimal level of privacy. Several works have tackled this concern by suggesting different schemes, such as implementing a secure buffer as seen in FedBuff [62], introducing staleness-awareness as in FedSA [63], and employing semi-asynchronous techniques like SAFA [64]. Finally, the server aggregates the shared parameters to update the global model. This involves performing model aggregation and computation of global model parameters at the server, executed as the follows: $W_t = \sum_{\kappa=1}^{\mathcal{K}} \frac{D_\kappa}{D} W_t^\kappa$, where D represents the entire dataset of all clients, i.e., $D = \sum_{\kappa=1}^{\mathcal{K}} D_\kappa$, (steps ④, ⑤). The steps from ② to ⑤ are repeated until the global model converges to a desired accuracy.

FL Aggregation Schemes

Gradient descent (GD) is an algorithm commonly used in various ML techniques, particularly NN models [15], to locate the minimum of a differentiable function. However, as the dataset size grows, the computational complexity of GD increases, rendering it unsuitable for FL systems due to its slow convergence rate. Stochastic gradient descent (SGD) provides an alternative to GD, allowing gradient calculations to be performed over subsets of data, significantly accelerating the convergence rate. In the FL setting, FedSGD is exploited as a method to quantify how frequently the global FL model needs to be updated [58]. Nevertheless, the FedSGD technique mandates numerous communication rounds, proportionate to the volume of nodes' datasets, which can impose strain on communication links and consume BW.

To address the above problem, the FedAvg strategy has been introduced as a means to alleviate the strain on communication resources [58]. FedAvg is a generalisation of FedSGD, where each node repeatedly conducts SGD on different local data subsets, subsequently determining optimal model parameters through an average of locally computed gradients. The performance of FedAvg hinges on three key parameters: (i) the fraction of selected nodes performing computations in each round, (ii) the size of data subsets, and (iii) the number of epochs a node traverses through its dataset during each round. In FedAvg, as opposed to transmitting computed gradients, each node solely transmits model parameters. Consequently, in contrast to FedSGD, the FedAvg algorithm involves more local computation and less communication with the server.

However, in real-world scenarios characterised by network devices of varying capabilities and non-identically distributed local datasets, FedAvg encounters challenges in terms of convergence behavior. Therefore, some variants of the FedAvg algorithm have been introduced to develop faster aggregation techniques. One such variant is federated proximal (FedProx), which was proposed to solve the heterogeneity issue in federated networks [65]. The FedProx principle resembles that of FedAvg, albeit with a small critical modification that improves performance. Instead of forcing every node to perform the same computation work, FedProx accommodates the system's heterogeneity by allowing each node to undertake a level of local computation proportional to its resources. Accordingly, enabling parameter aggregation across a diverse set of nodes with varying capabilities.

Another extension to the FedAvg scheme is the federated splitting (FedSplit) algorithm [66], which relies on the operator splitting procedure for convex optimisation problems. Operator splitting is an efficient method for solving large-scale convex problems by performing iterations of simple and computationally inexpensive operations. This technique transforms the primary problem into more manageable sub-problems, advancing their resolution separately. Motivated by the inability of FedAvg and FedProx to maintain the fixed points of the original optimisation problem, FedSplit introduces

itself as a splitting algorithm for federated optimisation, designed to achieve rapid convergence. Moreover, the work in [67] applies adaptive optimisers—namely ADAGRAD, ADAM, and YOGI—in the FL context, yielding FedAdaGrad, FedAdam, and FedYogi. Thorough experimental assessments compare these algorithms to the FedAvg approach. Furthermore, the Qsparse-local-SGD algorithm [68] tackles both local computation and communication reduction with distributed settings. Convergence analysis is undertaken for both synchronous and asynchronous FL scenarios, showing that the Qsparse-local-SGD algorithm achieves the same convergence rate as FedSGD.

The previously outlined approaches are primarily tailored for NN models, wherein the central elements for global model updates are the parameters—weights and biases. Despite numerous endeavors to refine the aggregation process, NN and DL models still incur substantial communication and computation costs. Therefore, a surge of investigations has been directed towards exploring alternative low-complexity techniques within the FL framework. Notably, ensemble learning strategies have gained traction in this regard, exemplified by federated boosting (FedBoost) [69] and FedraTrees [70]. It has been demonstrated in [69] and [70] that when the federated model is trained according to these algorithms, excellent performance is achieved in terms of accuracy, computation time, and communication rounds. This paves the way for exploring other ML techniques in the FL environment.

2.3.2 FL Applications in Wireless Networks

Since the inception of FL by Google in 2016, a substantial body of research has been dedicated to advancing, refining, and identifying optimal applications for this decentralised learning algorithm. Wireless networks are one of the forerunners to adopt FL in their architectures, as depicted in Fig. 2.3. This section will thoroughly present the key driving applications of FL within a range of diverse wireless networks.

Cellular Networks

The rollout of 5G in late 2020 has enabled operators to launch a plethora of commercial services leveraging the enhanced capabilities offered by this new technology [71]. In addition, the incorporation of FL in these networks has led to diverse applications across various domains, as detailed below.

a) *Homogeneous Cellular Networks:* This category refers to low-frequency wireless networks with macrocells, denoting their wide coverage. Two main concerns for FL at the network edge are heterogeneous devices with different computation and communication capabilities and securing local model updates. The work in [72] presented a blockchain-enabled FL framework to ensure security in a trustless environment. This approach utilises

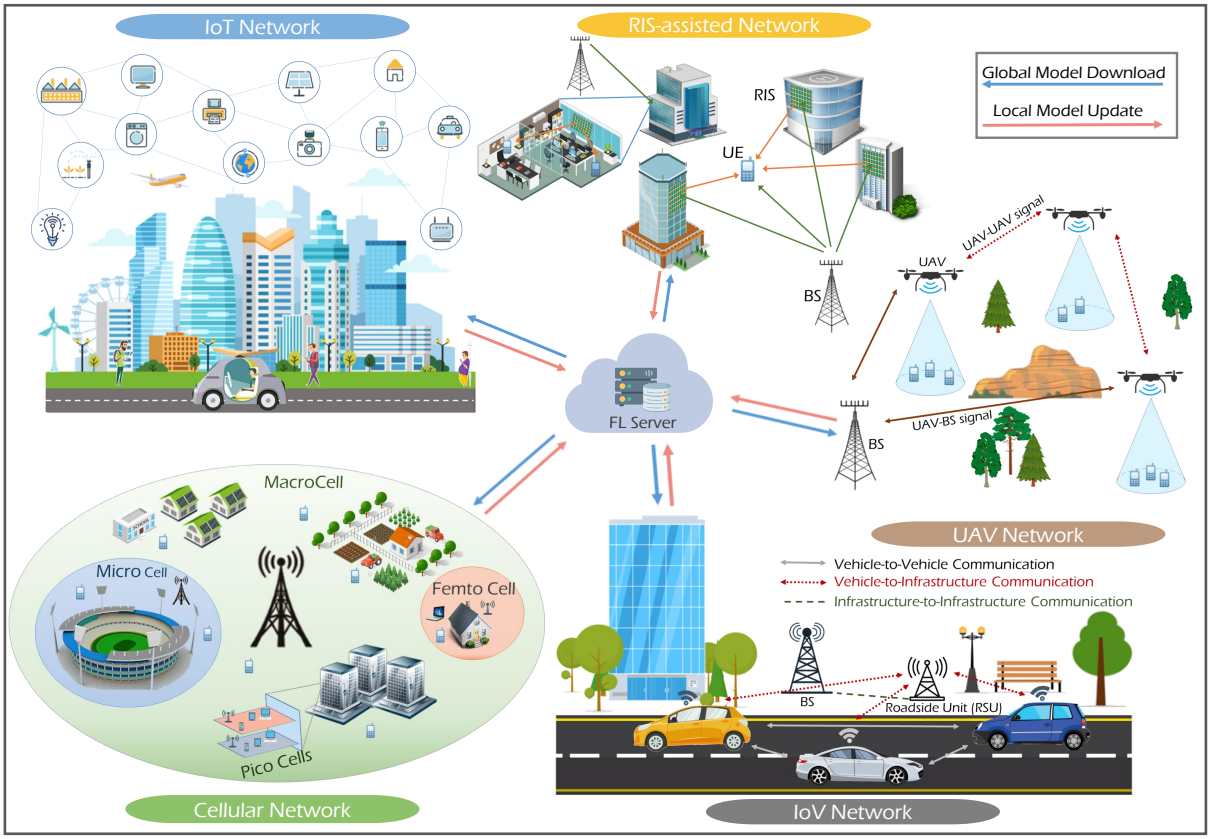


Figure 2.3: FL in various wireless networks; FL algorithm in the context of single or multiple wireless networks.

a distributed ledger between entities, positioning blockchain as an intermediary between the FL server and edge nodes to verify model parameters through a consensus process. Also, FL finds applications within network function virtualisation, an innovative concept that enables adaptive resource allocation for future wireless networks. The study in [73] leveraged the FL technique to build a model capable of proactively predicting auto-scaling settings for multi-access edge computing (MEC) virtual services. This approach ensures compliance with data protection policies while optimising resource allocation.

b) *Heterogeneous Cellular Networks:* Heterogeneous networks (HetNets), characterised by a variety of cell types, expand wireless networks' coverage and capacity. FL can be implemented in HetNets for resource allocation purposes. It was demonstrated in [56] that applying HFL by grouping the users and assigning the needed resources for transmission can reduce the end-to-end communication latency in HetNets. This optimisation is realised by clustering users and assigning each cluster to the closest SBS. On the other hand, 5G HetNets are vulnerable to attacks, like denial of service (DoS), evil twinning, and port scanning. The work in [74] proposed a node-edge-cloud framework empowered by HFL to detect attacks throughout the 5G HetNets. Specifically, multiple dedicated nodes are distributed inside the network, each of which performs model training by employing

reinforcement learning (RL) technique to facilitate adaptive learning that can capture the rapidly changing nature of the HetNets environment. In addition, [75] proposed an FL-empowered MEC framework to tackle the communication overhead and delay between the edge server and clients in FL to enhance the training efficiency.

c) MIMO*:* FL has many applications intertwined with MIMO technology. Given the dynamic nature of mmWave systems, the study in [76] enhanced the performance of massive multiple-input multiple-output (mMIMO) systems by estimating channel state information. FL is leveraged to conduct decentralised learning on the user side, utilising local pilot signals to predict channel matrix. This prediction aids in optimising beamforming design, consequently augmenting the system's overall performance. Moreover, the work in [77] presented an energy-efficient solution that supports multiple FL groups in future wireless systems. The approach involves employing mMIMO to facilitate model updates and ensure stable operation of various FL processes concurrently executed within the same coherence time.

d) *Fog-Cloud radio access network (*RAN*):* The ever-increasing number of connected devices in 5G and beyond networks has compelled a shift towards an ultra-efficient air interface. This transition has fostered the development of two distinct air interface structures: cloud radio access network (C-RAN) and fog radio access network (F-RAN). A notable observation from the literature survey is the prevalence of FL applications in F-RAN networks, whereas the utilisation of FL in C-RAN networks remains limited. For example, the study in [78] optimised the latency and BW resources when deploying FL in RIS-aided C-RAN systems. The RIS controls channel propagation conditions and supports over-the-air computation (AirComp) technique for coherent on-air aggregation of local models. This is achieved by allowing clients to transmit simultaneously to the parameter server. On the other hand, the F-RAN paradigm fully uses edge networks and endows crucial features such as content caching for optimal application performance and user experience. In similar vein, the authors of [79] proposed an FL-based mobility-aware content-caching framework in F-RAN-based networks. Mobility and content demand statistics are exploited to improve users' QoE by predicting and caching the most likely future content.

e) 5G-*new radio (*NR*):* 5G-NR stands as a new radio interface standard designed by 3GPP to satisfy the growing demands of 5G mobile networks. This pioneering radio access technology allows user equipment to dynamically switch between different resource blocks, each characterised by distinct BWs. Nonetheless, this technique raises resource allocation challenges in B5G networks. FL has many applications in resource allocation in terms of computation, communication, and energy efficiency. For example, [80] employed FL to formulate an ML model that facilitates distributed resource management across cellular networks, simultaneously reducing transmit power during uplink transmission.

Internet of Vehicular Networks

Internet of vehicles (IoV) has recently emerged as a key enabler for intelligent transportation systems (ITSs), combining two key concepts, namely, vehicle networking and AI [81]. Within this context, the IoV paradigm strives to attain intelligent information exchange between vehicles and all network entities. This objective is accomplished by harnessing vehicle computation capabilities, which leverage DL algorithms, cloud and edge computing, as well as big data analytics, to materialise vehicular intelligence.

a) FL *in* ITS*:* Communication reliability and latency are of particular significance in the area of ITS, owing to the severe consequences that might affect human safety. The work proposed in [82] capitalised on the integration of FL with blockchain technology to realise a distributed, privacy-aware, and efficient model designed for autonomous vehicular networks. The heterogeneous nature of the vehicles in ITSs is particularly appealing for FL applications. The varied data generated contributes to enhanced model accuracy by encompassing all network scenarios encountered by different vehicles. However, in addition to addressing latency, FL has shortcomings in terms of server centralisation. The process of exchanging large updates between participants and the server imposes a high overhead on the server itself. To overcome this challenge, [82] employed the blockchain technique. This approach entails a distributed ledger shared with each vehicle, thereby upholding copies of both the global and private models. This strategy, verified by each vehicle, alleviates the strain imposed on the central server. Similarly, [83] studied the use of FL setting within the context of URLLC in vehicular networks. The primary focus revolves around proposing a distributed framework for joint transmit power and resource allocation. This framework aims to reduce the power consumption of vehicular users while ensuring the facilitation of low-latency communications.

b) *Vehicular edge computing (*VEC*):* In line with the MEC concept, VEC exploits the communication and computation capabilities at the network edge. The authors in [84] implemented FL with VEC to perform image classification to support diverse applications in ITSs. A model-selective strategy is proposed to identify clients with the highest computational capabilities and select models that give the best image quality for aggregation. In an asymmetric FL setting, the server has no information about clients' data and resources. To this end, a two-dimensional contract mechanism was proposed in [84], in which the server designs contract bundles encompassing diverse attributes such as data quality, computation capability, and rewards. Clients, in turn, opt for bundles that optimise their utility. As part of IoV networks, electric vehicle (EV) networks are becoming more popular as the number of EVs increases; such networks are expected to take over from traditional vehicles in the coming years. In this context, [85] studied energy efficiency and profit maximisation at charging stations (CSs). It introduces an FL-driven economically efficient framework that investigates historical energy transactions to

increase CSs profit. Concretely, FL is used to train localised models using CSs' private data, predicting the energy demands of EVs. After that, these localised models from various CSs are aggregated and shared, facilitating information exchange and yielding more accurate results.

c) *Traffic Prediction:* Traffic prediction in smart cities brings up many benefits for ITSs, such as road safety, congestion avoidance, and shortest route selection. These gains are pronounced when exploiting edge-derived information in conjunction with FL. A noteworthy avenue for enhancing FL involves the selection of optimal hyperparameters for local models within edge devices. While a significant portion of existing literature has been directed toward FL's global optimisation, privacy concerns, and communication efficiencies, relatively few have delved into the optimisation of model parameters. The study in [86] introduced a technique rooted in particle swarm optimisation (PSO) for optimising local hyperparameters at the edge devices. Specifically, PSO optimises the local NN parameters, including the number of layers, neurons per layer, and epochs. The efficacy of this optimisation method is showcased within the domain of traffic prediction. The findings underscore a marked reduction in the number of client-server communication rounds required to find the optimal parameters. This technique is attractive due to its implementation's low complexity. However, it should be noted that its drawback lies in its reliance on random searches for the best initial parameters. This aspect entails an unpredictable time requirement, potentially impacting the whole learning process.

UAV Networks

The flying vehicles in unmanned aerial vehicle (UAV) networks have several appealing attributes, such as low cost, flexible mobility, and ease of deployment. These qualities enable them to engage in tasks that were previously considered challenging. The application of AI algorithms and the recent advancements in UAV technology have greatly broadened the scope of use cases for UAV networks [87].

a) AI-*Empowered* UAV: The interplay between AI and UAV networks opens new horizons for exploiting UAVs in more complicated tasks; however, data security and privacy remain a significant challenge. In UAV-enabled mobile crowd sensing (MCS) applications, FL is particularly appealing in maintaining the privacy of sensed data. In this regard, the authors in [88] integrated an FL-based UAV network with blockchain technology to eliminate the need for a central server. Moreover, blockchain enhances FL network security by excluding adversarial clients and facilitating secure model updates exchange among clients. On the other hand, the work in [89] proposed an FL-enabled framework for air quality monitoring in MCS, ensuring security. A swarm of UAVs is deployed to measure the air quality, with the gathered data utilised to train a lightweight model for predicting the air quality index. FL emerges as a promising candidate capable

of capitalising on the data silos collected by different agencies, thereby generating a global model while safeguarding data privacy.

Following the MEC concept, federated edge learning (FEEL) can potentially reduce the end-to-end latency and communication overhead in UAV networks. Yet, as demonstrated in [90], the efficient implementation of FEEL in UAV-based IoT networks is restrained by the battery lifespan of UAVs. In this respect, [42] has formulated optimisation of computation resources and bandwidth allocation to enhance FEEL performance within UAV networks. Also, [91] had harnessed FL as a supporting technique to reduce the communication cost between multiple UAVs and a ground fusion center. This application takes place within the context of image classification for remote area exploration missions.

b) *Flying ad-hoc networks (*FANETs*):* With the interest of accomplishing complex tasks in UAV networks, UAVs are ad-hoc grouped to establish a local network, facilitating cooperative efforts to tackle joint tasks. Recent advancements in trajectory design and remote monitoring predominantly rely on ML algorithms [92]. To recall, such classical algorithms do not align well with the context of UAV networks due to their high mobility and limited energy resources. FL is proposed to reduce the communication overhead as an efficient paradigm for FANETs. In this setup, all participating UAVs collaborate to estimate the initial model parameters. Then, initial model parameters from all UAVs are shared and leveraged for local model training. To facilitate model aggregation and foster the development of an improved global model, a FEEL server is deployed to capitalise on the potential of local models.

Attributed to the inherent decentralised nature of FANETs, these networks are vulnerable to several security threats that intend to disrupt their functionality, such as impersonation and jamming attacks [93]. Centralised attack detection and mitigation strategies are unfeasible due to the highly dynamic topology of FANETs. Thus, decentralised approaches become imperative for these network types. To this end, in [94], an FL-based device jamming detection for UAVs in FANETs was proposed. In addition to enhanced security, the framework outlined in [94] addressed the issue of data heterogeneity among different UAVs. In particular, a Dempster-Shafer technique classifies UAV clients into groups based on their data quality. Then, the FEEL server selects high-quality data group(s) for model training, contributing to both security enhancement and data quality considerations.

RIS-Assisted Networks

The emergence of numerous mmWave and THz applications has raised several concerns attributed to the vulnerability of such applications to signal blockage and shadowing effects. Motivated by this and with the recent advancements in the solid-state industry, RISs have emerged as enablers of future wireless networks [95]. An RIS, composed of

several reflective elements, can be artificially engineered to control the electromagnetic properties of wireless signals and enable diverse functionalities, including wave splitting, reflection, and absorption. Leveraging an RIS is particularly beneficial in AirComp-enabled FL scenarios, in which some clients may experience blockage or weak channel conditions, affecting the quality of global model training [96–98]. AirComp is a technique that exploits the superposition nature of the wireless channel to transmit simultaneous model updates from multiple clients. The authors in [96] used the AirComp technique assisted by RIS to expedite global model aggregation, which reduces the required radio spectrum for parameter transmission since the clients collectively send their updates using the same channel. Additionally, to further enhance and boost the quality of global model aggregation, an RIS is used to reduce aggregation errors by strengthening the quality of combined signals. In this respect, aiming to unlock the full potential of RIS in FL settings, the work in [97] formulated a joint communication and learning optimisation problem, considering device selection, transceiver design, and RIS parameters.

The aforementioned contributions have assumed perfect CSI at the server and clients' sides. However, acquiring channel state information at the transmitter (CSIT) is not always attainable due to dynamic channel conditions, leading to a significant delay in receiving the CSI information. This delay curbs the convergence of the FEEL global model. The proposed work in [98] investigated the CSIT-free over-the-air model aggregation based on RIS-assisted FEEL. The CSI at the transmitter side is assumed to be unavailable, while perfect CSI is assumed at the server side. Besides, the RIS adjusts and aligns the channel coefficients with the model aggregation weights. To this end, the successive channel coefficients are constrained as a function of RIS phase shifts, making them proportional to the weights of the local models. Moreover, the received scaling factor is optimised by minimising the aggregation mean square error. To solve this optimisation problem, a difference-of-convex algorithm is adopted. Furthermore, the RIS has proven its efficiency in converting wireless channels into a smart electromagnetic environment. To realise high-speed RIS-based communication, the authors in [99] proposed two FL-based RIS optimisation schemes: RIS-assisted outdoor and indoor IoT mmWave communications. In the former scenario, the RIS controller acts as the FL server, while the UE functions as a client. The clients' data represents the CSI corresponding to their location and the optimal RIS configuration. The trained model aims to optimise the achievable rate to enable high-speed mmWave communications. The latter scenario considers an access point (AP) connected to multiple IoT devices assisted by RISs, where the AP acts as an FL server, while the RIS and IoT devices act as clients. The FL model is trained based on location information and optimal RIS configuration. As a result, the trained FL model can achieve high transmission sum rates in IoT networks.

IoT Networks

High-dimensional data analytics will shift traditional IoT paradigms from simply connected things to connected intelligence. It is envisaged that FL will become an indispensable tool in intelligent IoT-based applications, which are spreading across diverse fields [100, 101]. This section outlines the usage of FL in various sectors associated with IoT networks.

a) *Industrial internet of things (*IIoT*):* The fourth industrial revolution, often referred to as Industry 4.0, is triggered by advancements in automation and manufacturing industries, coinciding with the emergence of IIoT devices. Despite the promising features of FL that benefit IIoT networks, the upsurge number of nodes participating in the training process may generate significant traffic that burdens the network. Implementing reliable participant selection schemes can reduce network overhead and alleviate communication costs. The work in [102] presented a budgeted client selection algorithm that enhances the global model accuracy by identifying the best clients. This algorithm finds random set of clients with the best test accuracy based on the secretary problem. More specifically, clients are interviewed sequentially, marked as selected or rejected, and then ranked from best to worst to streamline the selection process. Another critical design aspect in FL-empowered IIoT networks revolves around potential edge device failures, which can lead to severe fluctuations in production quality. Addressing this concern, the authors in [103] shed light on such aspects and proposed an anomaly detection framework that uses FL to train edge devices to predict abnormalities, thereby enabling enhanced communication efficiency.

b) *Healthcare Applications:* FL has become very popular in the field of healthcare applications [104]. Pandemics exert detrimental effects on human health and also cast shadows on the economy. A recent example is the global Covid-19 pandemic, which has resulted in health issues and fatalities. Covid-19 primarily manifests as pneumonia, detectable through X-ray scans. ML can play a pivotal role in such medical scenarios, utilising collected data to train an ML model capable of predicting infection states. Given the sensitivity of patient data across various medical institutions, the FL approach naturally emerges as a suitable solution for these applications. Therefore, the study in [105] applied FL on datasets of various clinical centers. FL clients exploit the local X-ray images of Covid-19 cases from each hospital to train a model aiding practitioners in determining a patient's infection status, all while preserving the privacy of personal information.

c) *Financial Perspective:* The financial sector plays a central role in every society. In particular, the reliance on credit cards has exponentially increased in recent years, facilitating day-to-day life. Security attacks constitute a major threat to credit card systems, resulting in critical information leakage and financial losses. At present, banks independently utilise their datasets to develop centralised ML algorithms for fraud

detection to mitigate such threats. However, this approach has proven ineffective due to the insufficiency of the datasets in creating accurate models. To overcome this challenge, the work in [106] presented a framework that leverages FL to build a collaborative fraud detection system by using datasets from multiple banks. A noteworthy challenge in this domain is the scarcity of fraudulent transactions in comparison to legitimate ones, which can hinder the performance of FL. To this end, the synthetic minority over-sampling technique (SMOTE) is used to oversample the minority class, generating synthetic datasets that bolster the training of the FL model and lead to improved model inference.

2.4 Summary

This chapter commenced with a thorough review of techniques proposed to tackle the blockage issue in UDNs. This spanned from traditional methods to cutting-edge, innovative-driven sensing-aided approaches. Subsequently, the discussion extended to cover the energy management issue in UDNs, involving the best energy forecasting solutions provided in the literature, whether based on centralised or decentralised approaches. Furthermore, the chapter delved into an introduction to the FL technique, explaining its core principles. This encompassed a detailed exploration of its architecture, diverse categorisations, operational mechanisms, and the commonly used aggregation algorithms. Following this, a comprehensive discussion unfolded, shedding light on the versatile applications of FL across a spectrum of wireless networks. The knowledge offered in this chapter is strategically positioned to equip readers with a comprehensive foundation, setting the stage for the subsequent chapters to seamlessly comprehend the novel contributions presented within this thesis.

Chapter 3

Proactive Beam Blockage Prediction in Vision-aided UDNs

In this chapter, the attention is focused on enhancing the reliability of UDNs by ensuring uninterrupted communication links. The primary objective, aligned with *O2*, is to anticipate potential blockages that pose a high risk of obstructing LoS links and disrupting ongoing communications. This improvement is facilitated through vision-aided wireless communication, where visual information captured by vision sensors equips UDNs with awareness of the surrounding environment. The incorporation of CV and ML plays a crucial role in predicting the status and timing of forthcoming blockages. This intelligent approach empowers the network to proactively identify and avoid blockages through performing PHO at an optimal point so that maintaining the QoE for a user at high levels.

3.1 Introduction

Next-generation wireless networks undergo a substantial design change when operating in high-frequency bands. The quest for high data rate services using mmWave and THz technologies demands a downscaled communication system, giving rise to a new network paradigm known as UDNs [5]. Moreover, the use of beamforming enhances the RSS by creating LoS beams. Nevertheless, UDNs encounter critical challenges due to the sensitivity of high-frequency beams to blockages. These signals experience high penetration loss and attenuation, leading to a sharp drop in RSS whenever an obstacle intercepts the LoS communication link.

In the literature, several techniques have been adopted to overcome the connectivity issue. For instance, researchers had examined the geometry of mmWave channels and the diffraction characteristics of signals in comparison to sub-6GHz frequencies to predict potential blockages in mmWave LoS connections [107, 108]. Other solutions relied on

ML and dual connectivity (DC) to maintain wireless communication and meet the required QoE for users [109, 110]. However, these solutions exhibit varying limitations in practicality, often leading to the inefficient use of network resources. Most importantly, they fail to completely mitigate the issue of link blockages, as the switching between links remains reactive in nature.

To effectively solve this problem, UDNs require a sense of the surrounding environment to move from reactive to proactive blockage measures. The direct view is essential for UDN communications and is equally important to CV, where visual information captures only direct and visible objects in the scene, enabling proactive detection of obstructing objects. Therefore, leveraging vision information collected from the served environment is envisioned to aid the operation of the network, moving beyond reliance solely on wireless information, which fails to address this problem [107–110]. Images are rich in detail that can effectively contribute to solving the blockage problem in UDNs. For example, in [28], depth images and a DL model were used to predict a user's RSS in the next few hundred milliseconds to assist in HO decisions. Further, [30] exploited RGB images to train a ResNet-18 model and then classify the images based on the blockage status. However, the approaches detailed in [28] and [30] fail to account for the associated latency until a successful HO is completed, thus rendering them unable to prevent link blockages—a critical concern in highly dynamic UDNs.

The previously mentioned studies have different assumptions that limit their applicability in real practical scenarios. The vision for B5G and 6G networks centers around meeting stringent requirements for sustaining high levels of QoS and QoE, ensuring consistent user connectivity to realise real-time services and applications. In response, a newly emerging research direction focuses on harnessing CV to enhance the performance of mmWave communication systems is envisioned to assist the operation of such systems, satisfy the stringent demands, and encourage their widespread implementation. Furthermore, the proliferation of next-generation high-frequency wireless networks is anticipated within smart cities, where video surveillance systems are widely available—particularly in densely populated areas, the main target of deploying UDNs. Given that future BSs will be miniaturised and installed on lampposts, the combination of wireless and vision information is recognised as an attractive approach to optimise the operation of high-frequency wireless networks [30,111]. Nevertheless, the integration of CV into the operation of UDNs is still in its early stages and necessitates dedicated research endeavors to extract its full potential. Hence, this study plays a leading role in this research domain, distinguishing itself from existing works in the literature by proactively predicting blockages and executing optimal PHOs. The study gives comprehensive attention to the framework's end-to-end latency, a critical factor ensuring effective PHO and seamless user transition to another stable connection.

3.2 Contributions

This chapter presents a novel CV-assisted PHO mechanism that integrates two modes of information: wireless and imagery information. The goal is to predict possible beam blockages in advance, enabling the network to perform HOs at a time that maximises the overall QoE and thus optimises network performance. An object detection and localisation (ODL) algorithm is employed to analyse RGB images captured by vision sensors. This algorithm detects obstacles and users, as well as determines their respective locations and speeds. Additionally, a simple NN model is trained using the multivariate regression method to predict the remaining time until a user becomes blocked by an obstacle. This study introduces a new HO event referred to as "blocking event" (BLOCK). This event is characterised by the presence of a blocking object and a user moving towards the blocked area. Once a BLOCK event is identified, the proposed framework calculates the optimal timing to trigger HO, seamlessly transitioning the user to another BS and thereby improving network reliability. the primary contributions and results of this work are outlined as follows:

- A novel solution to the problem of beam blockage and frequent HO in next-generation wireless networks is proposed through the utilisation of CV and NN algorithms. The application of CV enhances the network's understanding of its surroundings, while the NN model predicts instances of sudden RSS drops caused by stationary obstacles, which is a very common challenge in high-frequency networks.
- A new HO event termed "BLOCK" is introduced, which can be considered in B5G and 6G networks besides the standardised events defined by the 3GPP [9]. The BLOCK event is defined by detecting the presence of an obstacle and a user moving toward the blocked area.
- In order to ascertain the optimal timing for initiating and completing a HO once the BLOCK event is detected, this study conducts an analysis aimed at identifying the ideal handover trigger point. The objective is to maintain the user's QoE at a high level.
- The accuracy of the proposed framework is validated using state-of-the-art simulation tools. The results underscore the significance of this solution in ensuring uninterrupted connectivity.

3.3 System and Channel Models

This section offers a detailed description of the system and channel models and discusses the scenario under study.

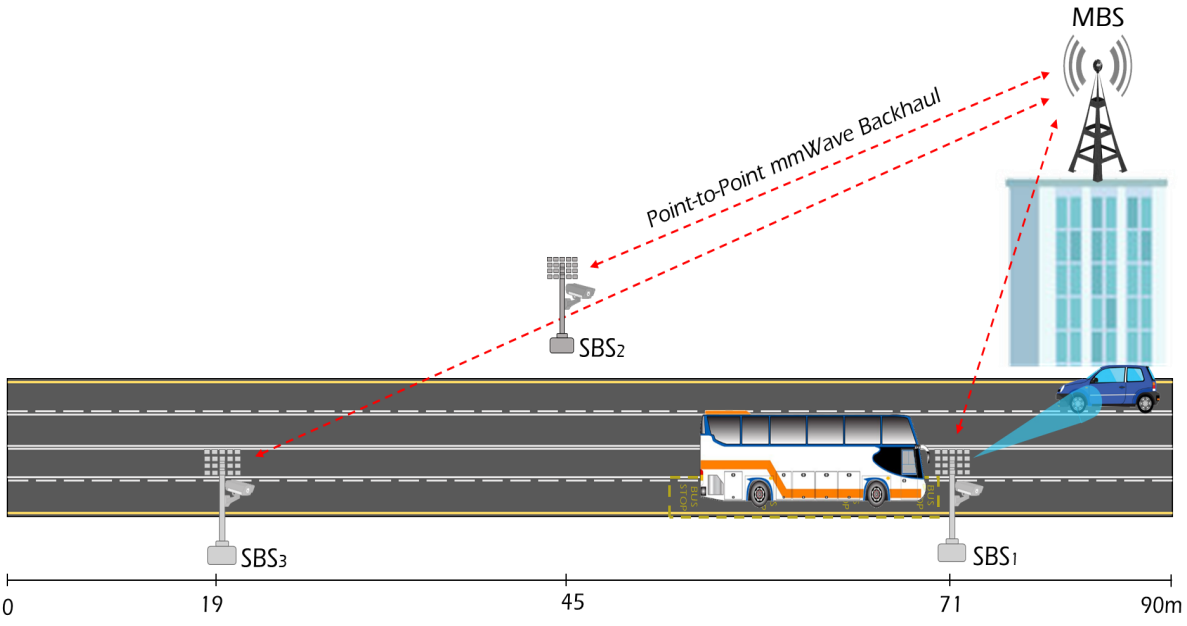


Figure 3.1: The proposed system model: portion of an UDN including one MBS and three SBSs each equipped with an RGB camera.

3.3.1 System Model

UDNs represent a paradigm shift in wireless communications, characterised by an exceptionally high deployment density of small cells. These networks aim to meet the escalating demands for data rates, low latency, and connectivity in densely populated urban areas. In UDNs, the traditional macrocell-centric approach is complemented by a multitude of small cells, such as SBSs, strategically positioned to enhance coverage and capacity. The increased spatial density of these cells allows for improved spectral efficiency and reduced path loss. The system model of this study encompasses one MBS and three SBSs[1], covering a 90×15m street, as illustrated in Fig. 3.1. The system adopts orthogonal frequency division multiplexing (OFDM) with K subcarriers and cyclic prefix of length Q, operating at 60 GHz [112]. Each SBS is equipped with a mmWave uniform linear array (ULA) composed of M antenna elements that enable beamforming technology to create LoS beams that can achieve high RSS at a single-antenna user. To minimise costs and power consumption, this study assumes an analog beamforming architecture with M phase shifters and a single radio frequency (RF) chain [113]. Further, to simplify the network operation, each SBS adopts a predefined beam codebook $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^{B}$, where $\mathbf{f}_i \in \mathbb{C}^{M\times 1}$ and B is the total number of beams in the codebook. Each beamforming vector $\mathbf{f}_i$ can be expressed as [114]:

[1]It is worth noting that: (i) extending this proposed framework to include a larger number of SBSs is straightforward, and (ii) the adoption of the three SBSs model is for the sake of simplicity.

$$\mathbf{f}_i = \frac{1}{\sqrt{M}} \left[1 \; e^{j\frac{2\pi}{\lambda}d\sin(\psi_i)} \; \ldots \; e^{j\frac{2\pi}{\lambda}(M-1)d\sin(\psi_i)}\right]^T, \tag{3.1}$$

where $\frac{1}{\sqrt{M}}$ is the normalisation factor, d is the inter-element distance of the antenna array, λ is the wavelength corresponding to the carrier frequency, and $\psi_i \in \{\frac{2\pi i}{B}\}_{i=0}^{B-1}$ is the steering angle.

To determine the best beam vector that maximises received power, the mmWave user will send a pilot message to the SBS. This pilot message is utilised to train the B beams in order to find the optimal beam $\mathbf{f}^\star$. Once $\mathbf{f}^\star$ is determined, the received downlink signal at the user's receiver for the kth subcarrier can be expressed as follows:

$$y_k = \mathbf{h}_k^T \mathbf{f}^\star s_k + n_k, \tag{3.2}$$

where $\mathbf{h} \in \mathbb{C}^{M\times K}$ represents the mmWave channel between the SBS and the user, s is the transmitted symbol, and $n \sim \mathcal{N}\left(0, \sigma^2\right)$ is the additive white Gaussian noise (AWGN).

In addition to the ULA, each SBS is equipped with a vision sensor – an RGB camera featuring standard definition (SD) resolution. This integration of cameras is aligned with the smaller coverage area of UDNs, ensuring their visual coverage coincides with the radio coverage area of the SBSs. By capturing real-time visual data from the targeted vehicular environment, this integration enhances situational awareness, enabling proactive decision-making, predictive analytics, and optimised resource allocation. The vision information is transmitted to a central server located at the MBS through 10Gbps point-to-point mmWave backhaul links [115]. The role of the central server is to collect, process, and use the visual information to train an ML model that can proactively predict possible beam blockages. This study revolves around Scenario (1) from Section 1.2.1, involving a single moving user (car) with a stationary blocking object (bus) that blocks the LoS communication between the SBS and the user in vision-aided UDNs.

3.3.2 Channel Model

This study adopts the geometric mmWave channel model for several reasons: i) it accurately represents the physical characteristics of signal propagation, and ii) it enables the direct use of accurate channel simulation tools, such as ray tracing, the tool selected to generate the wireless data used to validate this work. The mmWave channel model at the kth subcarrier can be written as [113]:

$$\mathbf{h}_k = \sum_{q=0}^{Q-1}\sum_{p=1}^{P} \alpha_p e^{-\mathrm{j}\frac{2\pi k}{K}q} \ell\left(qT_s - \tau_p\right) \mathbf{a}\left(\theta_p, \phi_p\right), \tag{3.3}$$

where P denotes the number of channel paths, α_p, τ_p, θ_p, ϕ_p are the gain, delay, azimuth and elevation angles of the arrival of path p, respectively. Also, T_s represents the sampling

time.

Optimal Beam and RSS: The maximum RSS value at any x location is associated with finding the optimal beamforming vector $\mathbf{f}_x^\star$ that can achieve this value. In other words, determining $\mathbf{f}_x^\star$ means obtaining the maximum RSS and vice versa. As a result, these can be mathematically expressed as:

$$\mathbf{f}_x^\star = \underset{\mathbf{f}\in\mathcal{F}}{\operatorname{argmax}} \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\left(\mathbf{h}_{k,x}\right)^T \mathbf{f}\right\|_2^2\right], \tag{3.4}$$

and

$$P_{r_x} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left\|\left(\mathbf{h}_{k,x}\right)^T \mathbf{f}_x^\star\right\|_2^2\right], \tag{3.5}$$

where $\mathbf{h}_{k,x}$ is the kth subcarrier's mmWave channel between the SBS and the user at the location x.

3.4 Proposed CV-assisted PHO Framework

The core concept of this work is to anticipate future beam blockage using CV and NN to facilitate timely PHO. Predicting beam blockage is a very challenging task due to its reliance on identifying the position of a mobile user and possible sources of blockage within a realistic wireless scenario. In CV, ODL is used to identify an object's class and its corresponding spatial coordinates. However, object detection alone does not suffice for predicting future blockages. This requirement demands two fundamental components. Firstly, an efficient system that can detect mobile users (wireless users) and potential source of blockage. Secondly, extracting augmented information, including speed, time, and distance from the blocked area. Guided by the above notions, the task of beam blockage prediction is divided into two sub-tasks. The first sub-task involves ODL to identify object types and their coordinates, aiding in speed calculation. The second sub-task employs a multivariate regression model to predict the remaining time until a user reaches the blocked area based on the information extracted from the RGB images. Before delving into a detailed discussion of the various components of the proposed framework, the assumptions made in this study are highlighed as follows:

1. It is assumed the availability of vision sensors integrated with SBSs. Moreover, the sensors can provide flat RGB images that exhibit uniform pixel widths for both the upper and lower segments of the street.

2. The vision sensors remain unaffected by the time of the day (day/night) and prevailing weather conditions.

3. The MBS has local processing units where optimisation and localised decision-making take place. In addition, the ODL consistently provides high performance in object prediction and bounding box determination, all within an acceptable level of precision.

4. The wireless user is identified in the image, and if necessary, the network correctly performs PHO for that user.

In the worst-case scenario where the assumptions are not fulfilled, the wireless network would lose the advantage of vision assistance in solving the beam blockage problem and would revert to functioning as a basic wireless network devoid of proactive blockage prediction capabilities.

3.4.1 Schematic Diagram of the Proposed Framework

This study aims to establish a mechanism for predicting blockages, enabling the network to proactively initiate HO procedures well in advance of a user entering a blocked area. *Once a BLOCK event is spotted out in the camera's field of view, the framework's main task is to predict the time required for the user to reach the shadowed region, denoted as* T_b. Determining this time interval enables the identification of the optimal moment for initiating HO, thus preventing service disruption that could occur if the user enters a blocked region. Fig. 3.2 demonstrates the schematic diagram of the proposed technique. Given that wireless networks frequently operate in dynamic and unpredictable environments like urban areas and smart cities, a heuristic approach is adopted. Moreover, increasing the network complexity usually leads to non polynomial (NP)-hard optimisation problems. Such problems demand a very high computation time, which is intolerable and challenges latency-sensitive applications, such as intelligent transportation systems. Therefore, the nature of the considered problem motivates the use of heuristic models [116, 117]. Multivariate regression is used to predict the T_b by modelling and training a two-hidden layer NN. Initially, the server compiles a complete view of the covered area (i.e., the street), detailing the exact coordinates and locations of each SBS. RGB cameras continually[2] capture images from the covered area, then every SBS adds its identification number and timestamp to each image before sending it to the central server through the mmWave backhaul link. Once received, the server performs the following tasks:

- First, it uses the ODL algorithm to detect blockages and users, consequently updating its view. If a BLOCK event is detected, the server proceeds to the next step; otherwise, it reverts to the detection phase.

[2]Note that certain cameras have a motion detection feature that can be activated to reduce the volume of visual data transmitted to the server [118].

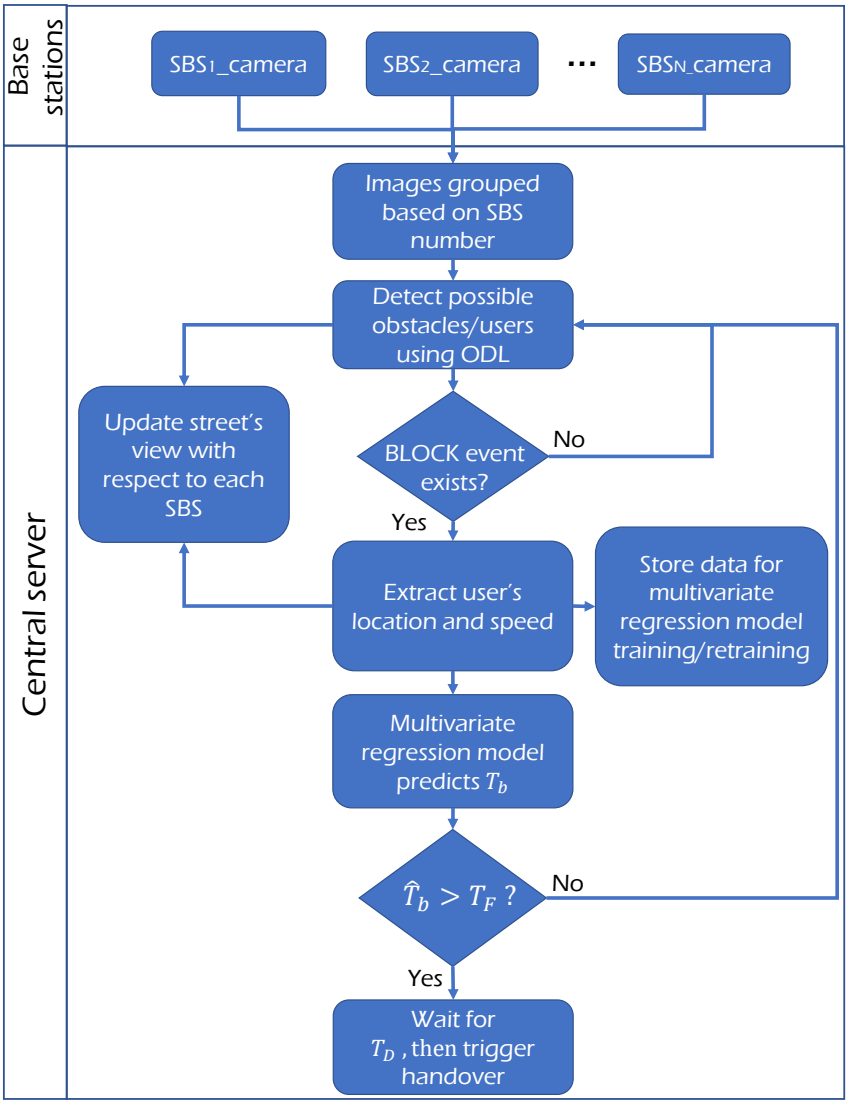


Figure 3.2: Schematic diagram of the proposed framework.

- Subsequent to that, the server identifies the user's exact location, thereby updating its view. By utilising the user's location information and calculating the time difference between consecutive images, the user's speed is determined.
- Following this, the server stores both the location and speed information for model training or retraining purposes. This collected data is also used to predict the T_b.
- Finally, if the T_b exceeds the execution time of the proposed framework (T_F), the server enters a waiting period before sending a HO trigger event to the network in order to HO the user to another SBS. Conversely, if this condition is not met, the server reverts to the initial detection phase.

The value of T_F *is defined as the time taken by the proposed framework to be completed, starting from the initial capture of* RGB *images to the finalisation of the* HO *process.* It's

expressed in (3.6) and can be calculated through the summation of four distinct time components: (i) The time required to send two consecutive RGB images to the central server (T_{RGB}), (ii) the duration taken for ODL on these images (T_{ODL}), (iii) regression model inference time (T_{Inf}), and (iv) PHO implementation time (T_{PHO}).

$$T_F = T_{RGB} + T_{ODL} + T_{Inf} + T_{PHO}. \tag{3.6}$$

In addition, a new time parameter is defined, denoted as (T_D), representing the interval between the completion of regression inference and the initiation of the HO triggering. The value of T_D is adaptable and dependent on the establishment of the optimal trigger region, a topic detailed in Section 3.4.4. However, it's essential to note that there exists an upper limit for T_D, as elaborated below:

$$T_D^{max} = T_b - T_F. \tag{3.7}$$

It's important to highlight that the values of all parameters specified in (3.6) remain constant, while the value of T_D varies based on the user's location and speed. Determining the values of T_{RGB}, T_{ODL}, T_{Inf}, and T_{PHO} relies on factors such as the capacity of the mmWave backhaul links, the chosen model type, and the specifications of the central server. While choosing the value of T_D is tied to the definition of the optimal HO trigger region where the framework identifies the most appropriate timing for executing the HO process.

3.4.2 Object Detection and Localisation

The central server necessitates the reception of a minimum of two consecutive images (frames) from each SBS to detect the objects' presence, position, and speed. Table 3.1 lists the most common camera resolutions for surveillance applications, alongside the corresponding image transmission times over 10 Gbps mmWave backhaul links. Notably, Table 3.1 illustrates that higher camera resolutions incur longer transmission times due to the production of larger image sizes. Since this study considers an SD camera resolution, T_{RGB} is equals the time required for transmitting two SD images plus 38.5 ms. This extra time accounts for the temporal gap between capturing consecutive RGB images, assuming an imaging rate of 26 frames per second for each camera [119]. Hence, the time required to transmit these images from the SBS to the server, T_{RGB}, approximately amounts to 40 ms.

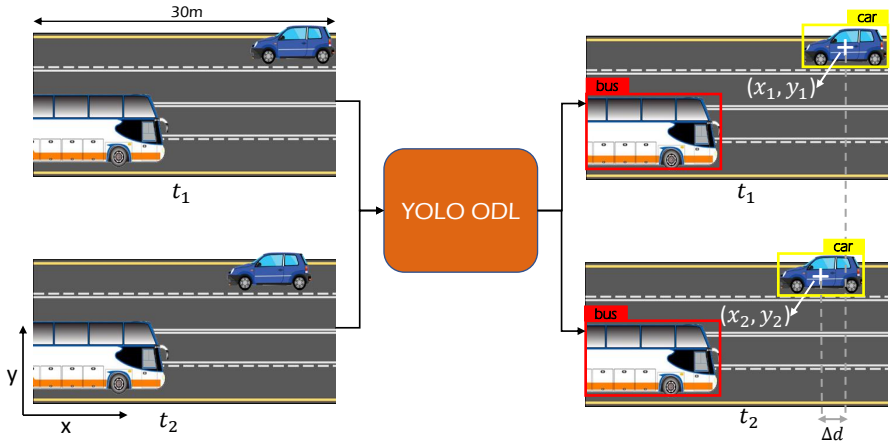


Figure 3.3: Using ODL to detect objects and determine their locations. This information is used to determine the speed of the moving object.

Table 3.1: The most common camera resolutions with associated image transmission times over 10 Gbps links.

Term	Resolution	Transmission time (ms)
2CIF	704×240	0.4
SD	640×480	0.7
4CIF	704×480	0.8
HD	1280×720	2.2
FHD	1920×1080	5

Once the server receives the visual information, the first step involves processing the data to obtain the location of the objects. This study adopts a state-of-the-art detection model "you only look once" version 3 (YOLOv3), renowned for its rapid and accurate real-time object detection capabilities [120]. Instead of developing and training an object detection model from scratch, YOLO models can be used directly without the need for any modification. Furthermore, the main objective of ODL is to identify if a BLOCK event exists and determine the objects' locations in the pixel-scale. This information is then converted to the meter-scale units for deducing the objects' speeds. The server feeds the RGB images to the object detection algorithm, which, in turn, detects the objects within the image by drawing bounding boxes around them and adding tags denoting their respective categories, as illustrated in Fig. 3.3. Moreover, this algorithm will provide the location information of the objects by determining the coordinates of the upper-left and lower-right corners of the bounding boxes. This information is utilised to determine the centre of the moving user, as shown in Fig. 3.3.

The location information obtained from the YOLOv3 model is in pixel units. However, to compute the user's speed, it's imperative to convert this information into a metric measurements. The cameras' field of view is set to 100 degrees, ensuring comprehensive

Table 3.2: Considered camera properties.

Parameter	Value
Resolution	standard definition (SD)
Field of view	100 degree
Frame rate	26 fps

coverage of the entire street while minimising any overlap. Given that the cameras capture images in two dimensions only, it is assumed that the image is flat and its width corresponds to a distance of 30m[3]. Therefore, the following formula can be used to determine the user's displacement in metres:

$$Travelled\_distance = \frac{W_m}{W_p} \times \Delta d, \tag{3.8}$$

where W_m is the image width in meters, W_p denotes the image width in pixels, and $\Delta d = |x_1 - x_2|$ signifies the user's x-axis displacement in the pixel scale, assuming the user is moving in a straight trajectory, as demonstrated in Fig. 3.3. For example, if a camera produces images with a resolution of 640×480, approximately every 21 pixels are equal to a distance of one meter. After evaluating the travelled distance, it is necessary to determine the associated travel time, which can be easily measured by exploiting the timestamp information of each image. Since the camera's frame rate is 26 fps, the time difference between two consecutive images will be about 38.5 ms. Now, the following speed formula can be used to deduce the speed of the moving user:

$$Speed = \frac{Distance}{Time}. \tag{3.9}$$

Since the proposed solution depends on the T_F to determine the feasibility of initiating PHO, as demonstrated in Fig. 3.2, it becomes imperative to find out the detection time of the YOLOv3 model (T_{ODL}). Benefiting from the high performance of MEC servers, the T_{ODL} can be reduced significantly to mere tens of ms, particularly when using the MEC server as a central server. Based on the analysis presented in [120], it is assumed that that the T_{ODL} requires 102 ms for detecting objects in two images. Table 3.2 summarises the camera properties considered for this study.

3.4.3 Multivariate Regression: Learning and Prediction

In this section, the reasons for selecting ML techniques to predict the value of T_b will be explored, along with an overview of the training and inferencing procedures employed for

[3]For practical implementation in real-world systems, determining image width and height in metric units can be easily accomplished by capturing an image from the camera and measuring its actual dimensions using image corners.

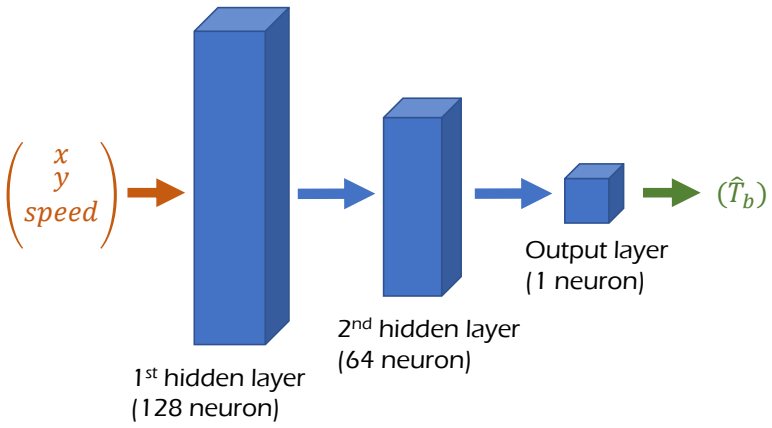


Figure 3.4: A two-hidden layer NN to perform regression.

the selected model. Using analytical methods can fulfil the task of determining the T_b in general. However, the primary aim of this work is to propose a flexible, scalable, and transferable framework that can be employed in complex and ever-changing environments. Therefore, a NN model has been utilised for the following reasons:

1. Analytical solutions generally perform very well in static systems, where the model's assumptions do not change. However, this study targets high-frequency wireless networks that are notably dynamic and unpredictable. For instance, predicting T_b greatly depends on the cameras resolution, which varies in practical wireless systems and provides diverse data. ML models learn from problem-specific data to automate the process of an analytical model and solve the associated tasks [121]. Therefore, NN is used to provide a more generalised and scalable solution.

2. Given the dynamicity and complexity of the targeted environment, a model that can scale well under these conditions is needed. NNs are flexible and adapt to the dynamics of the problem by learning from diverse data. Moreover, the transfer learning feature allows these models to be transferable, providing the proposed solution a new dimension to perform distributed learning.

DL algorithms have achieved breakthroughs in various areas but at the expense of high computing and energy consumption. Combining CV with DL to assist the operation of UDNs amplifies the computational complexity of the models, rendering this fusion inefficient [122]. Since this study depends mainly on vision information, the selection of an appropriate ML model is cruical, one that minimises training duration while achieving anticipated outcomes. In the previous section, the utilisation of the pretrained YOLOv3 ODL model is explored. This off-the-shelf solution can be used directly, eliminating the need for additional training efforts. Furthermore, for predicting the T_b, a simple NN

Table 3.3: Sample of the training dataset.

$x\ (m)$	$y\ (m)$	$Speed\ (m/s)$	$T_b(ms)$
68.47	9	1.5	60.00
74.21	10	5	1166.00
81.22	9	10	1284.00
71.88	10.5	12	291.67
75.92	9.5	15	502.67
81.22	10	18	713.33
86.91	11	20	926.50

model has been chosen to conduct multivariate regression, as shown in Fig. 3.4. The multivariate regression technique is a statistical approach that measures the relationship between dependent variables (i.e., T_b) and multiple independent variables (i.e., x, y, speed). Besides, instead of using information-rich RGB images in model training or inferencing, the proposed technique only requires extracting the user's location and speed to be fed into the NN model. This yields to significant savings in time.

Training Phase: The operations of the proposed regression model entail centralised model training and inference; both require the availability of data samples. For training the initial model, training datasets are readily generated using random (x, y) coordinates confined to the dimensions of the system model's street and using several speed values that reflect the expected vehicle speed in urban areas. The dependent variable T_b is derived by fixing the location of the obstacle (e.g, a bus) at an arbitrary location, for example at $x = 68.38$m as illustrated in Fig. 3.1, and using the speed formula given in (3.9). Table 3.3 shows a small sample of nearly ten thousand generated data samples that are divided into 70% training, 20% validation, and 10% testing. Moreover, Table 3.4 shows the selected model structure and the selected hyperparameters, where the adaptive moment estimation (Adam) is used as the optimiser in the training process. Moreover, other model hyperparameters, such as the number of epochs, batch size, metric, and activation function, are set to 50, 20, mean square error (MSE), and linear activation function, respectively. Fig. 3.5 shows the training and validation loss for the model in each epoch. This figure underscores the effectiveness of this model, as the loss consistently approaches zero with an increasing number of epochs. It is worth mentioning that the number of epochs depends on the number of data samples used for model training. A small dataset and a large number of epochs can lead to model overfitting, an unwanted behaviour for predictive modelling. To assess the model's performance, the coefficient of determination (R-squared) metric is employed. This metric measures the linear correlation

Table 3.4: Hyperparameters of the NN model.

Parameter	Value
Layers information	Input layer: 3 features 1st Hidden layer: 128 neurons, ReLU 2nd Hidden layer: 64 neurons, ReLU Output layer: 1 neuron, Linear
Loss functions	MSE
Optimiser	Adam
Batch size	20
Epochs	50
Data split	70% training - 20% validation - 10% testing

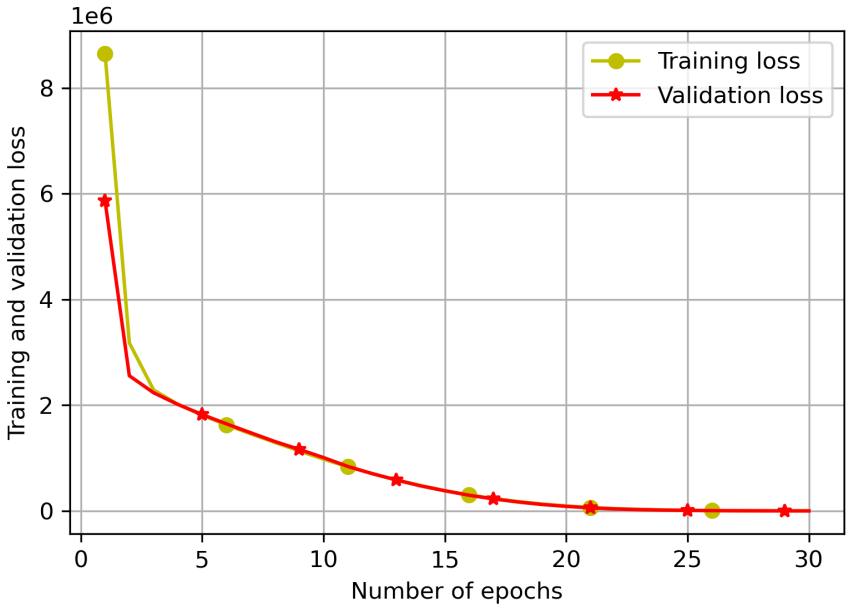


Figure 3.5: Multivariate regression model training and validation loss versus number of epochs.

between predicted and actual values using the test dataset, with values ranging between 0 and 1, where 1 signifies optimal alignment. Using the test dataset, the proposed model achieves 0.9998, highlighting its superb performance. Finally, based on the generated dataset and the hyperparameters delineated above, training the proposed regression model consumes about 20 seconds when utilising standard personal computer resources. However, leveraging the MEC server expedites the training time to under one second [123].

Inference Phase: With the trained regression model in place, the framework is now ready to predict T_b for users detected in images received from the cameras. Upon completing the street view and updating the blockage status, whenever the server receives RGB images from the SBSs, it will use the ODL to extract the location and speed of the

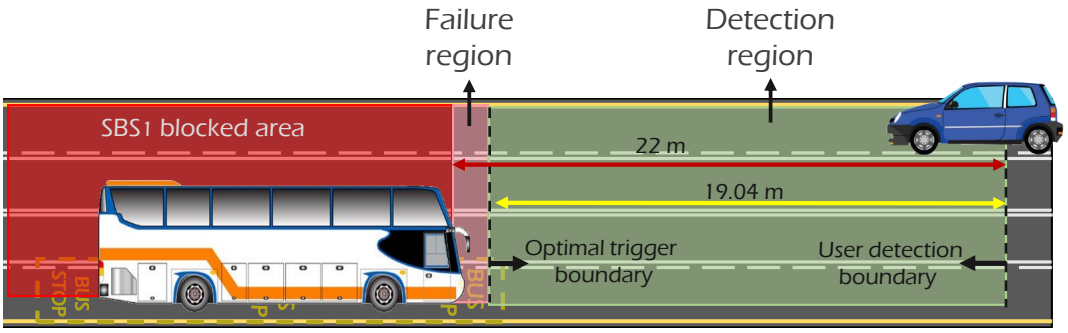


Figure 3.6: Optimal trigger distance for a user with a speed of 30 mph.

moving user as discussed in Sec. 3.4.2. The user's information is now ready to be fed to the input layer of the regression model, enabling the inference of the the anticipated time remaining until the user reaches the obstructed area. Furthermore, by using the same MEC server resources, the value of T_{Inf} is estimated at approximately 1 ms.

3.4.4 Optimal HO Trigger Distance

After attaining all requisite parameters, encompassing user location, speed, and T_b, the final step is to determine the optimal HO trigger location, aiming to maintain the user's QoS/QoE at their highest possible levels. This optimal location signifies the distance at which the central server initiates the HO request after detecting the BLOCK event, and the PHO is performed with minimum performance degradation. In this work, a threshold distance-based configuration is adopted, wherein the central server determines the optimal trigger distance using the following equation:

$$\Lambda_{opt} = v \times T_D^{max}, \tag{3.10}$$

where v is the speed of the user, which is already known from ODL. From equations (3.7) and (3.10), the variable trigger distance Λ can be inferred as:

$$\Lambda \leq v(\hat{T}_b - T_F), \tag{3.11}$$

where T_F is the sum of the four sub times i.e., T_{RGB} is 40 ms, T_{ODL} 102 ms and T_{Inf} is 1 ms whereas the T_{PHO} is 80 ms as will be discussed in the next section. The formula in (3.11) can be used to differentiate between early and optimal HO decisions.

To determine Λ_{opt}, the initial step involves identifying the detection region. This region is defined as the area within which the proposed framework is assigned the role of monitoring and detecting the presence of a wireless user. The detection region is confined within two boundaries, i.e. the user detection boundary and the optimal trigger boundary. This region is followed by the failure region located between the optimal trigger

Table 3.5: Optimal trigger distance based on different user speeds.

v (mph)	T_b (sec)	T_D (sec)	Λ_{opt} (m)
5	9.85	9.63	21.52
10	4.92	4.70	21.01
15	3.28	3.06	20.52
20	2.46	2.24	20.03
25	1.97	1.75	19.56
30	1.64	1.42	19.04
35	1.40	1.18	18.46

boundary and the blocked area, as shown in Fig. 3.6. The user detection boundary signifies the starting point of the algorithm's detection process for any potential BLOCK event. On the other hand, the optimal trigger boundary, denoted as Λ_{opt}, represents the minimal distance from the blocked area where a successful optimal HO can be executed. For instance, any HO undertaken prior to reaching the optimal trigger boundary, the user will experience undesirable performance degradation due to wireless channel path loss. Whereas, performing the HO beyond the optimal trigger boundary risks a LoS link being obstructed by the blocking object, owing to insufficient time for executing the PHO algorithm.

Therefore, the algorithm must determine the best T_D value to facilitate the seamless transition of the user to another SBS at the optimal point, thereby avoiding service disconnection. To successfully perform HO, the T_F is found to be equals 223 ms as discussed earlier. The formula in (3.11) serves as a tool to determine the optimal point for performing optimal HO, while also study the effects of early HO execution on overall system performance. For extensive analysis, optimal trigger distances based on different speeds are given in Table 3.5. For example, if a car is detected at $x = 90$m and is moving at 30 mph, the optimal trigger distance to perform PHO is 19.04 m, as shown in Fig. 3.6. Once a BLOCK event is detected, the optimal trigger distance can be calculated using (3.11). With v is already known, T_b is obtained using regression analysis, and T_F is also known, the potential for executing PHO within the detection region becomes viable. However, it's worth noting that early HO, while feasible within the detection region, can compromise the RSS value, an undesirable situation. Therefore, in the proposed model, the central server waits for T_D^{max} until the user reaches the optimal distance to complete the PHO request.

3.4.5 Proactive Handover Mechanism

T_{PHO} stands as a pivotal parameter within the proposed CV-aided PHO framework, acting as a determining factor in assessing the feasibility of performing HO while avoiding radio

links failures. Conventionally, in cases of user link disconnection, a series of steps precedes reconnecting the user to the same or an alternate SBS. This sequence entails beam failure detection, beam failure recovery, cell search, and contention-based/free random access [124]. Supposing the network employs proactive blockage prediction, the first two steps can be avoided, whereas the cell search can be performed while the user is still connected to the serving SBS. Therefore, T_{PHO} boils down to the latency accompanied by performing contention-based or contention-free random access. This study considers contention-based random access, which requires 80 ms according to the 3GPP specifications [124,125]. Since the values of all parameters in (3.6) are determined, the central server possesses knowledge of the time needed to execute the proposed algorithm—T_F equals 223 ms. If the central server detects a BLOCK event in the received RGB images, the server will predict the time needed until the user reaches the blocked area (T_b). If T_b is greater than T_F, then the algorithm has a high probability of successfully triggering and completing HO. Whereas when T_b is less than T_F, the time needed to complete the HO process and avoid radio link failure is insufficient, which means that the user undergoes a service interruption.

3.5 Performance Evaluation and Results

To investigate the effectiveness of the proposed CV-based PHO framework, the publicly available dataset known as vision wireless (ViWi) [126] is utilised. The ViWi dataset combines visual and wireless information generated using Wireless InSite ray-tracing software and 3D game modelling for mmWave wireless systems. It encompasses four distinct scenarios, categorised by camera location (collocated and distributed) and view (direct and blocked). Each data sample represents 4-tuple of user location, RGB image, depth image, and wireless channel. The distinctive feature of ViWi is that it is a parametric, systematic, and scalable data generation framework that can be used to produce data based on different scenario requirements. In the performance evaluation, the focus is to track the strength of the signal received by the moving user and test the value of the proposed framework in maintaining a strong received signal during movement compared to traditional mmWave systems (i.e., those lacking Proactive blockage prediction capabilities).

3.5.1 Simulation Setup

In the simulation, an environment containing a blocking object located near the SBS_1 and a single user moving at a speed of 30 mph is considered, as illustrated in Fig. 3.1. The user's trajectory is from left to right, with SBS_1 selected as the serving BS since the RSS from that BS is higher than other SBSs. Moreover, the static obstacle blocks the LoS communication between the SBS_1 and the user when it reaches the obstacle's

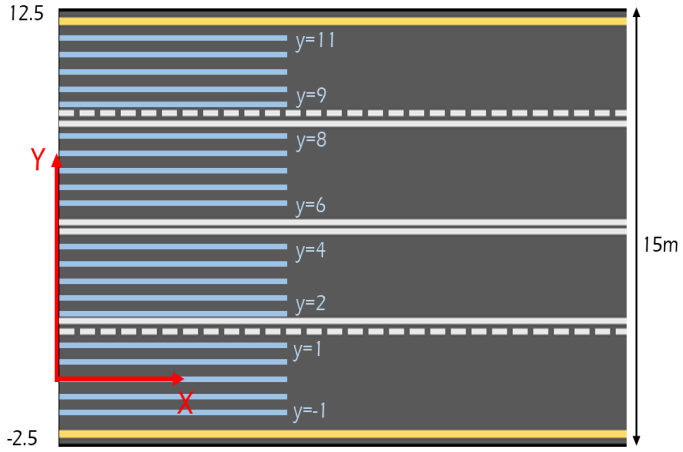


Figure 3.7: Locating the origin of the Cartesian coordinates in ViWi scenarios.

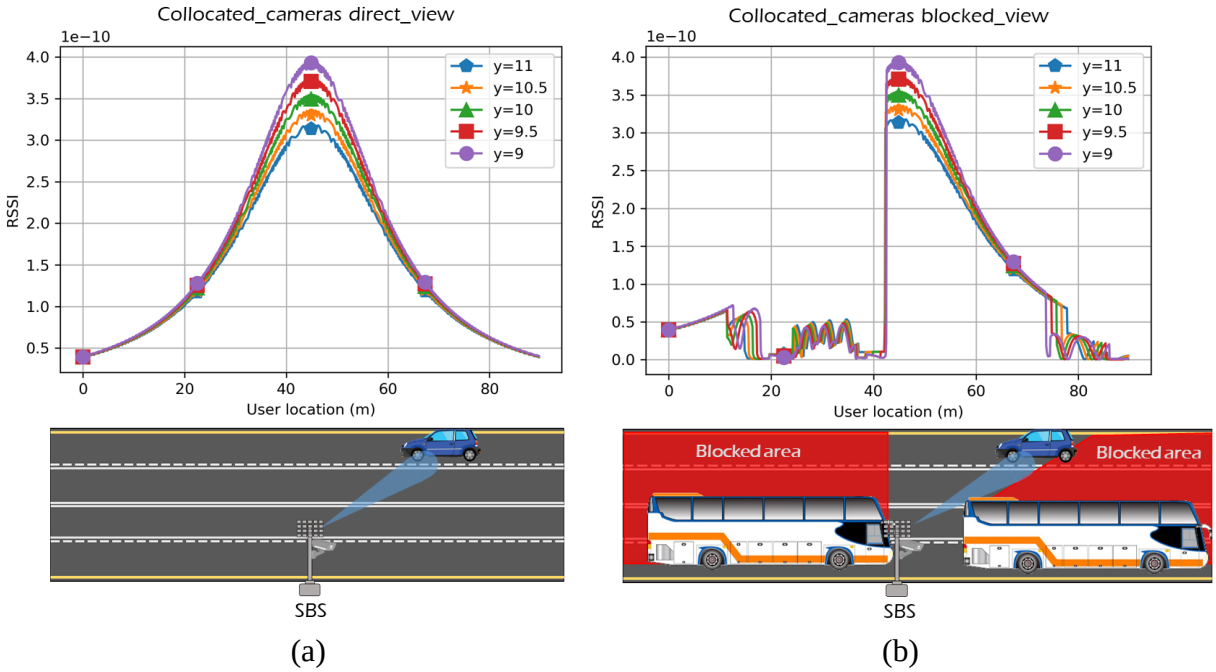


Figure 3.8: (a) Using ViWi information from colocated cameras direct view scenario to model SBS_2 in the system model, and (b) ViWi information from colocated cameras blocked view scenario shows a similar RSS pattern when there is no blockage.

Curve fitted formula:

$$RSSI(x) = \sum_{m=0}^{15} a_m x^m$$

Coefficients:

a_0	3.69140438e-11	a_8	-4.79406833e-18
a_1	2.38244809e-11	a_9	9.59204806e-20
a_2	-1.50142337e-11	a_{10}	-1.33887437e-21
a_3	4.42903748e-12	a_{11}	1.28057711e-23
a_4	-6.95624681e-13	a_{12}	-8.04694410e-26
a_5	6.59641831e-14	a_{13}	3.04183473e-28
a_6	-4.04638010e-15	a_{14}	-5.62049872e-31
a_7	1.67495163e-16	a_{15}	2.23997329e-34

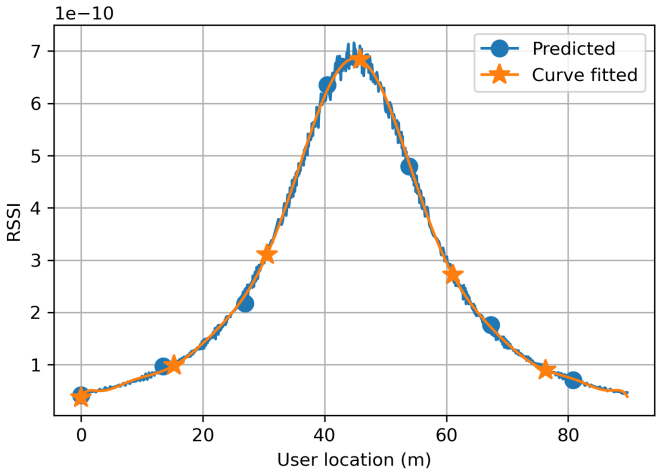


Figure 3.9: Determining the RSS from SBS_2 at trajectory y= 9 using the curve fitting tool.

blocked region. The system model is different from any of the scenarios introduced with the ViWi dataset. Nevertheless, visual and wireless data pertinent to this model were generated by merging two scenarios involving collocated cameras, direct and blocked views. Figures 3.7 and 3.8 show the analysis performed to construct the system model. The analysis begins by examining the information provided in the ViWi dataset to determine the place of origin of the Cartesian coordinate system, which is not explicitly stated. The ViWi trajectories (y= 9 to 11) were utilised to identify the origin's location and other trajectories, as shown in Fig. 3.7. Then, the RSS is plotted against the vehicle location for each trajectory considered in ViWi, as illustrated in Fig. 3.8 (a) and (b). These figures indicate a consistent signal pattern (bell shape) for each SBS across various trajectories, with the highest signal strength occurring at the same x-location of the respective SBS. The RSS modelling of SBS_2 involved using the information from Fig. 3.8(a) and applying curve fitting to establish a mathematical formula. First, RSS values were predicted for each other trajectory (y= -1 to y= 8) at each user location. Subsequently, a curve fitting process is employed to formulate the mathematical relationship between RSS and vehicle location (x). Fig. 3.9 illustrates the resulting mathematical formula and RSS from SBS_2 at y=9, which is the trajectory used in the evaluation. Finally, Python programs are used to conduct the simulations.

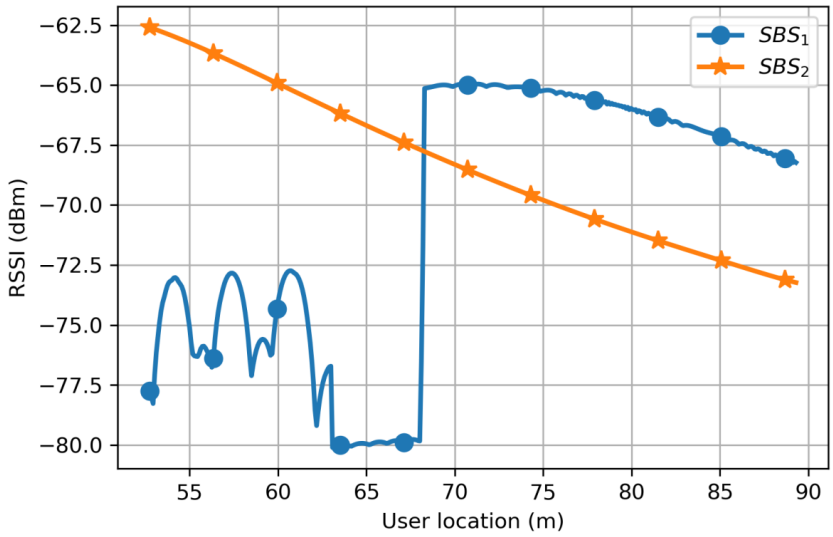
3.5.2 Simulation Results

In what follows, the usefulness of the proposed framework is examined in terms of maintaining physical link connectivity and ensuring a timely and seamless transition between SBSs. The received signal strength indicator (RSSI) is used as a metric to measure

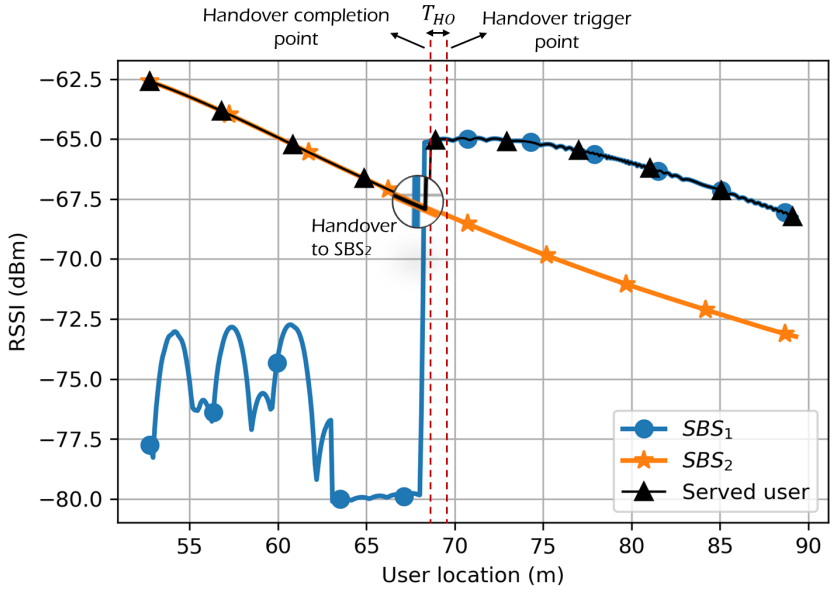
the quality of signals received from nearby SBSs. Fig. 3.10(a) plots the RSSI received from SBS_1 and SBS_2 along the y=9 trajectory on the street. It is noticed that the signal drops from SBS_1 when the user enters the area behind the blocking object, owing to severe signal attenuation. In contrast, the signal received from SBS_2 remains unaffected, as there are no interruptions in the LoS path between the user and SBS_2. In a traditional wireless network, which does not utilise the PHO algorithm, the user experiences connectivity disruptions when entering the blocking area. Such interruptions may lead to service drops and necessitate establishing new connections, resulting in delay and poor QoE. This contradicts the vision of 5G/6G wireless networks of providing URLLC. Fig. 3.10(b) demonstrates the capability of the proposed PHO algorithm in proactively predicting beam blockages. This figure reveals the efficiency of the proposed algorithm in identifying BLOCK events in advance and triggering a timely HO. The merit of determining the optimal point of triggering HO in the proposed algorithm is vital to maintain the QoE at high levels and avoid early HO, which could lead to bad system performance. The points of triggering HO and HO completion are also shown in the figure, which illustrates how this framework is QoE-aware.

Fig. 3.11 shows the results of the early/optimal HO trigger distance for the user moving at the speed of 30 mph. The optimal trigger boundary is the minimum distance where the central server can perform successful HO while minimising the degradation of performance, quantified by the percentage drop in normalised RSSI. In this specific case, the optimal distance for a user moving at a speed of 30 mph is found to be around 70 m from the origin. Upon detecting a BLOCK event, the central server performs HO, causing the user's connection resources to transition from SBS_1 to SBS_2. During the HO process, the user experiences a drop in RSSI due to path loss. For instance, if the central server performs an early HO, occurring 5 meters before the optimal trigger boundary, there is a power drop of approximately 20 %, as shown in Fig. 3.11. therefore, the optimal trigger distance provides the trade-off between the PHO success rate and the drop in RSSI to maintain the seamless connectivity.

Next, the investigation delves into the effectiveness of the PHO algorithm in improving the reliability of high-frequency wireless networks by considering a real-time application sensitive to service interruption and network latency. A moving user that is running a video call is considered, and the mean opinion score (MOS) is employed as a metric for QoE. MOS quantifies the overall perceived quality of media services, rated on a scale from 1 to 5 (1-bad, 2-poor, 3-fair, 4-good, 5-excellent) [127]. In Fig. 3.12, the relationship between MOS values and user location is depicted, capturing the variation in RSS due to different distances from the SBS and the presence of obstacles. To translate the values of RSS to the corresponding values of MOS, the mapping table in [128] is adopted. Fig. 3.12 reveals that without the PHO algorithm, service interruption occurs as the user enters the blocked area,



(a)



(b)

Figure 3.10: Performance evaluation of the proposed framework. (a) RSS from SBS_1 and SBS_2, (b) Using the CV-assisted PHO algorithm to detect BLOCK events and trigger PHO.

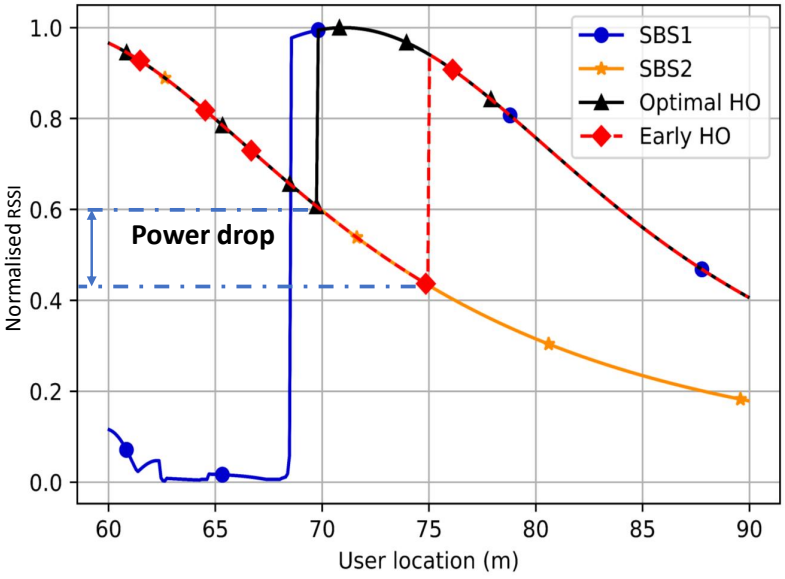


Figure 3.11: The normalised RSSI as function of the user location when the user speed is fixed at 30 mph.

causing a 40% drop in call quality, placing it within the poor MOS range. The call quality remains poor until the user disconnects from the serving SBS and connects to a new one. Conversely, leveraging the PHO technique, the algorithm intelligently detects blockage presence and countermeasures the possible signal blockage by triggering HO in advance. As a result, the perceived MOS remains excellent. This work enhances the reliability of UDNs that will facilitate the realisation of future latency-sensitive applications.

3.6 Summary

In this chapter, a novel CV-assisted PHO framework is introduced to address the challenge of beam blockage and frequent HO in next-generation wireless networks. The core concept revolves around enhancing network awareness of the surrounding environment by leveraging visual information and using CV to predict BLOCK events and facilitate PHO operations. The framework adopts a centralised training approach, utilising datasets gathered at the central server to conduct model training. Moreover, the framework employs a pretrained object detection model alongside a multivariate regression model to predict obstacle/user location and estimate the time remaining before the user reaches a blocked area. Additionally, the framework exhibits a QoE-awareness, featuring an analysis of optimal location and timing for HO execution while minimising QoE degradation. Evaluation results demonstrated that this framework is able to avoid 40% service reduction and maintain a high level of perceived QoE. Accordingly, this work improves the

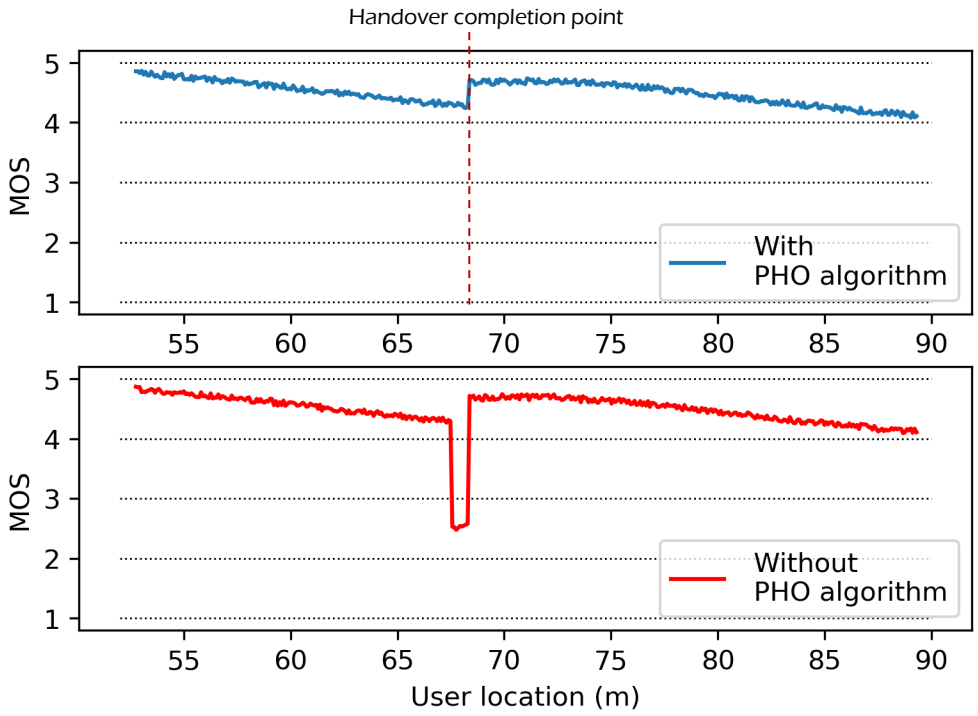


Figure 3.12: Measuring the QoE with and without PHO.

performance of UDNs rendering them more responsive to their environment, aligning with the vision of achieving low-latency and time-sensitive applications in B5G and 6G networks. The following chapter aims to further enhance the proposed CV-based PHO framework by embracing the FL training mechanism as an alternative to the centralised approach. This shift in training methodology is intended to address the challenges associated with the conventional training methods. Furthermore, the forthcoming chapters seeks to investigate more complex environments that encompass diverse users and multiple blocking objects.

Chapter 4

Blockages Prediction in Dynamic Vision-aided UDNs with Distributed Learning

After investigating the beam blockage problem and introducing a novel CV-aided PHO framework that enhances the reliability of UDNs, as discussed in the previous chapter, this chapter further improves this framework by considering complex communication environments. These environments may involve multiple static or dynamic objects and users. In this scenario, dynamic positioning table (DPT) is introduced to distinguish between wireless users and differentiate them from other objects in the captured vision information. Moreover, a distributed learning approach is embraced rather than the centralised learning technique, aligning with the aims specified in *O2*. This shift helps alleviate the pressure on communication channels, as it avoids sending large-sized and voluminous visual information to a central server, therby preserving the privacy of this data. The performance of the framework is evaluated based on critical parameters, including the speed of users and objects, average HO latency, and users' QoE.

4.1 Introduction

The challenge of beam blockage and frequent HOs becomes significantly more pronounced in highly dynamic environments characterised by a dense presence of users and obstacles, whether these elements are in motion or stationary. In these complex settings, wireless users can be blocked by other active wireless users or various passive objects. Building upon the previous study, this work takes a step forward by introducing an enhanced and more comprehensive FL-based CV-aided PHO framework, specifically developed to address the persistent issues of beam blockage and HO in scenarios involving multiple users and obstacles. By harnessing the power of FL and designating the MBS server

as the orchestrator for the FL process, this new framework incorporates collaborative model training at the network's edge, where SBSs are situated. The images captured by the cameras positioned at the SBSs are stored locally to be used for training a model designed to aid in predicting possible forthcoming blockages and assisting the operation of the networks. This approach offers numerous advantages, including safeguarding data privacy, efficient utilisation of network transmission resources, latency reduction, lower energy consumption, and eliminating the need for high storage capacity at a central server. Furthermore, the incorporation of CV equips the network with a comprehensive understanding of its surrounding environment, aiding in analysing the environment using the wealth of information derived from the captured RGB images. Nonetheless, this hinges on two key factors: detecting objects within the environment and identifying their mobility patterns, as elaborated in the previous chapter. Additionally, the need to distinguish wireless users from other passive objects in still images as will be discussed in detail in the subsequent sections.

4.2 Contributions

The following points summarise the main contributions of this chapter:

- The CV-aided blockage prediction problem is formulated for multi-user/object UDNs, and an end-to-end latency-aware framework is developed to leverage RGB images for proactive blockage prediction and PHO. This ensures that the QoE of users remains as high as possible.
- The framework incorporates FL as a distributed learning approach rather than the conventional centralised learning method. This approach involves training the model locally at each SBS where the visual information is stored, ensuring data privacy, and reducing communication overhead.
- Finally, the accuracy of the proposed framework is validated using modern simulation tools. The simulation results underpin the significance of this solution in maintaining seamless connectivity for highly dynamic UDNs.

4.3 Network Model

This study targets UDNs prevalent in smart cities, where the environment is challenging due to numerous mobile users and obstacles. The outdoor mmWave system in consideration comprises an MBS and many SBSs, as depicted in Fig. 4.1. For clarity, this figure shows only two SBSs and a segment of a street as a small portion of a UDN. OFDM

with K subcarriers is adopted as the modulation scheme based on 28GHz. Each SBS has a camera that monitors objects within its field of view. Moreover, it has an M-element antenna array that enables beamforming to serve single-antenna mobile users with beams selected from a predefined beam steering codebook $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^{B}$, where $\mathbf{f}_i \in \mathbb{C}^{M\times 1}$ and B denotes the total number of beams.

The network's primary focus is to identify the optimal beam that achieves the highest RSS at the user end ($\mathbf{f}^\star$). Given this, an area of interest (AoI) is defined as the coverage area that achieves the optimal RSS when the users are connected to the corresponding SBS. QoE serves as the key performance metric that this study aims to maximise. Consequently, the study assumes that the network will perform HOs when users cross the boundaries of AoIs. Moreover, the chosen channel model for the mmWave environment is the geometric model, as it effectively captures the geometrical distribution of the environment and is commonly used in practical mmWave systems [129]. Therefore, the downlink received signal at subcarrier k is:

$$y_k = \mathbf{h}_k^T \mathbf{f}^\star s_k + n_k, \tag{4.1}$$

where $\mathbf{h}^T$ is the transpose of the downlink channel, s is the transmitted symbol, and n represents the AWGN. In addition, the received power at the user side can be determined as follows:

$$P_r = \frac{1}{K}\sum_{k=1}^{K}\left|\mathbf{h}_k^T \mathbf{f}^\star\right|^2. \tag{4.2}$$

4.4 Problem Formulation

The beam blockage problem can be formally defined as follows. The camera captures frames of RGB images, and image processing is applied to produce flat red-green-blue (F-RGB)[1] images focused on the AoI. Each F-RGB image is assumed to contain O objects, and every object $o \in O$ will be monitored until it leaves the SBS's AoI. The F-RGBs are fed into an ODL algorithm to obtain boundary-boxes information about every object. This information is then converted into a 6-dimensional metric vector $[x_{ul}, y_{ul}, x_c, y_c, x_{lr}, y_{lr}]$, where the subscripts ul, c, and lr indicate the upper left, center, and lower right coordinates of the boundary boxes, respectively. The complete mobility vector ($\mathcal{L}$) of any object o is shaped by adding its movement direction ($\vec{d}$) and speed (v), as follows: $\mathcal{L}_o = [x_{ul_o}, y_{ul_o}, x_{c_o}, y_{c_o}, x_{lr_o}, y_{lr_o}, \vec{d}_o, v_o]_{o=1}^{O}$.

Assuming the number of wireless users in any F-RGB is U, $U \subseteq O$, and a user $u \in U$ is identified from one of all objects (as will be discussed in Section 4.5.2), then the $\mathcal{L}$ vector for that user is represented as $\mathcal{L}_u = [x_{c_u}, y_{c_u}, \vec{d}_u, v_u]$, given that the UE is located in the

[1]The term "Flat" is used to indicate a 2D image that has uniform metric width throughout.

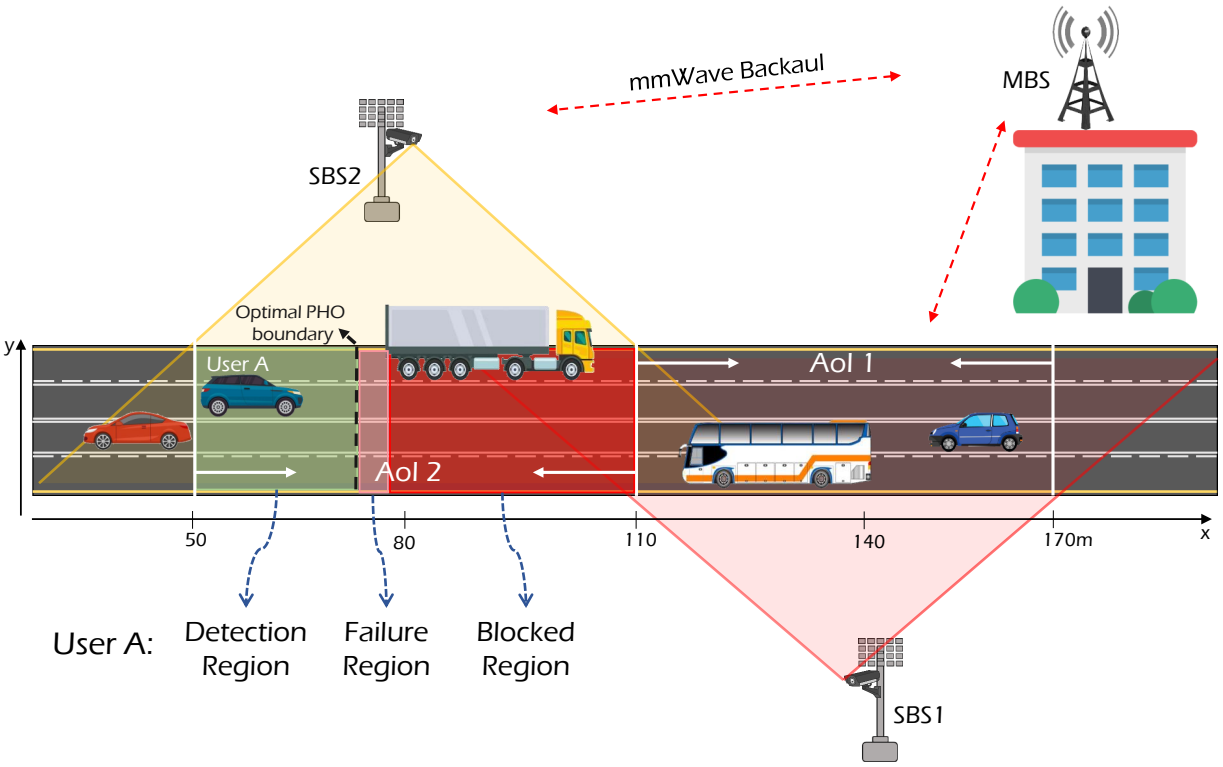


Figure 4.1: The proposed system model: portion of an UDN including one MBS and two SBSs each equipped with a vision sensor.

middle of the object, and other objects are potential blockages. Let $S_{u,o}$ represents the combination of the wireless user and a single object, $S_{u,o} = \{\mathcal{L}_u, \mathcal{L}_o\}, u \in U, o \in O \backslash \{u\}$. Therefore, the goal is to classify whether this sample leads to a possible future blockage $b \in \{0, 1\}$, where 0,1 indicate beam non-blockage or blockage, respectively. Moreover, the study predicts the remaining time until the user gets blocked if a link blockage is expected, denoted as T_b, which could be defined as:

$$T_b = \begin{cases} i \, , b = 1, \ \ \forall i \in \mathbb{R}^+ \\ \text{-1}, b = 0 \end{cases} \tag{4.3}$$

where -1 means not applicable when the sample $S_{u,o}$ does not lead to a future blockage. Thus, $s_{u,o} = \{b_{u,o}, T_{b_{u,o}}\}$ is defined as the labels associated with each data sample $S_{u,o}$.

The objective of this study is achieved by using an ML model denoted as $\Psi_\Theta(S)$, which is capable of performing both classification and regression in parallel. This model takes in the user-obstacle vectors S as input and produces predictions $\hat{s}$. The model's predictions are governed by a set of parameters Θ, which are adapted based on dataset of labelled samples $D = \{S_{u,o}, s_{u,o}\}, u \in U, o \in O \backslash \{u\}$. This dataset is used to train the ML model to reach high accuracy in blockage status classification and accurate time prediction for blockage events. The following mathematical formulas represent the purpose of the model, which aims to maximise the probability of link status prediction while minimising

the prediction error for blockage time.

$$\max_{f_\Theta(S)} \prod_{u=1}^{U} \mathbb{P}\left(\hat{b}_{u,o} = b_{u,o} \mid S_{u,o}\right), \ \forall o \in O \backslash \{u\} \tag{4.4}$$

$$\min_{f_\Theta(S)} \sum_{u=1}^{U} (|\hat{T}_{b_{u,o}} - T_{b_{u,o}}|), \ \forall o \in O \backslash \{u\} \tag{4.5}$$

4.5 CV-assisted Dynamic Blockage Prediction and PHO

4.5.1 Key Idea and Schematic Diagram

This study focuses on a more practical scenario that considers multiple dynamic users and objects, extending the previous work that considers a single user and a stationary blocking object. The framework is divided into several subtasks, as illustrated in the schematic diagram in Fig. 4.2. Initially, the camera at each SBS captures sequences of time-tagged RGB images, which are processed to focus on the respective AoI. Then, one of the leading-edge ODL algorithms is used to recognise objects and extract the necessary augmented information. Next, it becomes essential to differentiate the wireless users from other obstacles to form $S_{u,o}$ data samples. At this point, the data samples are labelled[2] based on blockage status and time. The complete dataset is then stored to train the multi-output model using FL. Once the model is ready, the unlabelled data samples are directly fed into the model for inference. If the predicted T_b ($\hat{T}_b$) is greater than the time required by the proposed framework (T_F), it is highly possible to avoid such blockages by requesting a PHO. The following formulas illustrate the main time parameters of the proposed solution:

$$T_F = T_{ODL} + T_{Inf} + T_{PHO}, \tag{4.6}$$

$$T_D \leq T_b - T_F, \tag{4.7}$$

where T_{ODL} is the time associated with using the ODL algorithm on two successive F-RGB images. T_{Inf} is the model's inference time. T_{PHO} is the time required for performing PHO, and T_D is the time defined to delay triggering the PHO to the point that yields the best QoE.

[2]Labels of data samples can be obtained analytically in the absence of prior information, or through observation by monitoring and recording the users blocking status and time.

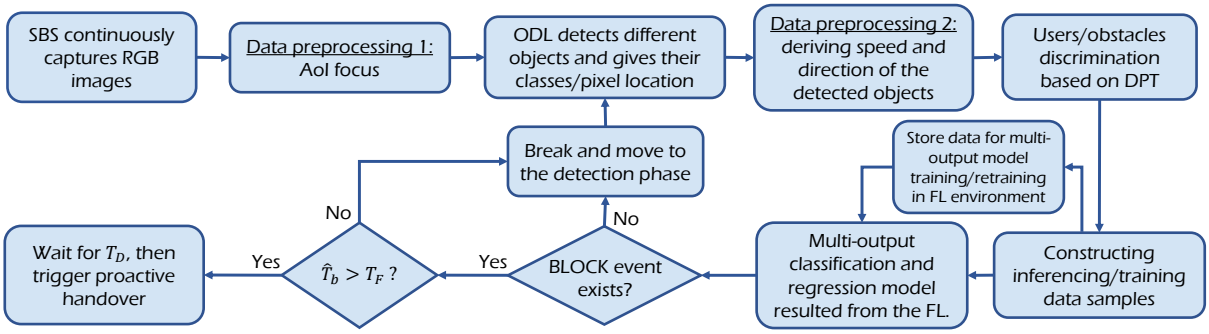


Figure 4.2: Schematic diagram of the proposed framework.

4.5.2 Objects Detection and Users/Obstacles Discrimination

ODL algorithms have recently undergone many advancements, enabling super-fast, real-time, and accurate detection. In this study, a state-of-the-art YOLOv3 algorithm is adopted to detect various objects in the F-RGB images and produce boundary boxes indicating the positions of the objects in pixel scale [130]. These boundary boxes are then converted to metric scale using the conversion ratio $W_m : W_p$, where W_m and W_p refer to the width of F-RGB images in meters and pixels, respectively. This process is followed by extracting objects' speed and direction to build the $\mathcal{L}$ vector for every object. Performing ODL on two successive F-RGB images is essential to determine the speed and direction. The direction is determined by noting the displacement in the x location, whether to the left or the right. This offset distance is divided by the time difference between corresponding timestamps to calculate the object's speed. The study in Chapter 3 shows that performing ODL on two F-RGB images requires 102ms, i.e. $T_{ODL} = 102ms$. This time can be reduced if edge computing resources are utilised in SBSs.

Identifying wireless users: Moving from a single-user scenario to a multi-user environment necessitates distinguishing each particular user from other objects present in the F-RGB image. This study uses a mapping technique wherein the exact location of the wireless user in the environment is reflected onto the F-RGB images and then compared with all boundary boxes. The object possessing a boundary box with a centre closest to the user's location is identified as the wireless user within the F-RGB. Several techniques have been employed to obtain the user's position in the wireless environment, such as GPS and RSS triangulation. However, these methods often fail to provide an accurate localisation. The shift to higher operating frequency is foreseen to improve the positioning accuracy based on the cellular networks [131]. Moreover, several studies have explored this research direction, proposing innovative techniques that offer highly accurate user localisation [132, 133]. Given these developments, this work assumes that the radio access network adopts one of these highly accurate methods to provide the location and track the users. Therefore, this study proposes the DPT within each SBS to continuously

track user locations, which are also converted and reflected on the pixel scale. With DPT tables, it is now possible to differentiate wireless users from other objects and generate the necessary user-obstacle data samples, $S_{u,o}$.

4.5.3 Model Training and Inference: FL Approach

The nature of the defined problem is best solved by a model capable of simultaneously performing both classification and regression tasks. Hence, this work develops a multi-output two-hidden layer NN model. This model is fed by user-obstacle samples to predict both blockage status and time. In addition, this study adopts FL rather than centralised learning to protect the privacy of the data and relieve the pressure on communication channels.

Training Phase: The proposed framework requires a well-trained model before execution. During the FL process, the NN model is used as the base model for training across SBSs. The number of participating clients is set to three, although the framework can be easily extended to incorporate a larger number of SBSs if needed. A parameter server situated in the MBS orchestrates the training process by selecting the number of SBSs participating in each round and sends them the model to start the training. Each SBS then utilises its respective dataset to conduct local model training. Subsequently, each SBS then sends the model's parameters to the parameter server for aggregation. Furthermore, an early stopping patience technique is developed to avoid suboptimal performance or excessive rounds of unnecessary training.

Inference Phase: Once the model has been trained in the FL environment, it becomes prepared for inference. The user-obstacle data samples that have been generated can then be input into the trained model to predict blockage status and time. Furthermore, the time associated with model inferencing requires approximately 1 ms, i.e., $T_{Inf} = 1$ms.

4.5.4 Optimal PHO Trigger Point

The proposed framework alerts the network to initiate PHO when a blockage is anticipated. The principal question revolves around determining when and at what distance the PHO should be triggered. Similar to the previous study, this work identifies two key regions: the detection region and the failure region. The detection region is defined as the region where the proposed framework actively monitors and detects objects. In this system model, this region is the same as the AoI. On the other hand, the failure region marks the area where there is minimal possibility of avoiding link interruption, primarily due to insufficient time remaining to complete the PHO. The failure region is located just before the blocked

region. Its width and location vary because of the environment's dynamicity and the user's speed. Fig. 4.1 illustrates these regions.

Determining the optimal distance at which to trigger a PHO is closely tied to ensuring high levels of the user's QoE. If a BLOCK event is detected, the optimal strategy involves waiting for the maximum delay (T_D^{max}) obtained from equation (4.7) before executing the PHO. Translating this time value into a distance gives the optimal distance (Λ_{opt}) defined as follows:

$$\Lambda_{opt} = v \times T_D^{max}. \tag{4.8}$$

In other words, the optimal PHO must be triggered just before reaching the failure region in the boundary between the detection and failure regions. While triggering the PHO anywhere within the detection region and before the optimal PHO boundary could potentially avoid the link blockage, it might come at the expense of compromising the perceived QoE. In addition, doing an early PHO may impact the balance and the allocation of network resources. Hence, the objective of the proposed framework is to always trigger the PHO within the vicinity of the optimal PHO boundary.

4.5.5 PHO Latency

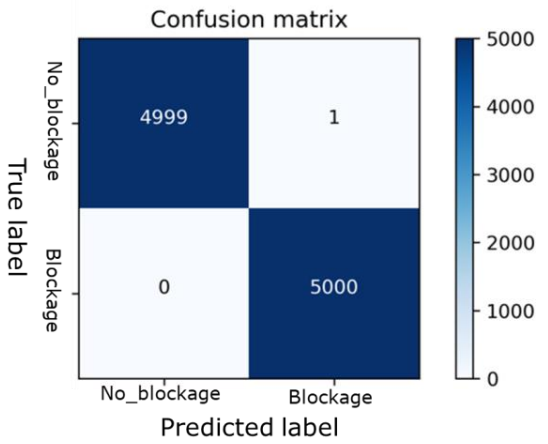
The final essential parameter for the proposed framework is T_{PHO}. In a conventional network employing beamforming, when a user loses the connection with the SBS, several steps are taken to reestablish the connection. These steps are beam failure detection, beam failure recovery, cell search, and contention-based/free random access [114]. According to the 3GPP specifications, each step is associated with a time duration until completion. The cumulative delay time is ∼312.2 ms, indicating the time associated with performing reactive HO when contention-based random access is assumed. By employing proactive blockage prediction, this time can be significantly reduced. For instance, in the context of this study targeting urban areas with dynamic wireless environments, where contention-based radio access is assumed, the T_{PHO} is estimated to be 80ms [134]. Consequently, the proposed framework's T_F is now determined and equals 183ms. If the predicted T_b is greater than 183ms, the framework has a high chance of avoiding link interruption. Otherwise, a link interruption will happen.

4.6 Performance Evaluation and Results

This section first discusses how the NN model is developed and then delves into describing the simulation setup, followed by a discussion of the simulation results.

Table 4.1: Hyperparameters of the NN model.

Parameter	Value
Layers information	Input layer: 10 features 1st Hidden layer: 128 neurons, ReLU 2nd Hidden layer: 64 neurons, ReLU 1st Output (Classifier): 2 neurons, Softmax 2nd Output (Regressor): 1 neuron, Linear
Loss functions	MAE, sparse_categorical_crossentropy
Optimiser	Adam
Learning rate	0.001
Batch size	1000
Data split	70% training - 30% testing



T_b categories

Category	Samples	Blockage?
1) Actual $T_b < T_F$	230	✓
2) Predicted T_b <= Actual T_b	2695	✗
3) Predicted T_b > Actual T_b	2075	✓

Figure 4.3: Classification and regression model performance.

4.6.1 FL-based Multi-output Model Development

The NN model is trained under the FL setup using the FedAvg algorithm. The complete information about the model structure and the selected hyperparameters can be found in Table 4.1. Model performance is tested using ten thousand samples forming 50% blocking and 50% nonblocking. Fig. 4.3 displays the testing results in which the confusion matrix demonstrates the near-optimal classification accuracy of the model. Additionally, the table divides the T_b into three categories, providing the corresponding blockage status for each category.

To give a better understanding, the PHO success rate is defined as $\mathcal{S}_{PHO} = N_S/N_T$, where N_S denotes the number of samples with successful PHO, and N_T indicates the total number of blocking samples. Therefore, it can be concluded from the table in Fig. 4.3 that the success rate is unsatisfactory at 54%. However, this result can be improved by making a trade-off between the $\mathcal{S}_{PHO}$ and the QoE. To achieve this, a new parameter called percent shift ($\mathcal{P}_{Shift}$) is introduced. The $\mathcal{P}_{Shift}$ is defined to reduce the predicted T_b by this percent. This parameter aims to move as many samples as feasible from the

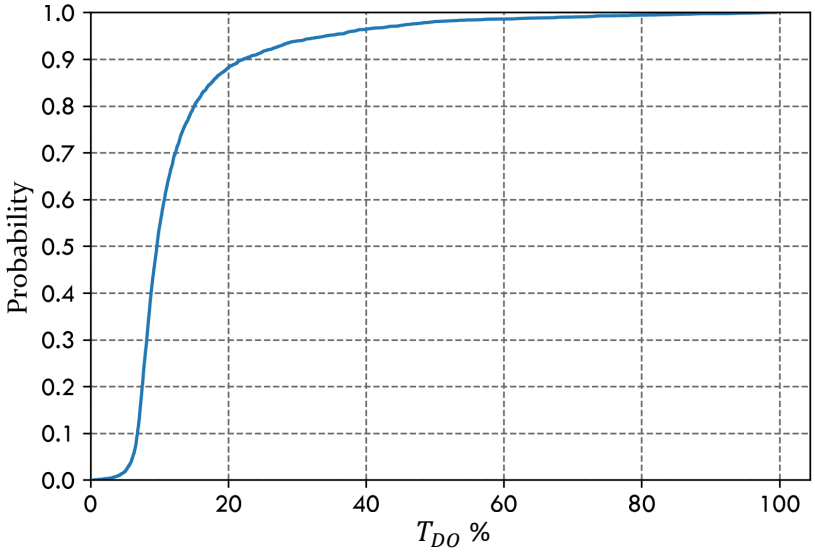


Figure 4.4: The distribution of the TD offset of the samples with successful PHO.

Table 4.2: $\mathcal{S}_{PHO}$ versus $\mathcal{P}_{Shift}$.

$\mathcal{P}_{Shift}$ %	0	1	3	5	7	9	11	13	15
$\mathcal{S}_{PHO}$ %	54	77.1	91	93.4	93.6	93.3	92.8	92.4	92

third to the second category, thereby improving the $\mathcal{S}_{PHO}$ at the cost of a slight decrease in the QoE. Table 4.2 provides insight into how altering the values of the $\mathcal{P}_{Shift}$ influences the $\mathcal{S}_{PHO}$, with the best value being identified as 7%. Accordingly, Fig. 4.4 illustrates the cumulative distribution function (CDF) of the samples with successful PHO in relation to T_D offset, desfined as (T_{DO}). This offset indicates the disparity between the predicted T_D ($\hat{T}_D = \hat{T}_b - T_F$) and the actual one, defined as:

$$T_{DO} = \frac{T_D^{max} - \hat{T}_D}{T_D^{max}} \times 100\%, \quad \forall \hat{T}_D \leq T_D^{max}. \tag{4.9}$$

The optimal PHO point occurs when the T_{DO} equals zero. The closer the samples to this point signifies improved performance. Moreover, deviating from this point means initiating PHO earlier, which could potentially affect the QoE. However, performing earlier PHO with some QoE reduction is preferable to experiencing a complete loss of connection followed by re-establishment. The later scenario would incur much overhead and increase network latency. Finally, the framework is now ready to be used under the considered scenario that will be discussed subsequently.

4.6.2 Simulation Setup

The overall performance of the proposed framework is evaluated by considering a practical outdoor environment. The scenario considered in this study is inspired by the ViWi ASU

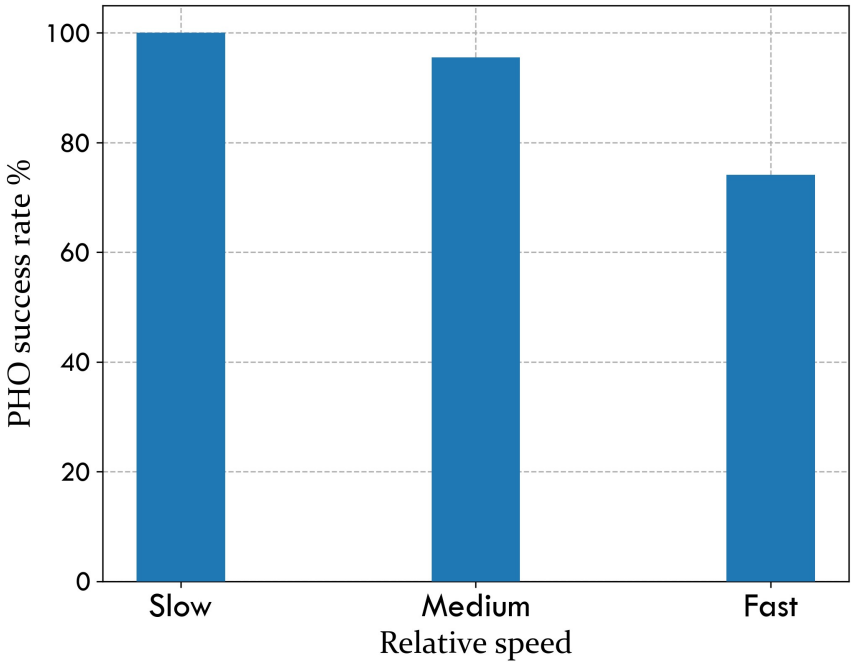


Figure 4.5: The impact of different relative speeds on the PHO success rate.

downtown scenario (ASUDT1) [135], which closely resembles the system model adopted for this study. ASUDT1 comprises two mmWave SBSs operating at 28GHz and located 60 m apart on opposite sides of the street. Each has an antenna array that forms LoS beams to serve 60 users moving in straight trajectories. Users are UEs placed in the center of vehicles of different sizes, such as cars, buses, and trucks. These vehicles move at different speeds and directions, and each can be seen as a potential obstacle for other users. At each time instance, referred to as a scene, ASUDT1 provides raw data for every user u consisting of a 4-tuple of concurrent information including user location, RGB images, mmWave channel, and link status from each SBS. This dataset helps in the evaluation of proposed solution's performance. The simulation experiments are conducted using Python programs, with the key performance metrics are the PHO success rate, network latency, and the perceived QoE.

4.6.3 Simulation Results

Several aspects are considered to examine the efficacy of the proposed framework. First, given the dynamic nature of the considered environment, the impact of vehicles' speed on performing a successful PHO is studied. The speed of the vehicles is set within the range of 1.5 to 20 mph. In this context, a new parameter called *relative speed* is introduced. For every blocking sample in $S_{u,o}$, the relative speed parameter is defined as follows: if the user and the obstacle are moving towards each other, the relative speed is computed as the sum of their speeds; if they are moving in the same direction, it is calculated as the

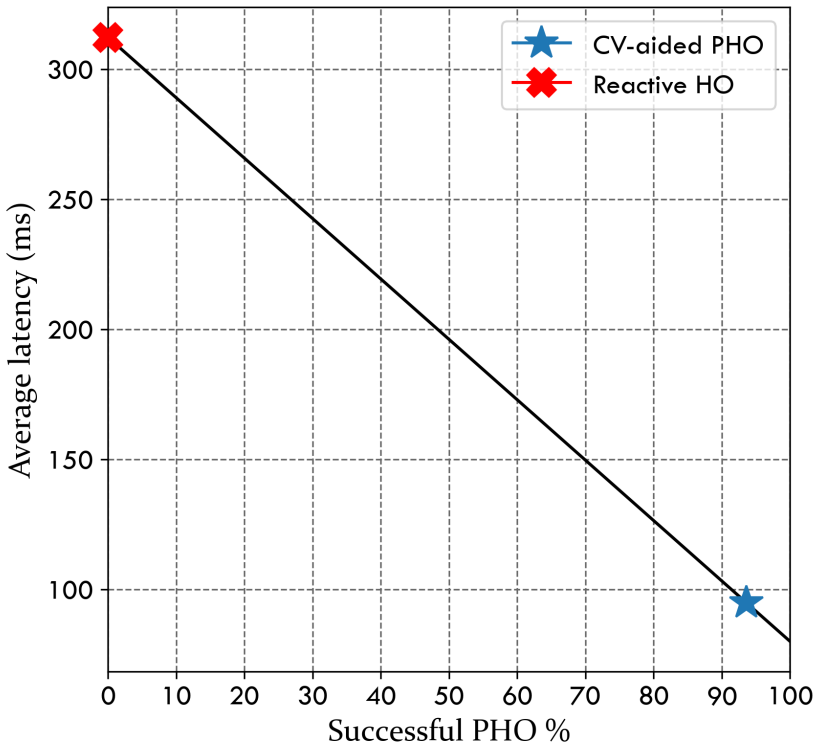


Figure 4.6: Comparison of the average latency between the reactive HO and the proposed CV-aided PHO.

difference in their speeds. This parameter essentially quantifies how fast a blockage occurs and will be used to investigate its impact on performing successful PHO. Consequently, the relative speed derived from all blocking samples within the testing dataset is divided into three groups: slow, medium, and fast based on their mean and standard deviation. Samples with a relative speed below 10 mph are classified as slow, those with a relative speed between 10 mph and 29 mph are categorized as medium, and those exceeding 29 mph are labeled as fast. Fig. 4.5 demonstrates the results of this study. It can be observed that the PHO success rate is high when the relative speed is low and medium. Conversely, the success rate decreases as the relative speed increases. This outcome is in line with expectations, given that higher relative speeds lead to reduced T_b values for blockages, consequently diminishing the likelihood of a successful PHO.

Next, the latency associated with performing HO on both reactive- and proactive-based HO approaches is investigated. Section 4.5.5 delves into this analysis, revealing that the latency linked with reactive HO amounts to approximately 312.2 ms. In contrast, the PHO entails a mere 80 ms, assuming contention-based random access. Following a similar approach from [114], the average HO latency for 5000 users is calculated as:

$$\zeta = \frac{\{\epsilon \times U\} \times 80 + \{(1-\epsilon) \times U\} \times 312.2}{U}, \tag{4.10}$$

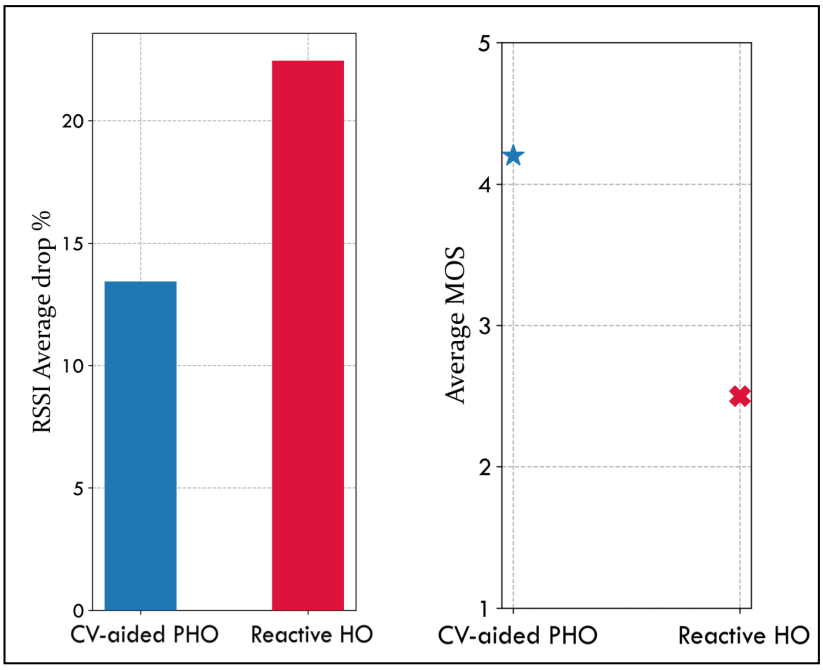


Figure 4.7: The RSS percentage drop due to performing reactive and PHO, and how much this drop affects the QoE measured through MOS.

where U signifies the total number of users, and $\epsilon \in [0, 1]$ is the proportion of users who successfully performed PHO. Fig. 4.6 shows the average latency improvement achieved by the proposed PHO framework compared to the reactive mechanism. The calculated average latency is 94.8 ms for the proposed CV-aided PHO solution, which outperforms the reactive HO approach by a factor of 3.3. This improvement is pivotal for maintaining uninterrupted connectivity in real-time applications.

The final study adopts a similar approach from the previous work of considering moving users running an real-time protocol (RTP)-based application and measures the average QoE/MOS of a group of users with prospect blockages. This study takes advantage of the mmWave channel information provided by the ASUDT1 scenario. Since ASUDT1 provides plentiful information represented as scenes for every location point, this study spotlights on the portion of the street between the two SBSs and only focuses on the blockages within the scene interval from 680 to 980. For every blocking, the RSS percentage drop when a user is handed over to another SBS is recorded and is done for all users who encounter blockages in between the two SBSs. Then, the average percentage drop of the RSS is calculated and mapped to the corresponding value of the MOS. However, in the case of the reactive-HO approach, no proactive measures are undertaken. In this scenario, the percentage drop in RSS is measured to find the average RSS percentage drop. Fig. 4.7 illustrates the outcome of this study in which the proposed framework can keep the MOS at a high level despite the small drop in the average RSS. On the contrary, the reactive HO fails to keep users at high MOS during interruption time. This result confirms the

potential of the proposed framework for improving the reliability of high-frequency wireless networks, rendering them more suitable for latency-sensitive applications.

4.7 Summary

This study explored the potential of leveraging vision information to improve the reliability of high-frequency networks by predicting dynamic blockages in advance and taking measures to perform PHO. A NN multi-output model is developed that, combined with CV technology, propose a novel framework capable of accurately predicting blockages and the time needed before the user reaches the blocked region. Moreover, the model is trained using FL to protect data privacy and conserve bandwidth resources. Simulation results indicated that the proposed framework achieves a high PHO success rate of 93.6%, outperforms the reactive-HO approaches by a factor of 3.3 in terms of latency, and maintains the QoE at higher levels. These results highlight a promising solution for beam blockages in multi-user mmWave/THz networks. However, it is important to note that the performance of CV-aided PHO frameworks may be impacted by low-light conditions and adverse weather. Therefore, the subsequent chapter will explore an alternative approach involving radar sensors to provide more enhanced proactive beam blockage solution.

Chapter 5

Radar-aided Dynamic Blockages Recognition in UDNs with Distributed Learning

The preceding two chapters have effectively addressed the beam blockage problem by proposing CV-aided PHO frameworks applicable in diverse environments. However, the deployment of vision sensors faces challenges, including regulatory constraints and compromised image quality in low-light or adverse weather conditions. This chapter addresses these issues by proposing RaDaR, a framework that leverages radar sensors to monitor the surrounding environment and identify potential blocking objects in advance. Moreover, the FL distributed learning approach is employed, providing the framework with advantages such as scalability, knowledge sharing, and resource efficiency. This aligns with the aims specified in *O3*. Considering critical metrics like RSS, latency, and throughput in UDNs, simulation results affirm that RaDaR, with its blockage-aware approach, enhances the QoE for users in highly dynamic environments with low-latency requirements.

5.1 Introduction

Incorporating CV technology to predict the forthcoming beam blockages has introduced innovative frameworks that significantly enhance the operation of high-frequency networks. Nevertheless, deploying vision sensors may not always be feasible due to regulatory and privacy concerns. Moreover, image quality may be affected by low-light and adverse weather conditions. These restrictions have spurred a new research direction that involves the adoption of radar technology for predicting link blockages, achieved by detecting obstacles through radar fingerprints. The reasons for this shift to radar-based solutions are due to their low cost, ability to capture useful object features like range, velocity, and direction. In addition, radar technology offers lower privacy risk and, most importantly,

enables low-latency transmissions as it operates at high-frequency bands. To date, few studies have considered the use of radars to address beam blockages problem in high-frequency networks [136, 137]. The work in [136] is leading in harnessing radar sensors to enhance the reliability of mmWave communication systems in indoor environments. The study proposed a concept, called radar-based medium access (RadMAC), which leverages reflected radar signals to enable intelligent beam-steering decisions based on blockage prediction and avoidance. Likewise, in [137], a mmWave BS was integrated with a low-cost frequency-modulated continuous wave (FMCW) radar to gather information such as range and velocity for obstacle prediction. To enhance the network's intelligence and enable it to anticipate possible beam blockages up to one second ahead, a combination of CNN and LSTM models were trained based on radar range-angle maps. However, these studies are preliminary and did not adequately consider the dimensions of obstacles, which are crucial factors in determining the potential blockage of an LoS link. Additionally, existing studies overlooked the significance of the height of both the BS and the UE, which significantly influence the status of the mmWave channel. Consequently, further research is required to fully explore the potential of using radars to proactively predict blockages and achieve promising reliability and latency gains for the practical realisation of next-generation networks.

5.2 Contributions

This chapter introduces a novel framework named radar-aided dynamic blockages recognition (RaDaR), developed to handle the challenge of frequent beam blockages in high-frequency outdoor networks. RaDaR employs radar sensors to enhance the network's situational awareness by monitoring and tracking the movement of objects to generate range-angle and range-velocity maps. These maps are useful for analysing the scene and making accurate blockage predictions. Unlike previous related works, RaDaR leverages radar information to predict the height of objects, a critical factor in determining whether an object will obstruct the LoS link or not. Additionally, the framework's end-to-end execution time is measured to provide the network with a proactiveness merit in predicting blockages and performing PHO, to avoid link interruption and ensure high QoE. In addition, RaDaR incorporates the FL training mechanism, which brings three important benefits: scalability, knowledge sharing, and resource efficiency. As a result, RaDaR supports network expansion to new development areas and consolidates knowledge from multiple SBSs by performing distributed learning rather than centralised learning. This consequently reduces the overhead on the network's transmission resources. Benefiting from the real-world DeepSense dataset [138], a set of non-independent and identically distributed (non-IID) features is extracted to collaboratively train a dual-output NN model

offline using a group of SBSs. The trained model is then used online at each SBS to predict the occurrence of BLOCK events and the remaining time before obstacles obstruct the link. With this information, RaDaR can preemptively decide the optimal instant to switch the beam or initiate PHO to maintain the users' QoE as high as possible. The contributions of this work can be summarised as follows.

- A novel framework, called RaDaR, is presented, aiming to improve the reliability of federated mmWave networks by integrating radars for the anticipation of LoS link blockages while considering latency and QoE metrics.
- The FL algorithm is employed to facilitate collaborative model training at each SBS by using information acquired from radars placed at the top of the SBSs. FL provides the framework with vital features, including scalability, knowledge sharing, and conserving network resources.
- The effectiveness of the proposed framework is evaluated using the large-scale real-world DeepSense dataset. In particular, Scenario 30 is augmented, and diverse environments are created to reflect practical scenarios.

5.3 System Model

The considered system comprises several SBSs and stationary[1] users in a vehicular environment, as shown in Fig. 5.1. The SBS and users are equipped with a GPS that is supported by a real-time kinematic network. This enables accurate three-dimensional geolocation with sub-centimetre precision [139]. Moreover, the SBS is equipped with two primary components: (i) a phased array antenna that produces LoS beams to serve the user and (ii) an FMCW radar mounted at the SBS to detect and track mobile objects in the operating vicinity. In the following subsections, the signalling models are explained for the network and radar components.

5.3.1 Channel and Blockage Models

The considered system operates at 60 GHz using OFDM with K subcarriers and a cyclic prefix of length Q. Each SBS is equipped with a ULA consisting of M antenna elements. These elements are utilised to produce narrow directive beams that maximise the receive beamforming gain at single-antenna UEs. Moreover, it is assumed that each SBS has a predefined beamforming codebook $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^{B}$, $\mathbf{f}_i \in \mathbb{C}^{M\times 1}$, where B denotes the total

[1]The term 'stationary' is used to indicate that the user will remain within the effective coverage range of the beam.

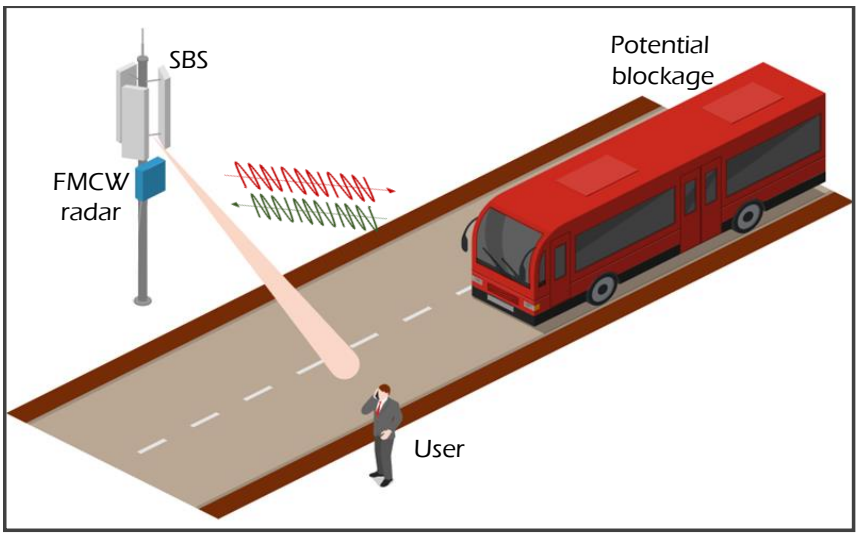


Figure 5.1: The proposed system model of the RaDaR framework.

number of beams in the codebook. Any beamforming vector can be represented as:

$$\mathbf{f}_i = \frac{1}{\sqrt{M}} \left[1 \; e^{j\frac{2\pi}{\lambda} d \sin(\psi_i)} \; \ldots \; e^{j\frac{2\pi}{\lambda}(M-1) d \sin(\psi_i)} \right]^T, \tag{5.1}$$

where $\frac{1}{\sqrt{M}}$ represents the normalisation factor, d denotes the distance between adjacent antenna elements, λ is the wavelength of the carrier frequency, $\psi_i \in \{\frac{2\pi i}{B}\}_{i=0}^{B-1}$ is the steering angle, and T indicates the transpose notation.

The objective of the network is to determine the optimal beam $\mathbf{f}^\star$ that yields the highest RSS at the UE. To accomplish this, the SBS acquires a pilot message from the UE and uses it to train all the beams in the codebook to identify the best beam. Once selected, the received signal at the UE side on the kth subcarrier can be represented as:

$$y_k = \sqrt{\mathcal{E}_k} \mathbf{h}_k^H \mathbf{f}^\star s_k + n_k, \tag{5.2}$$

where $\sqrt{\mathcal{E}}$ is the transmitter gain, $\mathbf{h} \in \mathbb{C}^{M \times K}$ indicates the narrow band channel between the SBS and the UE, $(\cdot)^H$ denotes the Hermitian transpose, s is the transmitted data symbol, and $n \sim \mathcal{CN}\left(0, \sigma^2\right)$ is AWGN with zero mean and σ^2 variance.

Assuming that the multi-path components of the signal arrive at the receiver through P distinct paths. The shortest path is the LoS path, which is the path of interest in this study and is denoted by $p^\star$, and the other paths represent the NLoS components. Hence, the channel between the transmitter and the receiver can be mathematically expressed as:

$$\mathbf{h}_k = \mathbf{h}_k^{\text{LoS}} + \mathbf{h}_k^{\text{NLoS}}, \tag{5.3}$$

where the LoS and NLoS channels are given respectively as follows [140]:

$$\mathbf{h}_k^{\mathrm{LoS}} = \sum_{q=0}^{Q-1} \alpha_{p^\star} e^{-\mathrm{j}\frac{2\pi k}{K} q} \ell\left(qT_s - \tau_{p^\star}\right) \mathbf{a}\left(\theta_{p^\star}, \phi_{p^\star}\right), \tag{5.4}$$

$$\mathbf{h}_k^{\mathrm{NLoS}} = \sum_{q=0}^{Q-1} \sum_{p=1}^{P\setminus\{p^\star\}} \alpha_p e^{-\mathrm{j}\frac{2\pi k}{K} q} \ell\left(qT_s - \tau_p\right) \mathbf{a}\left(\theta_p, \phi_p\right), \tag{5.5}$$

where α_p, τ_p, θ_p, ϕ_p are the gain, delay, azimuth, and elevation angles of the arrival of path p, respectively. Here, $\ell(\cdot)$ represents the pulse-shaping filter. T_s is used to denote the sampling time, and $\mathbf{a}$ is the receiver array steering vector [141].

This study employs a blockage fading channel model to include the impact of blockages on the communication system. Let b denote the LoS blockage status and is defined as:

$$b[t] = \begin{cases} 1, \text{ LoS path is blocked} \\ 0, \text{ LoS path is not blocked.} \end{cases} \tag{5.6}$$

Therefore, the mmWave channel for any subcarrier k at the time instant $t \in \mathbb{Z}^+$ can be updated as follows:

$$\mathbf{h}_k[t] = (1 - b[t])\, \mathbf{h}_k^{\mathrm{LOS}}[t] + \mathbf{h}_k^{\mathrm{NLOS}}[t]. \tag{5.7}$$

It is noteworthy that, in practical scenarios, NLoS signals exhibit a relatively negligible impact on high-frequency communication systems due to their considerably lower signal-to-interference-plus-noise ratio (SINR), rendering them less desirable for robust and reliable communications [142, 143]. Furthermore, these systems naturally have few NLoS links with significantly inferior channel gains compared to LoS counterparts, even in the presence of blockages [144]. Thus, it is reasonable to assume that wireless networks in mmWave and THz bands predominantly hinge on LoS beam communications for the delivery of reliable and highly efficient data transmission.

5.3.2 Radar Model

In this system, an FMCW radar is installed in each SBS to obtain measurements of the surrounding environment and leverage them to develop a proactive mechanism for predicting potential network blockages. In each measurement, the radar transmits a frame of C frequency-modulated chirps that represent continuous waves of radio signals separated by pause time τ_P. Each chirp (a.k.a. ramp) has a linearly varying frequency that starts from f_c and ends with $f_c + \mu t$, given as [137]:

$$X_{chirp}(t) = A_t \exp\left(j\left(2\pi f_c t + \pi \mu t^2\right)\right), 0 \le t \le \tau_c \tag{5.8}$$

where A_t denotes the transmitter gain, $\mu = B_c/\tau_c$ is the slope of the chirp signal, which has bandwidth B_c and duration τ_c. Upon transmitting all chirps, the radar system remains inactive until the initiation of a new frame. However, during the frame time, the radar receives the signals that are reflected by the target objects. The received signals are subsequently directed to a quadrature mixer, which combines the transmitted and reflected chirps to generate in-phase and quadrature components. The mixed signals are then processed by a low pass filter to produce intermediate frequency (IF) signals. The IF signal captures the variations in frequency (a.k.a. beat frequency) and the phase between the transmitted and reflected signals. The IF signal can be mathematically represented as [145]:

$$Y_{IF}(t) = A_t A_r \exp\left(j\left(2\pi f_c \tau_{rt} + 2\pi\mu\tau_{rt}t\right)\right), 0 \leq t \leq \tau_c \tag{5.9}$$

where A_r is the receiver gain, $\tau_{rt} = 2r/c$ is the round-trip delay of the radar signal reflected from the object, which depends on the distance r between the object and the radar, as well as the speed of light c.

The IF signal then undergoes analog-to-digital converter (ADC) and is subsequently sampled at a rate of f_s, producing S samples per chirp. Assuming the radar is fitted with M_r receive antennas, the number of samples per measurement will be $M_r \cdot S \cdot C$. These samples are represented as $\mathbf{R} \in \mathbb{C}^{M_r \times S \times C}$, and they constitute the fundamental information used to infer object-related information in RaDaR.

Range and velocity calculation: Suppose an object has a time-varying distance $r(t) = r_0 + x(t)$, where $x(t) = vt$ is a function denotes the distance variation of an object moving at v velocity. Thus, the round-trip time can be written as:

$$\tau_{rt} = \frac{2r_0 + 2vt}{c}. \tag{5.10}$$

After substituting (5.10) in (5.9) and performing some mathematical manipulations, the IF signal can be written as follows:

$$Y_{IF}(t) = A_t A_r \exp\left(j\frac{4\pi}{c}\left(\mu v t^2 + \mu r_0 t + \frac{v}{\lambda_R}t + \frac{r_0}{\lambda_R}\right)\right), \tag{5.11}$$

where λ_R denotes the wavelength that corresponds to the operating frequency of the radar. Note that the first and last terms of the exponent have limited usefulness in extracting range and velocity information. The first term is very small, and the last term remains constant. In contrast, the second term provides valuable range information, while the third term enables the extraction of velocity information.

After sampling the IF signal with the ADC converter, a fast Fourier transform (FFT) is applied along the time samples direction, referred to as *Range*-FFT, to determine the range

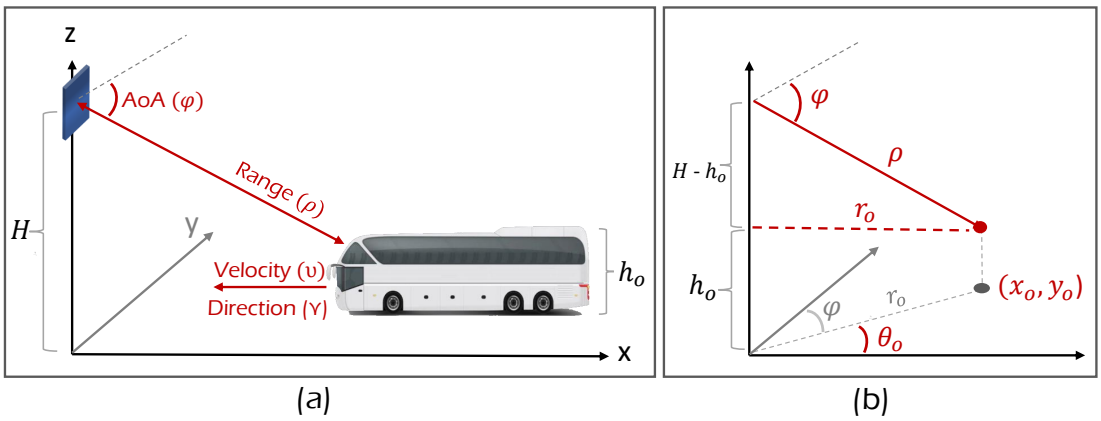


Figure 5.2: (a) Radar is used to gather situational information, (b) this information is used to generate the localisation vector of the object.

of the object[2]. The peak of the power spectral density output reveals the range information of the target object. Additionally, performing a second FFT on the chirp samples, known as *Velocity*-FFT, helps determining the target object's velocity by observing the peak of the output spectrum.

Angle and direction estimation: The utilisation of MIMO antennas in radar systems enables the estimation of the angle of arrival (AoA). By implementing an additional FFT in the direction of the receive antenna samples, referred to as *Angle*-FFT, the angular information can be extracted. Specifically, the variation in distance between the object and each receiving antenna causes a phase shift in the FFT peak, which corresponds to angular information. On the other hand, the object movement direction can be easily identified by observing changes in the AoA, or alternatively by checking which of the in-phase or quadrature components of the complex beat signal is leading in phase. Finally, performing range, velocity, and angle FFTs would result in *radar cube*, which can be viewed as the stack of range-angle maps of each velocity value.

5.4 Problem Description and Formulation

The main objective is to utilise radar measurements $\mathbf{R}$ for detecting objects and predicting forthcoming LoS beam blockages. This decision is facilitated by knowing the location information of the users and the potential obstacles. Once a stationary user u is connected to the network, the SBS activates the radar sensor to monitor the surrounding area. Assuming that a number of objects denoted as O are detected in the ith measurement $\mathbf{R}_i$. Each object $o \in O$ will be continuously monitored until it exits the radar's field of

[2]Multiple ranges can be obtained for an object, but typically, the shortest one, often generated by the nearest upper edge, is considered

view (FoV). During this monitoring process, the object's situational parameters, which include its range ρ, velocity υ, AoA φ, and direction Υ described in Fig. 5.2(a) are extracted. Based on the height of the radar H and the height of the object h_o, which will be discussed in Section 5.5.1, the situational information is utilised to construct a six-dimensional localisation vector $\mathcal{L}$ for the object, given by $\mathcal{L}_o = [r_o, x_o, y_o, \theta_o, v_o, \vec{d_o}]$, where:

$$\begin{aligned} r_o &= \sqrt{(\rho^2 - (H - h_o)^2)} \\ x_o &= r_o \sin(\varphi) \\ y_o &= r_o \cos(\varphi) \\ \theta_o &= \tan^{-1} y_o / x_o \\ v_o &= \upsilon \\ \vec{d_o} &= \Upsilon. \end{aligned} \tag{5.12}$$

These features are derived by placing the SBS at the Cartesian origin and considering the x-y plane. Specifically, r_o is defined as the distance from the SBS to the object, and x_o and y_o as the object's x and y coordinates, respectively. Additionally, θ_o is used to represent the angle between the positive x-axis and the line passing through (x_o,y_o) and the origin. The object's speed is denoted by v_o, and $\vec{d_o}$ signifies the direction of its movement, which can be either "left" or "right". Please refer to Fig. 5.2(b) for further illustration.

Given a set of stationary users U, each user $u \in U$ is assigned a four-dimensional localisation vector denoted as $\mathcal{L}_u = [r_u, x_u, y_u, \theta_u]$. These features are similar in their definition to those related to the detected object and are obtained by converting the user's GPS information to Cartesian using the universal transverse Mercator tool [146]. Then, for each user u, $S_{u,o} = \{\mathcal{L}_u, \mathcal{L}_o\}$ is generated, with the objective of classifying whether this data sample results in a future blockage $b \in \{0, 1\}$. 0,1 denote beam non-blockage or blockage, respectively. Moreover, the remaining time until the obstacle obstructs the LoS connection is estimated and denoted as T_b and is defined as:

$$T_b = \begin{cases} i \ , b = 1, \ \ \forall i \in \mathbb{R}^+ \\ -1, b = 0, \end{cases} \tag{5.13}$$

where -1 indicates that the value is not applicable due to the absence of potential blockage. Therefore, $s_{u,o} = \{b_{u,o}, T_{b_{u,0}}\}$ is defined as the labels associated with each data sample $S_{u,o}$.

The user/object localisation information could be leveraged to intelligently handle channel disruptions and enhance network reliability. To formulate that, the objective is to design an ML model, denoted as $\Psi_\Theta(S)$, that can simultaneously perform classification and regression. This model takes in user-object data samples S and generates prediction $\hat{s}$. These predictions are governed by a set of parameters Θ adapted based on a labeled

dataset $D = \{S_{u,o}, s_{u,o}\}$. The model aims to maximise the probability of accurately predicting link disconnections while minimising the error associated with predicting the blockage remaining time, as given below:

$$\max_{\Psi_\Theta(S)} \prod_{u=1}^{U} \mathbb{P}\left(\hat{b}_{u,o} = b_{u,o} \mid S_{u,o}\right), \ \forall u \in U, \forall o \in O \tag{5.14}$$

$$\min_{\Psi_\Theta(S)} \sum_{u=1}^{U} (|\hat{T}_{b_{u,o}} - T_{b_{u,o}}|), \ \forall u \in U, \forall o \in O \tag{5.15}$$

5.5 The Proposed RaDaR Framework

This section presents RaDaR, an ML-based approach for predicting beam blockages in B5G and 6G networks using radar data. RaDaR utilises a dual-output NN model trained on bundles of radar measurements to provide the system with real-time blockage handling intelligence, thereby improving the performance of next-generation networks. In the next subsections, an in-depth explanation of the proposed solution is provided.

5.5.1 Overview and Schematic Diagram

This study focuses on practical communication systems, particularly in the context of UDNs. UDNs densify SBSs and LoS links per unit area. However, downscaling communication systems have complicated the challenge of mobility, particularly in dynamic areas like smart cities. SBSs use narrow directive beams to connect users to the network, making the UE's position relative to the SBS critical for service continuity. However, the presence of mobile objects can obstruct LoS links, leading to fluctuations in data rates. Conventional wireless networks can only detect the presence of blockages when the user's throughput fluctuates or when the link is disconnected. This concludes that the network is reactive to blockage events, resulting in poor performance. To overcome this challenge, wireless networks must shift from reactiveness to proactiveness by having a sense of their surroundings. Proactiveness must be integrated as a key dimension using existing sensing modalities to improve wireless networks. Hence, radar sensors are deployed, with each SBS equipped with an FMCW radar to monitor the coverage area. The information obtained from radars is vital in dealing with link blockages and controlling the communication system. Fig. 5.3 illustrates the proposed framework's schematic diagram, which consists of three main phases: obstacles detection phase, training and inferencing phase, and PHO decision phase.

The obstacles detection phase

Not every object detected by the radar will necessarily act as a blockage disrupting the communication channel. The occurrence of a blockage event in the system model is highly dependent on the position of the antenna array, the UE, and nearby objects, the direction of object movement, and the dimensions of the objects, especially their height. Therefore, it is imperative to accurately identify actual obstacles to prevent unnecessary HOs. To achieve this, when a stationary user is connected to the network, the serving SBS uses the 3D location information of the user (x_u, y_u, z_u) and the antenna array (x_A, y_A, z_A) to determine the formula of a 3D line crossing them. This equation can be expressed as a vector equation in mathematical terms as follows:

$$\begin{aligned} &< x, y, z > = < x_A, y_A, z_A > + \beta < \varkappa, \iota, \varsigma > \\ &(\varkappa, \iota, \varsigma) = (x_u - x_A, y_u - y_A, z_u - z_A) \end{aligned} \tag{5.16}$$

where β is a parameter describing a particular point on the line, and $< \varkappa, \iota, \varsigma >$ is the direction vector. At the same time, the SBS activates the radar. The radar collects and forwards the reflected signals to a processing unit, where the FFT process is executed to generate a radar cube. Upon detecting an object, the framework utilises the radar cube information to determine the object's range, velocity, AoA, and movement direction. Subsequently, the framework must establish the object's localisation vector. However, the object's height remains unknown, a key parameter for constructing the object's localisation vector, as highlighted in (5.12).

To determine the height of a detected object, one of the several studies available in the literature that tackled this issue through radar technology is employed [147,148]. The most appropriate technique involves using radar fingerprints and residual networks (ResNets) to classify objects in real time. Then, this technique determines the objects' heights by referencing a predefined table that includes the dimensions of each classified object [148]. This study is primarily concerned with predicting the timing of LoS link blockage, a task that can be accomplished by knowing the range and height of an object. While comprehensive object dimensions, such as length and width, are of lesser importance for this purpose, they will become crucial for estimating the duration of link blockages [149]. Investigating these full object dimensions represents a potential area of research expansion in the future, building upon this work. It is assumed that the framework adopts the aforementioned approach to determine the height of the detected object, which, in turn, is used to generate the object's localisation vector. Subsequently, the framework generates the plane $y = y_o$ and calculates the point of intersection between the 3D line equation and the plane, denoted as (x_I, y_I, z_I) as illustrated in Fig. 5.4(a). If the direction of the object's movement is toward the 3D line, the framework compares z_I with the h_o to

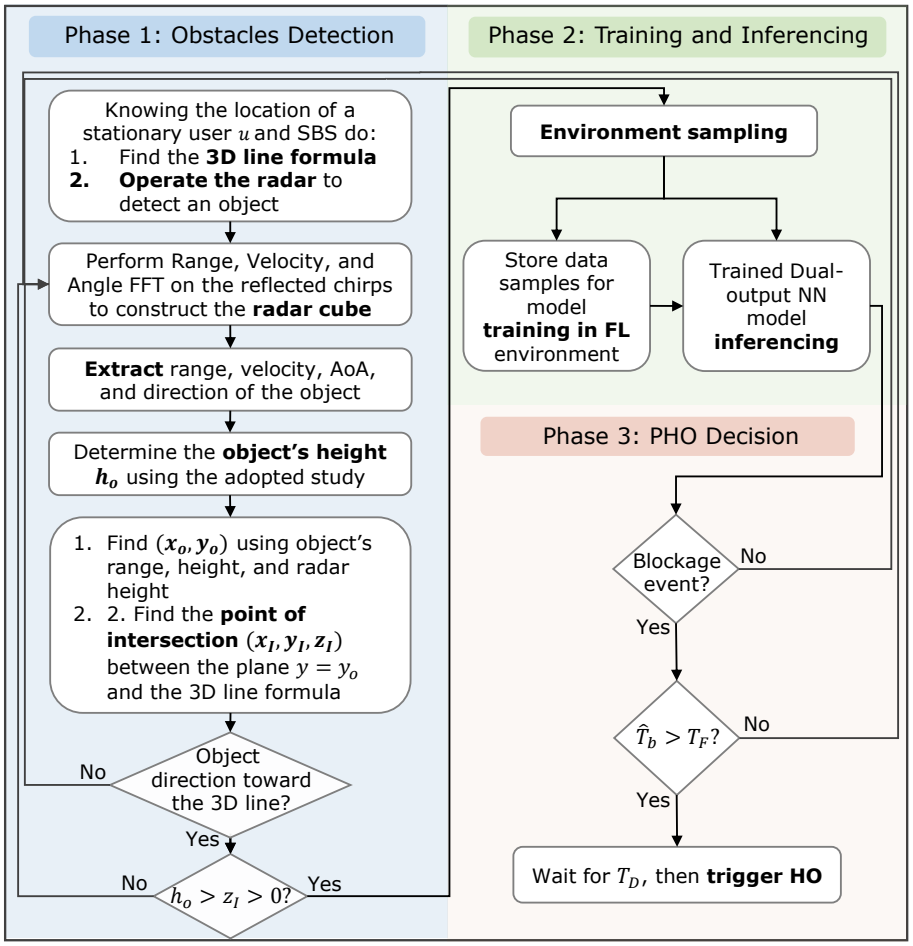


Figure 5.3: Schematic diagram of the proposed framework.

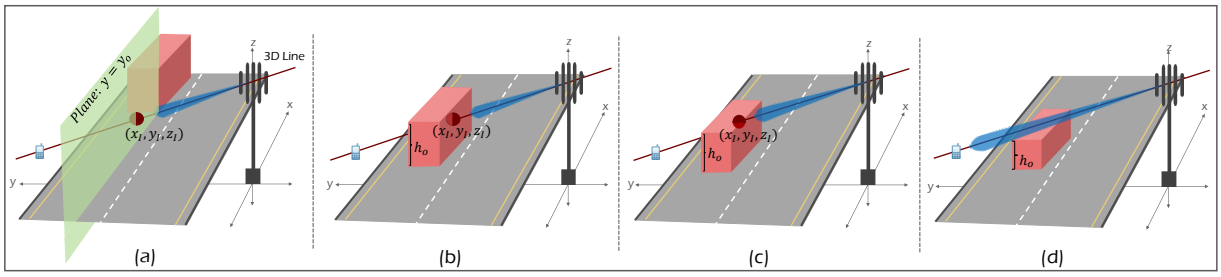


Figure 5.4: Obstacle detection analysis: (a) the use of 3D line equation and $y = y_o$ plane to determine the point of intersection. Assuming object's height is h_o, then (b) $h_o > z_I$, means a blockage, (c) $h_o = z_I$, means a blockage (d) $h_o < z_I$, means no blockage.

assess the potential for obstructing the LoS link. When h_o is greater than or equal to z_I, this means the object will block the connection, as demonstrated in Fig. 5.4(b) and 5.4(c). Conversely, if h_o is less than z_I, this means the object most likely will not block the connection, as shown in Fig. 5.4(d).

Training and inferencing phase

The next phase involves sampling the surrounding environment and collecting the necessary data samples for supervised model training. The framework acquires the requisite dataset to train the NN model for predicting the status and time of blockages by localising users and objects. The training process is performed offline using the FL algorithm, in which each SBS contributes its collected datasets to collaborate on model training. Afterward, the trained model is deployed online to perform inferencing.

PHO decision phase

If a blocking event is anticipated, the primary objective of the framework is to prevent user shadowing and maintain the connection by proactively deciding to perform a HO. This decision is supported by predicting the remaining time until the obstacle blocks the LoS links $\hat{T}_b$. Knowing this time enables better planning for the optimal time to perform HO and maintain the QoE at its highest possible level. However, it is important to measure the total time required by the proposed framework, denoted as T_F, which comprises three main sub-time parameters as follows:

$$T_F = T_R + T_{Inf} + T_{PHO}, \tag{5.17}$$

where T_R is the time duration that begins when the radar is activated and continues until the radar cube information for a single measurement is processed, including the classification of the detected objects. T_{Inf} represents the model inferencing time, and T_{PHO} indicates the time required to switch the user to another stable connection. Moreover, picking the optimal time instant for initiating proactive HO is facilitated by introducing a new time parameter called the delay time (T_D). This parameter represents the idle time between the completion of inferencing and the triggering of HO, and is defined as follows:

$$T_D \leq \hat{T}_b - T_F. \tag{5.18}$$

5.5.2 Radar Measurements Processing

The T_R parameter entails several sub-processes that the framework executes to prepare for the next stage of generating the data samples. Table 5.1 presents the typical radar system parameters considered in this study [138]. Initially, the duration of each measurement,

Table 5.1: Radar system parameters.

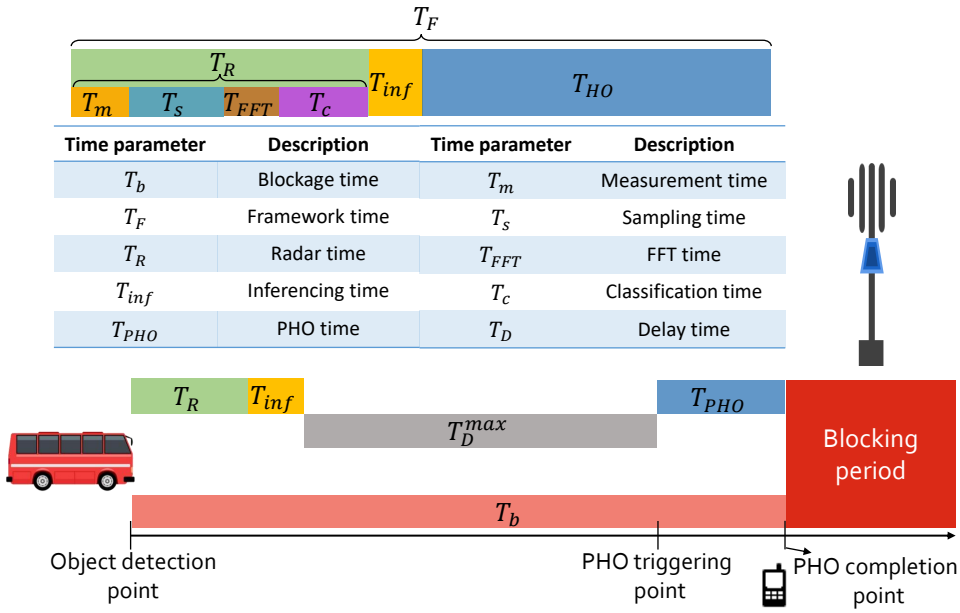
Parameter	**Value**
No. of transmitters	1
No. of receivers	4
No. of chirps	128
Start frequency (f_c)	77 GHz
Chirp slope (μ)	15015 GHz/s
Chirp duration (τ_c)	60 μs
Chirps pause time (τ_P)	5 μs
No. of samples per chirp	256
Sampling rate	5 MHz
Max range	100 m

denoted as T_m, is determined. This measurement involves transmitting 128 chirps, with each chirp lasting for 60 μs, totaling 8.3 ms. Next, the maximum time a radar signal can remain in the air is calculated, given that the maximum radar range is set to 100 meters. This time is expressed in microseconds; however, it can be disregarded since the framework's timescale is in the millisecond range. Then, the sampling time T_s required to perform ADC and sampling of the received reflected signals is measured. T_s is calculated by dividing the number of samples per measurement by the sampling rate and is found to be 26.2 ms.

After acquiring the radar measurements in the form of samples, the next step is to measure the time taken to perform FFT and generate the radar cube, represented as T_{FFT}. Assuming an FFT process has a complexity of $\mathcal{O}(m \log m)$, and a single token m requires one nanosecond to execute, it is estimated that generating a radar cube per measurement requires performing three FFTs, resulting in a processing time of 6 ms [150]. Furthermore, the time required to classify the detected object, denoted as T_c, within the adopted work, is identified. To ensure real-time object classification, the previous study [151] demonstrates that the ResNet-50 model's inference time will be 26 ms, which is considered near real-time. Therefore, it is assumed that the adopted work requires a similar time to classify the detected object. In summary, the following equation defines the sub-processes times covered under T_R:

$$T_R = T_m + T_s + T_{FFT} + T_c. \tag{5.19}$$

Based on the above assumptions and discussion, T_R is calculated to be around 66.5 ms for the illustrative realistic scenario. Fig. 5.5 provides a summary of the defined time parameters in the RaDaR framework and their respective roles.



Time parameter	Description	Time parameter	Description
T_b	Blockage time	T_m	Measurement time
T_F	Framework time	T_s	Sampling time
T_R	Radar time	T_{FFT}	FFT time
T_{inf}	Inferencing time	T_c	Classification time
T_{PHO}	PHO time	T_D	Delay time

Figure 5.5: Diagram of time parameters and descriptions.

5.5.3 FL Design for Model Training

This study aims to select a model that meets various requirements, including high prediction accuracy, low inference latency, collaborative training capability, and simultaneous classification and regression ability. Through careful investigation, a three-hidden layer NN model is designed to process user-object data samples and produce dual predictions of blockage status and time, as depicted in Fig. 5.6. Furthermore, this study aligns with the current research trend of using the FL approach instead of the centralised learning mechanism to provide several benefits, such as safeguarding data privacy, improving bandwidth efficiency, and promoting scalability and knowledge sharing. This approach enables the framework to be generalised by learning from different scenarios, facilitating its deployment in new development areas.

Offline learning phase

The proposed dual-output NN model is trained offline using the FL mechanism. The model comprises an input layer that receives ten features $\{f_i\}_{i=1}^{10}$ of user-object data samples, followed by three hidden layers with 128, 64, 32 neurons, respectively. The model has two output layers: a classification layer with two neurons activated by the softmax function and a regression layer with a single neuron activated by the linear function. The model's architecture is depicted in Fig. 5.6. The loss functions used are mean absolute error (MAE) and sparse categorical crossentropy. Additionally, the model's optimiser, learning rate, batch size, and epochs are set to Nadam, 0.001, 100, 10, respectively. The training process involves SBSs acting as clients participating in model training. This study usees five

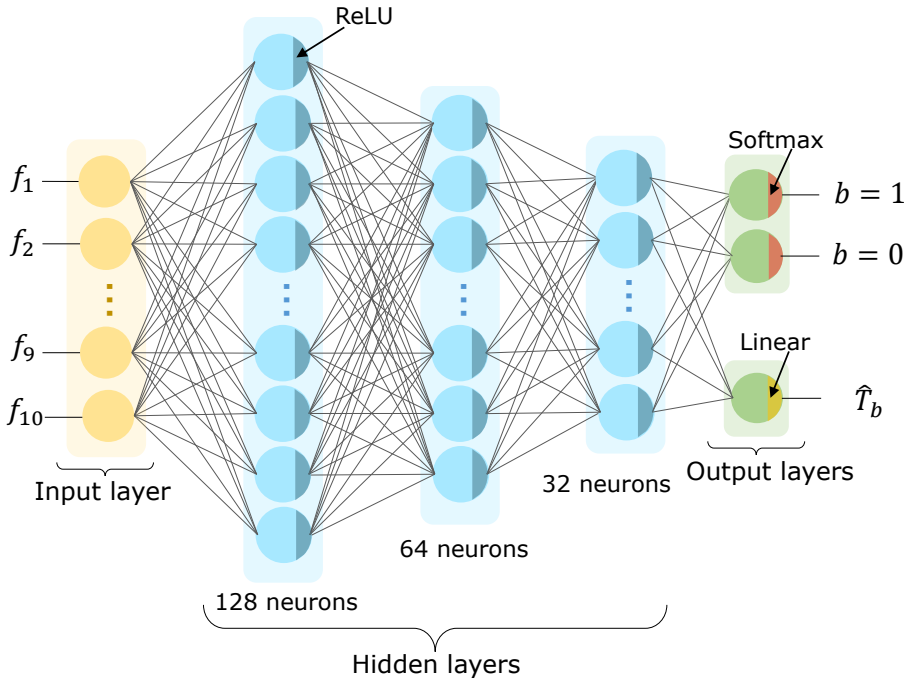


Figure 5.6: The structure of the developed dual-output NN model.

clients, but the framework is scalable and can accommodate large number of clients. Each SBS utilises its user-object data samples to iteratively train the NN model and reports the model updates to a central server located, for example, in a MBS or the cloud. Moreover, the server follows the FedAvg method [58] to aggregate the shared model parameters by computing their weighted average. The weight of each update is proportional to the number of data points of each client. In addition, the number of FL communication rounds is controlled using a stopping technique to avoid suboptimal or unnecessary communication rounds.

Development dataset The NN model is trained and evaluated by exploiting the real-world data from scenario 30 of the DeepSense 6G testbed [138]. This testbed closely resembles the adopted system model, wherein a transmitter and receiver located on opposite sides of a two-way city street. The scenario offers diverse data modalities, including radar measurements and blockage information with associated timestamps. By computing the difference between timestamps, user-object samples can be labelled and determine when the detected object will obstruct the LoS link. However, in scenario 30, every object is blocking the LoS link due to the low height of the BS, which does not necessarily reflect practical network deployments. In contrast, the proposed system considers more practical communication systems, where the detected objects may or may not cause blockages depending on the height of the SBS. Additionally, scenario 30 only involves a single BS communicating with a single stationary user, whereas this study

Table 5.2: The parameters adjusted in the testbed to reflect a practical wireless communication system. The SBS is positioned at the Cartesian origin in the middle of the street, and [· , ·] indicates the range from which values are chosen based on the corresponding distribution.

SBS	**Distrib.**	**Street width(m)**	$\boldsymbol{H}$(m)	$(\boldsymbol{x_u}, \boldsymbol{y_u})$(m)	$(\boldsymbol{x_o}, \boldsymbol{y_o})$(m)	$\boldsymbol{h_o}$(m)	$\boldsymbol{v_o}$(mps)	**No. of samples**	**Block: Non-block**
SBS1	Uniform	40	3	([-20, 20], 12)	([-20, 20], [1, 11])	[1, 4.5]	[3, 9]	10,000	10% : 90%
SBS2	Gaussian	60	4	([-30, 30], 13)	([-30, 30], [1, 12])	[1, 4.5]	[3, 11]	15,000	25% : 75%
SBS3	Gamma	80	5	([-40, 40], 14)	([-40, 40], [1, 13])	[1, 4.5]	[3, 13]	30,000	50% : 50%
SBS4	Binomial	100	6	([-50, 50], 15)	([-50, 50], [1, 14])	[1, 4.5]	[3, 15]	25,000	75% : 25%
SBS5	Poisson	120	7	([-60, 60], 16)	([-60, 60], [1, 15])	[1, 4.5]	[3, 17]	20,000	90% : 10%
SBS6	Beta	100	5	([-50, 50], 13)	([-50, 50], [1, 12])	[1, 4.5]	[3, 9]	2,000	50%: 50%

targets wireless networks that involve multiple SBSs and users.

To overcome these limitations, scenario 30 is analysed and augmented with the objective of generating multiple distinct environments featuring diverse datasets that replicate the complexity found in real-world data distributions. More specifically, in real-world context, data collected from distinct endpoints often exhibit substantial variations in data characteristics, class distributions, and statistical properties. Non-IID datasets offer a more accurate representation of these real-world data distributions. Therefore, the goal is to deliberately introduce the non-IID attribute in the developed wireless environments by generating diverse datasets for each SBS participating in the FL process. The augmentation process commences with the comprehensive examination and understanding of all parameters within scenario 30 testbed, especially those contributing to suboptimal user link connectivity with the BS. Subsequently, new and distinct wireless networks are thoroughly constructed using programming tools. For each augmented network, various testbed parameters are adjusted, including users and objects positions, objects heights, and objects speeds. These adjustments are based on different distributions to encapsulate the non-IID characteristic, which is further validated by ensuring that the mean and variance of each parameter differ across the SBSs. Other constant parameters, such as street width, H, and y_u, are also given different values across different networks to foster environmental diversity. Using the programming tool, user-object datasets S are then generated for each SBS, with varying numbers of data samples and blockage-to-non-blockage ratios. Table 5.2 provides a summary of the modified parameters employed to generate these new non-IID datasets.

Through these modifications and augmentations, a more realistic communication system is established with six SBSs, each having its own user-object data samples. If $h_o \geq z_I$, $[\Upsilon, y_o, y_u, \theta_o, \theta_u]$ is used as a set of features to assess blockage status. The table in Fig. 5.7(a) presents all possible cases, from which it can be observed that only two cases will result in blockages. Fig. 5.7(a) also depicts one of the blockage cases. Similarly, to determine the T_b, the position information, movement direction, and speed are important.

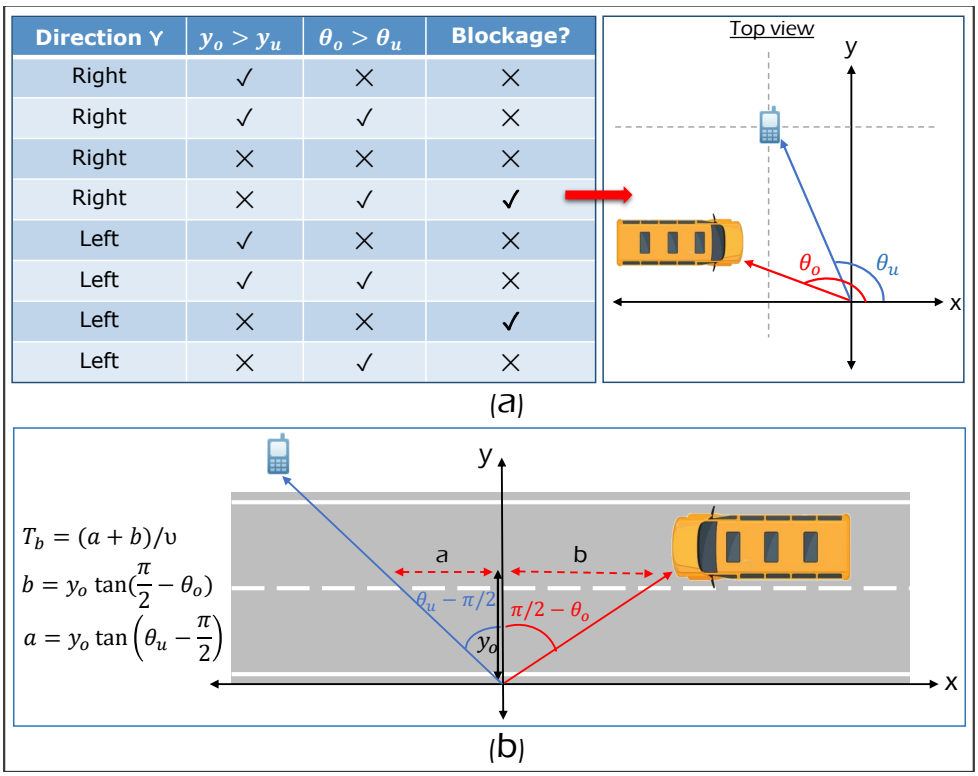


Figure 5.7: Labeling user-object data samples: (a) blockage status of various cases, (b) A case demonstrating the calculation of the T_b.

Using trigonometry, T_b can be determined, as shown in Fig. 5.7(b).

Model evaluation After preparing the data samples with their respective labels for each SBS, 2000 samples were allocated for model evaluation, while the remainder were reserved for training. From the evaluation dataset, a small dataset is randomly selected to form a total of 5000 samples used for evaluating the model's performance at the server. Fig. 5.8(a) displays the classification and regression performance of the dual-output NN model at the SBSs and the server after 30 communication rounds of training using FL. While the classification accuracy was near optimal for all SBS, there were variations in the MAE values, which is attributed to the non-IID properties of the training datasets, where each dataset was obtained from a different distribution representing a distinct environment. To further enhance the performance, each SBS was permitted to personalise the model by utilising 500 data samples and performing a few rounds of tuning. Fig. 5.8(b) illustrates the final performance of the trained models, depicting a further enhancement in the prediction accuracy of the model for each SBS as a result of personalisation.

Knowledge transfer case study During the augmentation process, an additional environment named SBS6 has been created, representing a new deployment area with a limited number of data samples, as outlined in Table 5.2. The purpose is to assess

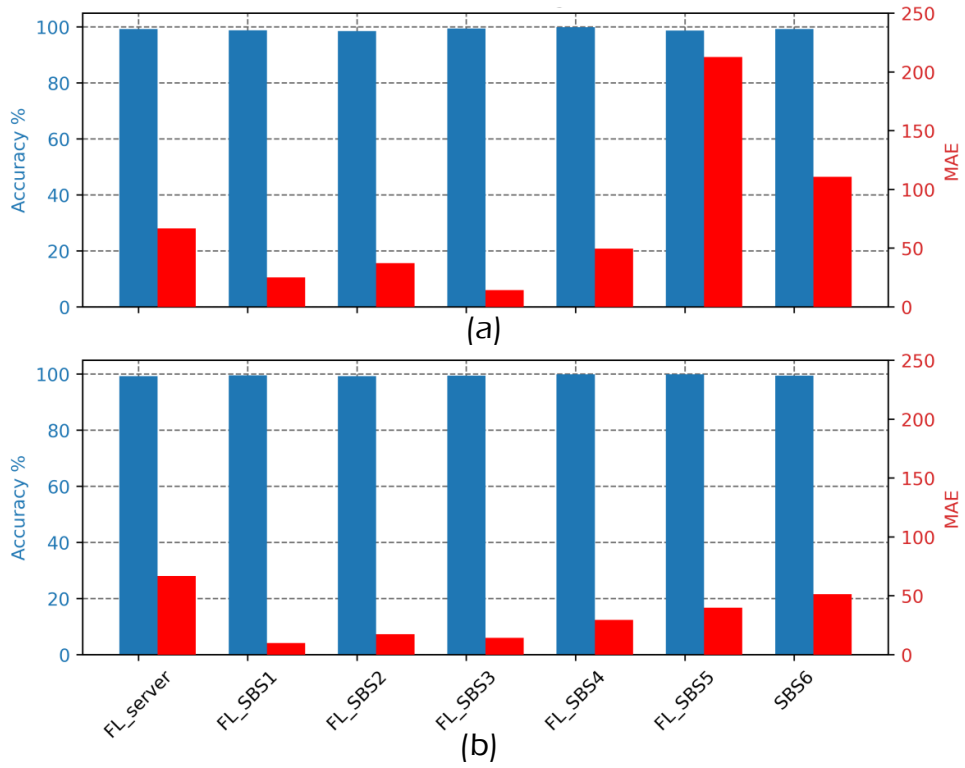


Figure 5.8: The classification and regression performance of the dual-output NN model (a) without and (b) with tuning.

the generalisation and scalability of the proposed framework for supporting knowledge transfer and the rapid deployment of new SBSs. SBS6 did not participate in the FL training process, and the server pushed the latest version of the trained NN model to this SBS to initiate its operation. The performance of SBS6 is presented in Fig. 5.8, showing that the regression accuracy is inferior to that of the first four SBSs, though it is better than SBS5, which was part of the training process. Notably, the classification accuracy is remarkable and can be attributed to the transferability, generalisation, and robustness features of the NN model. These features demonstrate the model's ability to capture fundamental relationships and universally applicable patterns, enabling accurate predictions even in new deployment areas. After tuning, the regression performance is significantly improved by a percent of 53%. These results demonstrate that the proposed framework is scalable and can support the rapid deployment and operation of new network sites.

Online inference phase

After training and personalising the models, each SBS is ready to use its personalised model to predict the occurrence of blockages and, hence, the stability of the LoS beams. When a stationary user is connected to the network, the SBS creates samples of user-object data and feeds them to the model for inference. The model then predicts the status of the

blockages and the time until the object blocks the link. Given that the dual-output NN model is much simpler than ResNet-50, the T_{Inf} should be much lower than the 26 ms inference time of ResNet-50. The T_{Inf} is measured using a standard personal computer and found it to be around 1 ms. Although it should be even lower, for the sake of analysis, it is assumed that T_{Inf} equals 1 ms.

5.5.4 Optimal HO Trigger Point

After detecting an obstacle and determining the blockage time, the framework must notify the network to perform a HO and ensure user connectivity. The crucial question is, what is the best time instant or distance point to trigger the HO process and switch the user to alternative stable links? Defining the T_D parameter in equation (5.18) should help in finding such instant/point by introducing the following formula:

$$\Lambda \leq v \times T_D \tag{5.20}$$

where Λ represents the points at which the network can trigger PHO and maintain stable connections for the user each time an obstacle crosses that points. Since the proposed framework is QoE-aware, it aims to delay the HO process until reaching the optimal point (Λ_{opt}), which corresponds to the maximum tolerable T_D and is given by:

$$\Lambda_{opt} = v \times T_D^{max}. \tag{5.21}$$

By adopting this approach, the proposed framework ensures a seamless user experience while avoiding obstacle's disruption.

5.5.5 PHO Procedure and Latency Minimisation

In conventional wireless networks, when there is a degradation in the signal quality of a user's connection, a HO mechanism is initiated based on predetermined events detected through measurement reports [152]. If beamforming technology is being utilised and the LoS beam is disconnected, then several steps must be taken to re-establish a stable connection. The steps are beam failure detection, beam failure recovery, cell search, and contention-based or free random access. Each of these steps requires execution time that combined would result in high latency $\sim$ 312.2 ms, affecting the reliability of the communication system.

In this study, if a BLOCK event is detected, the network should take preemptive measures to prevent channel interruption. The proposed framework is proactive in nature, eliminating the first two steps of the HO process and performing the cell search step in advance while the user is still connected. Therefore, the HO latency boils down to the

latency associated with experiencing either contention-based or free random access. This framework considers the worst-case of contention-based random access that requires about 80 ms [153]. As a result, the last parameter of equation (5.17), i.e., T_{PHO} equals to 80 ms. Now that all the values of the parameters in equation (5.17) have been determined, the total is 147.5 ms, representing the execution time of the proposed framework. Finally, it is important to note that the performance of RaDaR framework depends heavily on the hardware specifications of the radar, server, and network. As hardware specifications improve, the framework's execution time will reduce, resulting in a further enhanced framework.

5.6 Performance Evaluation and Results

This section examines the efficacy of the proposed RaDaR framework in improving the operation of high-frequency communication systems, such as UDNs, by effectively and preemptively predicting the occurrence of beam blockages and implementing appropriate measures to guarantee uninterrupted connectivity for users while preserving high levels of QoE.

5.6.1 Dual-output Model Development

After training the dual-output NN model in an FL environment and performing personalisation, the resulting models are now ready to be tested in practical scenarios. To further evaluate the models, a new parameter called the PHO success rate ($\mathcal{S}_{PHO}$) is defined, representing the percentage of successfully detecting beam blockages and performing PHO. $\mathcal{S}_{PHO}$ is calculated by dividing the number of samples successfully performing PHO by the total number of samples. Using the same evaluation dataset of the SBSs, the $\mathcal{S}_{PHO}$ results for each SBS are presented in the 0% column of Table 5.3. The figures reveal variations in the results, where the $\mathcal{S}_{PHO}$ of some SBSs is unsatisfactory, while others exhibit better performance. After careful investigation, it was discovered that the inconsistent behavior is due to the inaccurate prediction of the blockage time $\hat{T}_b$. Occasionally, the proposed model predicts the value of $\hat{T}_b$ greater than the actual blockage time. Considering the QoE-aware nature of the RaDaR framework, it waits for the maximum $\hat{T}_D^{max}$ ($\hat{T}_D^{max} = \hat{T}_b - T_F$) before initiating a PHO. As a result, the value of $\hat{T}_D^{max}$ may exceed the actual T_D^{max}, leading to a scenario where the user encounters a blockage by a vehicle before completing the PHO process.

To address this issue, a new parameter named "percent shift" ($\mathcal{P}_{Shift}$) is introduced and designed to mitigate the impact of blockages that occur when a model predicts $\hat{T}_b$ larger than the actual blockage time. The utilisation of $\mathcal{P}_{Shift}$ involves reducing all predicted $\hat{T}_b$ values for all samples by a specific percentage, determined the value of $\mathcal{P}_{Shift}$. For

Table 5.3: Study of $\mathcal{S}_{PHO}$ [%] versus $\mathcal{P}_{Shift}$ for different SBSs.

SBS	$\mathcal{P}_{\textbf{Shift}}$					
	0%	2%	4%	6%	8%	10%
SBS1	55.4	69.6	80.1	87.2	91.4	**93**
SBS2	36.9	66.5	81.5	87.4	89.8	**90.6**
SBS3	70	82.4	85.6	86.7	87.7	**88.2**
SBS4	96.9	99.1	99.1	99.1	**99.3**	99.3
SBS5	41.1	87.9	96.6	98.2	**98.8**	98.8
SBS6	30.5	62.3	85.3	93.2	95.7	**97.3**

example, when the $\mathcal{P}_{Shift}$ is set to 10% and the predicted $\hat{T}_b$ is 100ms, applying the $\mathcal{P}_{Shift}$ parameter will reduce the $\hat{T}_b$ value to 90ms. This may result in slightly reduced QoE for the user due to earlier PHO. However, maintaining a stable connection with slightly reduced QoE is believed to be preferable to losing the connection and having to reestablish it. The findings show that the introduction of $\mathcal{P}_{Shift}$ has significantly improved the $\mathcal{S}_{PHO}$ for all SBSs, as shown in Table 5.3. Nevertheless, there exists a trade-off between selecting the values of $\mathcal{P}_{Shift}$ and the perceived QoE. It is imperative to note that higher values of $\mathcal{P}_{Shift}$ can yield further improvements. However, such improvements may necessitate even earlier PHOs, thereby potentially degrading the user's QoE. Hence, the value of 10% has been selected for SBS1, 2, 3, and 6, and the value of 8% for SBS4 and SBS5.

To investigate the effect of the $\mathcal{P}_{Shift}$ parameter, Fig. 5.9 plots the CDF for every sample $i \in D$ results in successful PHO, with respect to time delay offset T_{DO}. Here, T_{DO} is defined as:

$$T_{DO_i} = \frac{T_{D_i}^{max} - \hat{T}_{D_i}}{T_{D_i}^{max}} \times 100\%, \; \forall \hat{T}_{D_i} \leq T_{D_i}^{max}, i \in D \tag{5.22}$$

where T_D^{max} is the maximum actual delay time before triggering PHO, and $\hat{T}_D$ is the predicted time delay given as $\hat{T}_D = \hat{T}_b - T_F$. T_{DO} indicates how far the $\hat{T}_D$ from the actual one; the closer the values to zero, the better the performance. Fig. 5.9 illustrates the overall performance of the SBSs in predicting blockages and successfully executing PHO, with only small variations in the values of the T_{DO}. All SBS have more than 80% of their samples with T_{DO} is less than 20%, which highlights the superiority of the proposed framework in proactively predicting blockages and performing PHO at the time/point that maintains the user's QoE as high as possible. Thus, the framework of each SBS is now ready for deployment and further investigation.

5.6.2 Simulation Setup

A hybrid approach is employed that integrates real-world and simulated environments to assess the overall performance of the proposed RaDaR framework. The simulated environments are derived from scenario 30 to reflect more practical wireless networks and

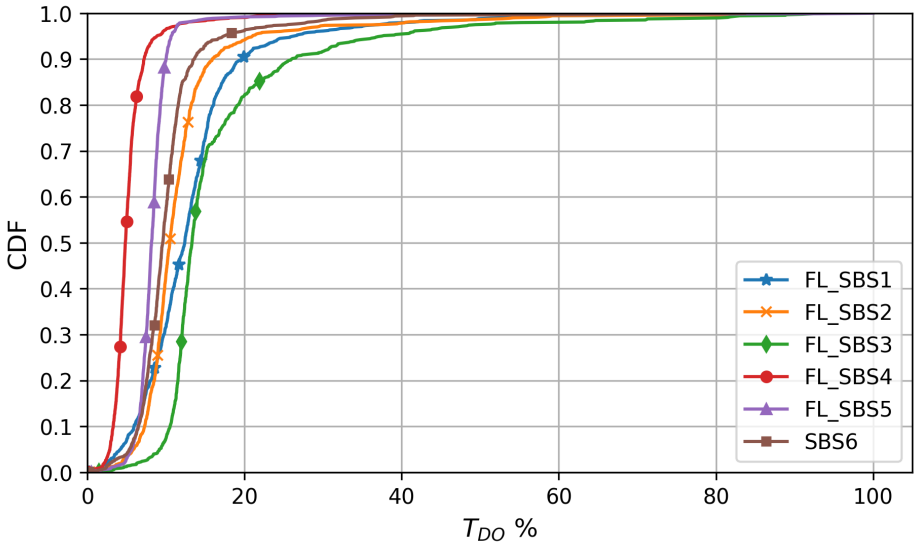


Figure 5.9: The distribution of the T_{DO} samples that lead to a successful PHO for different SBSs.

generate multiple distinct environments. The simulation experiments are implemented based on Python installed on a Windows operating system with an Intel Xeon CPU E5-2620 @ 2GHz and 16GB RAM. The key performance metrics are the RSS, network latency, and throughput.

5.6.3 Performance Analysis

This section reveals the effectiveness of the proposed RaDaR framework in comparison to a conventional wireless network that lacks proactive blockage prediction techniques and only responds reactively to blockage events, referred to as Reactive-HO. Initially, the effect of blockages on the RSS at the user's end in Reactive-HO communication systems is analysed, using the diverse information available in the DeepSense scenario 30 testbed. Fig. 5.10 illustrates the normalised RSS of a stationary user, blockage events, and the best beam index at each data sample. The results demonstrate unstable performance, as the RSS deteriorates each time an obstacle obstructs the LoS beam serving the user. This performance is unsuitable for time-sensitive services and data-intensive applications, such as high-definition video streaming, which require fast and reliable wireless connections to guarantee an uninterrupted and smooth user experience. Furthermore, the figure demonstrates that in the absence of obstacles, the best-serving beam is consistently limited to a few fixed beams. However, when the communication channel is blocked, the best beam index varies and can be any beam from the beamforming codebook, determined by the beam that gives the highest received power after reflection from the environment. These findings highlight the importance of proactive blockage prediction techniques in achieving reliable and stable wireless communication performance.

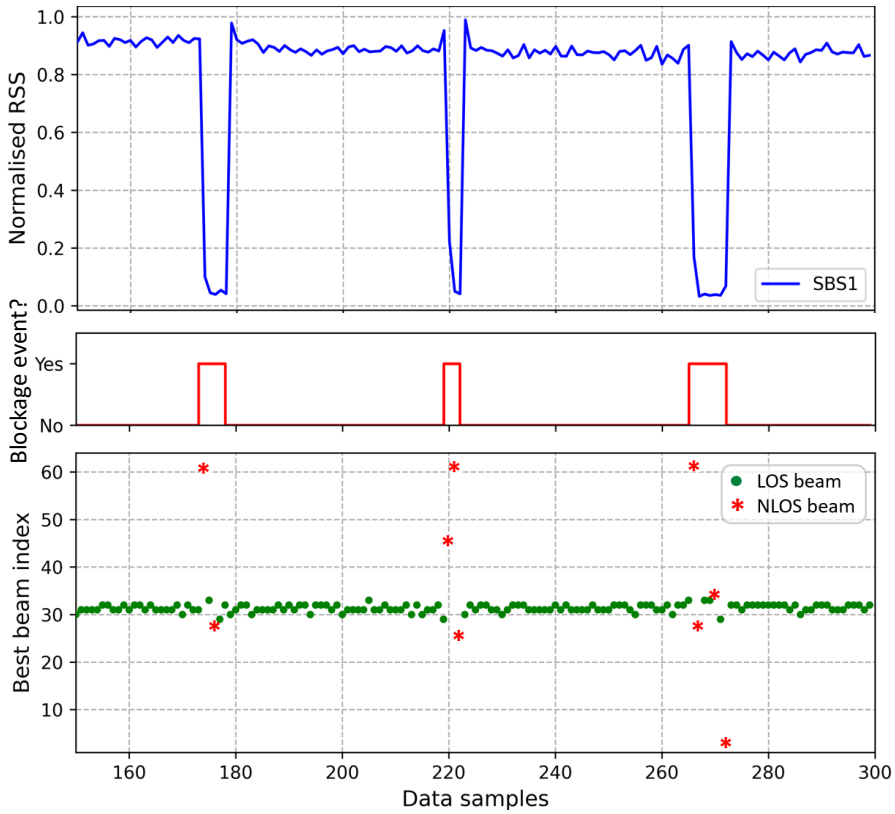


Figure 5.10: The normalised RSS, blockage events, and best beam index in a Reactive-HO communication system.

Next, the efficacy of the proposed RaDaR framework is examined by augmenting the scenario 30 testbed using Wireless InSite ray-tracing software. Specifically, a new SBS (SBS2) with the same specifications as the existing SBS1 is introduced, placed 80 meters apart on the opposite side of the street. Fig. 5.11 depicts the normalised RSS from both SBS and the SBS that serves the wireless user at different time instances. Once the user is connected, the RaDaR framework begins to function by monitoring the surrounding area using radar sensors. Prior to the first blockage event, the user is connected to SBS1 since it is within the SBS's coverage area and receives a higher RSS from SBS1 than SBS2. When an obstacle enters the FoV of the radar, the framework detects the object, and the communication system becomes aware of this potential blockage. The framework classifies the object as a blockage and predicts the blockage time. It then determines the optimal time to perform PHO. The figure reveals how the RaDaR framework can detect blocking objects and switch the user to SBS2, which offers a more stable communication channel. It is essential to note that the QoE of the user when served by SBS1 is better than that when connected to SBS2. However, scarifying the perceived QoE slightly is preferable to experiencing disconnection and engaging in undesired network operation to resume the connection, which impacts network latency and affects its reliability.

Finally, two key performance metrics for wireless networks are examined, namely the

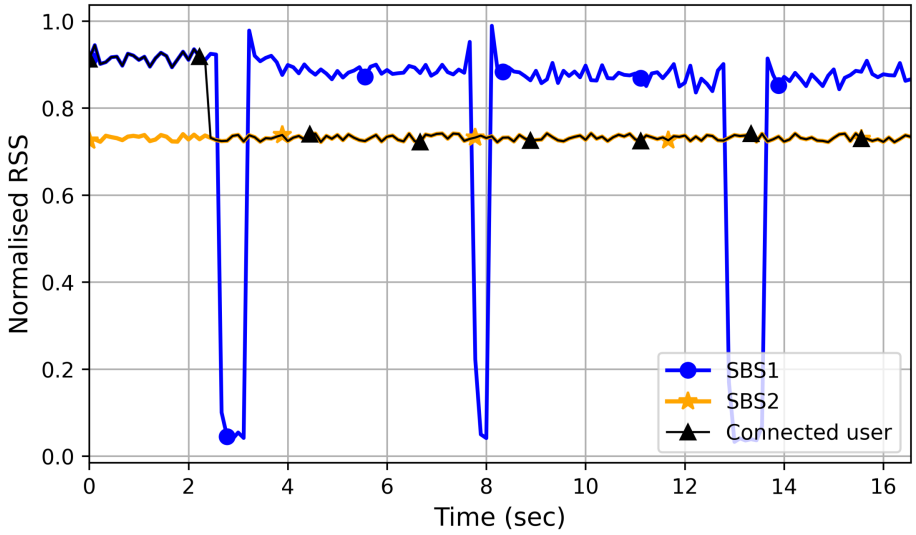


Figure 5.11: The impact of blockages on the user's RSS and how RaDaR is able to detect blockages and ensure seamless connectivity.

network latency associated with performing handovers and average user throughput for Reactive-HO and radar-aided PHO networks. The proposed study involves monitoring the evaluation environments over a certain period of time and considering multi-users by treating the stationary user as a new user whenever a new obstacle is detected. Fig. 5.12 shows the normalised results of average latency and throughput for both Reactive-HO and radar-aided PHO networks, represented by SBS1-SBS6. In terms of average latency, as discussed in Section 5.5.5, Reactive-HO communication systems must perform four steps that take 312.2 ms each time a link is disconnected and a user needs to be switched to another SBS. In contrast, the proposed framework eliminates the first three steps, reducing PHO latency to 80 ms. Making use of $\mathcal{S}_{PHO}$, the average PHO latency per user is measured as follows:

$$\zeta = \frac{\{\mathcal{S}_{PHO} \times U\} \times 80 + \{(1 - \mathcal{S}_{PHO}) \times U\} \times 312.2}{U}. \tag{5.23}$$

Overall, SBSs adopting the RaDaR framework outperform Reactive-HO networks lacking proactiveness in detecting blockages. Moreover, the average latency decreased with an increase in $\mathcal{S}_{PHO}$ as the probability of detecting blockages and performing successful PHO increased. The variations in the average latency values across SBSs are attributed to the differences in the $\mathcal{S}_{PHO}$, which heavily depend on the serving environment of the SBS.

Regarding average throughput, this study is performed by monitoring the environments for a specific period of time and recording the user's throughput at every time instant, irrespective of the presence of obstacles. Reactive-HO networks experienced a significant drop in user throughput due to LoS beam disconnection, which required users to switch to other reflected beams with reduced throughput. However, radar-aided PHO systems have a higher probability of predicting obstacles in advance and switching users to an SBS,

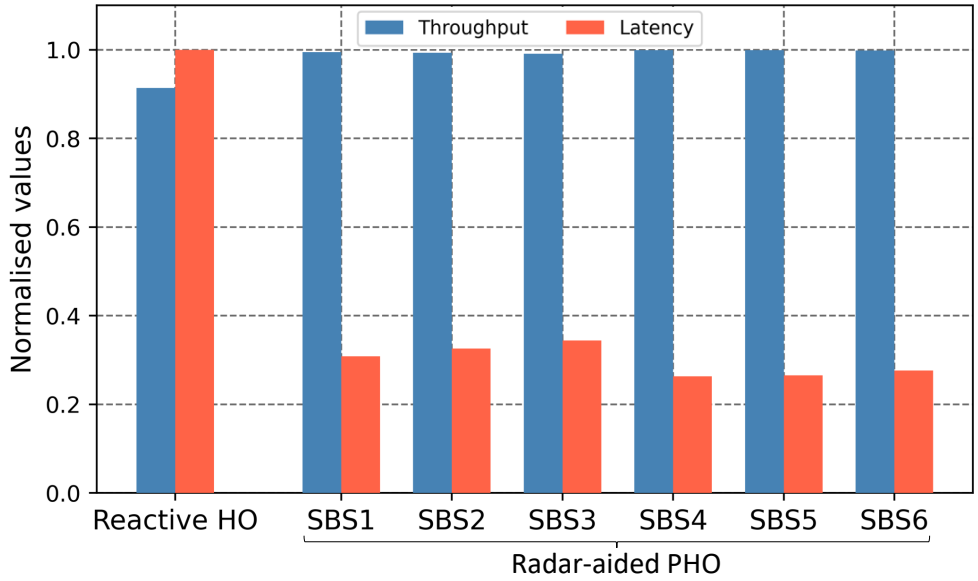


Figure 5.12: Latency and throughput study.

offering a stable connection, thereby maintaining high throughput levels.

5.7 Summary

In this chapter, a radar-aided dynamic blockage recognition framework called RaDaR is proposed. The main objective is to increase high-frequency networks' awareness of their surrounding environment and improve network reliability. Radar measurements are utilised for training a dual-output NN model using FL to predict forthcoming link blockages and determine the optimal time to perform PHO, thereby avoiding link disruption. To compare the effectiveness of RaDaR, a conventional wireless network lacking proactive blockage prediction mechanisms, named Reactive-HO, is considered. The performance of the suggested framework is assessed by utilising co-existing modalities derived from both real-world and simulated environments. The simulation results confirmed that RaDaR, with its blockage-aware approach, enhances the QoE for users who require low latency and operate in extremely dynamic environments. Compared to Reactive-HO networks, evaluation results indicated that the framework significantly improves the operation of next-generation wireless networks by offering high RSS, maintaining high throughput levels, and reducing network latency, enabling future latency-sensitive applications. Moving forward, the thesis shifts its focus to energy management in UDNs. The next study aims to establish more sustainable energy management plans by introducing an efficient framework capable of improving energy forecasting.

Chapter 6

Optimising Power Consumption in UDNs: Introducing FedraTrees for Lightweight Distributed Learning

The previous three chapters have addressed a critical challenge related to mobility management in UDNs, specifically the beam blockage problem. However, the dense deployment of SBSs, coupled with the adoption of sensing-aided wireless communication concept and ML methods, results in increased power consumption in these networks, exacerbating energy management issues. Therefore, this chapter focuses on the energy management aspect, specifically aiming to achieve the energy efficiency requirement of next-generation wireless networks, aligning with the objectives outlined in *O4*. This chapter introduces FedraTrees, a novel lightweight framework designed to facilitate the use of DT-based models within the FL setting. While FedraTrees is developed for energy forecasting tasks, its applicability extends to various applications. Using real-world power consumption datasets, FedraTrees demonstrates remarkable performance in predicting short-term power usage with significantly fewer communication and computation resources compared to the commonly used FedAvg algorithm.

6.1 Introduction

In UDNs, energy efficiency is crucial due to surging wireless service demands, making sustainable network operation paramount. Thus, energy forecasting becomes a strategic approach to achieving optimal energy efficiency within UDNs. While various categories of methodologies, such as infrastructure optimisation, spectrum management, and power-saving techniques, play pivotal roles in mitigating energy consumption, the integration of energy forecasting adds a layer of proactive adaptability. By leveraging predictive analytics and advanced ML techniques, energy forecasting empowers UDN to anticipate future

energy demands accurately. Accurate short-, mid-, or long-term energy forecasting enables network operators to orchestrate energy resources, fine-tune operational parameters, and allocate resources judiciously, thereby averting unnecessary energy wastage. The synergy between energy forecasting and the broader array of energy-efficient strategies paves the way for smarter UDN, where proactive management ensures not only seamless connectivity but also a substantial reduction in the network's carbon footprint.

Various techniques have been considered for efficient and reliable load forecasting. Statistical forecasting methods, such as MLR, AR, and MA, have been used to project past and present load profiles into future predictions. Later, the introduction of smart metering and the evolution of AI technology paved the way for replacing traditional prediction techniques with various ML algorithms due to their ability in analysing large datasets in short periods of time while providing impressive accuracy levels [154]. advanced metering infrastructure (AMI), a system of smart meters connected to a communication network, represents the first step toward smart energy, enabling the collection and analysis of smart-meter data. However, collecting load profiles into a central entity to conduct energy forecasting raises privacy and security concerns [155]. Load profiles hold sensitive information that can be used in various dimensions like inferring operation patterns of networks, unauthorised data use, and data selling, among others [156]. Moreover, sending massive amounts of constantly generated power consumption data to a central location burdens communications resources and is costly.

To address these issues, FL has emerged as the best choice for handling enormous datasets and developing efficient and scalable systems. The use of FL in energy forecasting is still in very early stages, and few studies have considered this approach [49–53]. In these studies, the authors focused on utilising LSTM architectures, a type of RNN used in the field of DL, due to their remarkable performance in predicting time-series data sequences. However, these works overlooked a critical issue: DL models are extremely resource-intensive (energy, memory, processor, etc.), and the lengthy and extensive underlying mathematical operations demand resource-rich hardware. Considering such schemes of combining FL with LSTM models requires extended computation time to reach the desired precision and impedes their scalability.

6.1.1 Contributions

In this chapter, FedraTrees, a novel light aggregation algorithm developed to utilise DTs under the FL setting, is proposed. Specifically, the light gradient boosting model (LGBM) model [157] is used, one of the boosting techniques in ensemble learning, to be sent and trained across the clients of the FL framework. The main reasons for considering LGBM models are their rapid training speed, lower memory consumption, higher efficiency, and accurate predictions. FedraTrees aims to minimise computations

and the number of communication rounds while guaranteeing high training performance; hence, it is envisioned to play a crucial role in a wide range of FL-based applications in several fields, such as smart energy. Moreover, this work considers the common challenge of optimising the number of communication rounds that may lead to suboptimal performance or excessive rounds of unnecessary training, thus consuming computation and communication resources. Therefore, a delta-based FL stopping technique is developed, a dynamic algorithm that monitors the FL training process and stops it when no further enhancement is possible. Furthermore, this work examines the importance of each feature used in the training process and offers a study of the effect of using a different number of features on the final training performance. Additionally, the performance of the FedraTrees algorithm is benchmarked against the popular LSTM-based FedAvg and the naïve Persistence model. Finally, the proposed framework is evaluated based on state-of-the-art metrics and by conducting extensive simulations. The main contributions of this study are summarised as follows:

- FedraTrees, a novel, light algorithm that employs DT-based models within the FL setup, is introduced. Using LGBM models, FedraTrees shows improved performance in terms of the required number of communication rounds and computation time compared to LSTM-based FedAvg aggregation scheme when applied to energy forecasting.
- Unlike other FL-based energy forecasting frameworks that rely on fixing the number of communication rounds, this study develops a delta-based FL stopping technique to reach the best possible accuracy while reducing the computation and communications costs.
- A feature importance evaluation study is conducted on load profiles to achieve optimal forecasting performance.
- Finally, the performance of FedraTrees is compared with LSTM-based FedAvg and a naïve Persistence model as benchmarks, using state-of-the-art metrics. The results demonstrate a significant improvement in overall performance when utilising the FedraTrees framework.

6.2 Preliminaries and background

This section provides an overview of the fundamentals of LSTM, LGBM, and FedAvg necessary for a better understanding of the study conducted in this chapter.

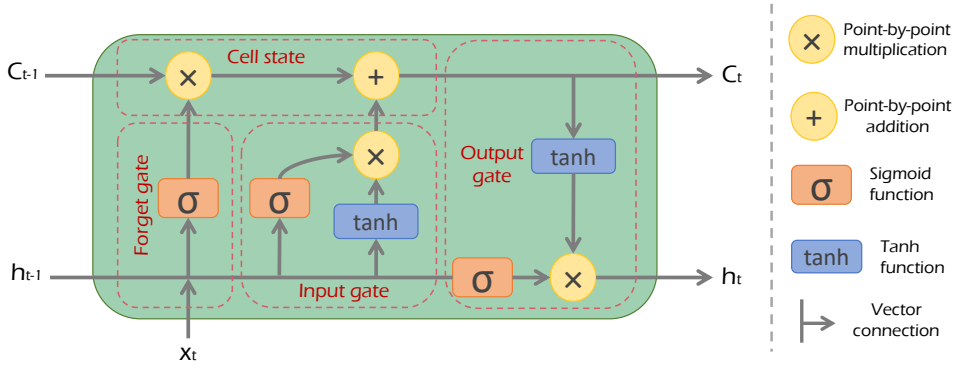


Figure 6.1: LSTM memory cell with gating units.

6.2.1 Long Short-Term Memory Networks

An RNN is a class of artificial neural networks designed to process sequential data, recognising patterns, understands temporal dynamic behaviours, and making predictions based on previous states [158]. Like NNs, RNNs are gradient-based learning algorithms that rely on the backpropagation mechanism to update the weights of neurons. However, RNNs face two main challenges related to performing partial derivatives across the network to find model weights: vanishing and exploding gradients. These problems can prevent the network from further training, causing the network to struggle to learn long-term dependencies. Therefore, LSTM networks were introduced as more robust models without being affected by the unstable gradients problem by improving the gradient flow [158]. This was achieved by introducing artificial memory (cell state) and three gates: a forget gate, an input gate, and an output gate. The three gates can be thought of as filters that regulate the flow of information into and out of the cell to help predict the output sequence. A visual representation of an LSTM unit is given in Fig. 6.1.

6.2.2 Light Gradient Boosting Model

DTs are supervised ML algorithms that can perform regression and classification tasks by iteratively splitting data according to specific rules. The simplicity of DTs has contributed to their popularity, and they have been applied in many applications. However, DTs suffer from several challenges that limit their use in more complex situations, such as overfitting, instability, and bias. Therefore, the concept of ensemble learning was adopted with the aim of combining many weaker learners (i.e., DTs) to produce a more robust ensemble. Ensemble learning includes three main classes: bagging, stacking, and boosting. For more details on ensemble learning, readers are referred to [159].

In 2017, Microsoft introduced one of the most powerful boosting models named LGBM [157], as depicted in Fig. 6.2. What distinguishes LGBM from other boosting models is its efficiency, fast processing, and scalability. These highly desirable features are gained

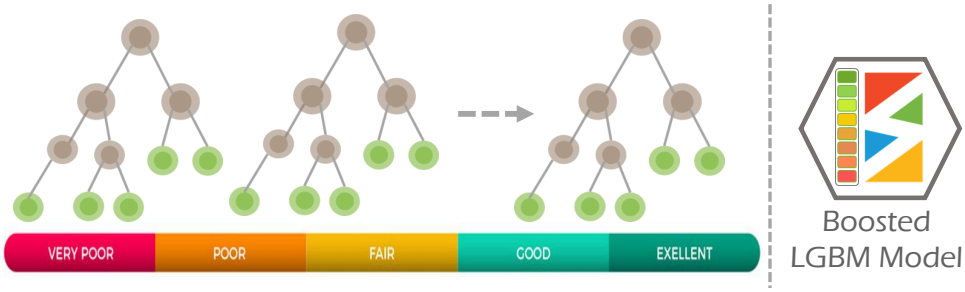


Figure 6.2: Ensemble of DTs are combined to boost and form the LGBM model.

through the incorporation of two techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). GOSS selectively downsamples data instances with small absolute gradients while keeping those with higher absolute gradients, as they contribute more to the training process. At the same time, EFB reduces the number of data features by bundling mutually exclusive ones. These characteristics, including reduced computations, lower memory consumption, fast training, the ability to handle large-scale data, and scalability, position LGBM algorithms as powerful tools for solving numerous real-world problems.

6.2.3 Federated Averaging

ML is a data-driven approach that relies on several quantities of datasets to train a model for specific tasks. Conventionally, model training occurs centrally by collecting the required datasets in a central location. However, centralised training methods threaten data privacy and security, and contradict the legislation in data protection laws. Therefore, FL has emerged as a promising solution that addresses privacy concerns by pushing the model to the locations where the data is generated and resides [160]. Several aggregation algorithms have been introduced, such as FedAvg [161] and federated distance (FedDist) [162]. However, these algorithms are designed to aggregate the parameters of NNs models and cannot be used with other ML algorithms, like DTs. Furthermore, the FL framework heavily relies on the resources of clients and networks, and using NN models under the FL framework can lead to substantial resource consumption. Therefore, developing new efficient aggregation schemes is imperative. To tackle this problem, this study proposes a novel DTs-based FL framework termed FedraTrees, which will be discussed in detail in the next section. FedraTrees performance is compared with FedAvg, the most widely used aggregation algorithm in the literature.

The FedAvg algorithm is best suited for NNs, where network parameters (weights and biases) can be extracted and aggregated to form updated parameters. FedAvg is characterised by three main parameters: a subset of total client count (C), breaking down the dataset into small-sized mini-batches (B), and the number of epochs (E) through which

Algorithm 1 Federated Averaging [161]

Require: Communications limit (C_l), Clients count (C_c), Number of mini-batches (M_b), Epochs number (E_n)

Server executes:

$w_0 \leftarrow$ initialise weights

for $t = 1, 2, ..., C_l$ **do**

$R_s =$ Random set of C_c

for $k = 1, 2, ..., R_s$ **in parallel do**

$w_t^k \leftarrow \text{ClientUpdate}(k, w_{t-1})$

$w_t \leftarrow \sum_{k=1}^{R_s} \frac{n_k}{n} w_t^k$

ClientUpdate (k, w)**:**

for $e = 1, 2, ..., E_n$ **do**

$batches \leftarrow$ Split dataset into M_b batches

for $b = 1, 2, ..., batches$ **do**

$w \leftarrow w - \eta \nabla l(w; b)$ *//Compute gradients and update weights*

Return w to server

the client passes over its dataset per round. FedAvg allows each client to perform multiple rounds of SGD locally on different subsets of its local data. Then, it derives the optimal model parameters by averaging the locally computed gradients at the FL server. The complete pseudo-code of FedAvg is given in Algorithm 1. In addition, LSTM networks have been widely employed in predicting time-series datasets, making them suitable for federated optimisation problems. In previous studies focusing on energy forecasting tasks under the FL environment, LSTM networks were the predominant model choice. Although LSTM models deliver accurate prediction performance, these studies often overlook crucial problems associated with FL distributed learning: computational and communication costs.

6.3 Proposed FedraTrees Algorithm

6.3.1 FedraTrees in LGBM-based FL

As mentioned earlier, in the federal environment, the use of NNs is becoming increasingly popular. However, many other ML algorithms have not been explored in such an environment, despite possessing desirable merits like simplicity and efficiency. Recently, the research community has begun to investigate the use of alternative ML techniques, such as DTs [163], [164]. This study follows a similar concept of employing DTs under the FL setting, but it is the first research to explore the use of LGBM models with FL. Additionally, this study pioneers LGBM-based FL within the context of smart energy. Inspired by the federated voting (FedVoting) algorithm presented in [164], where the

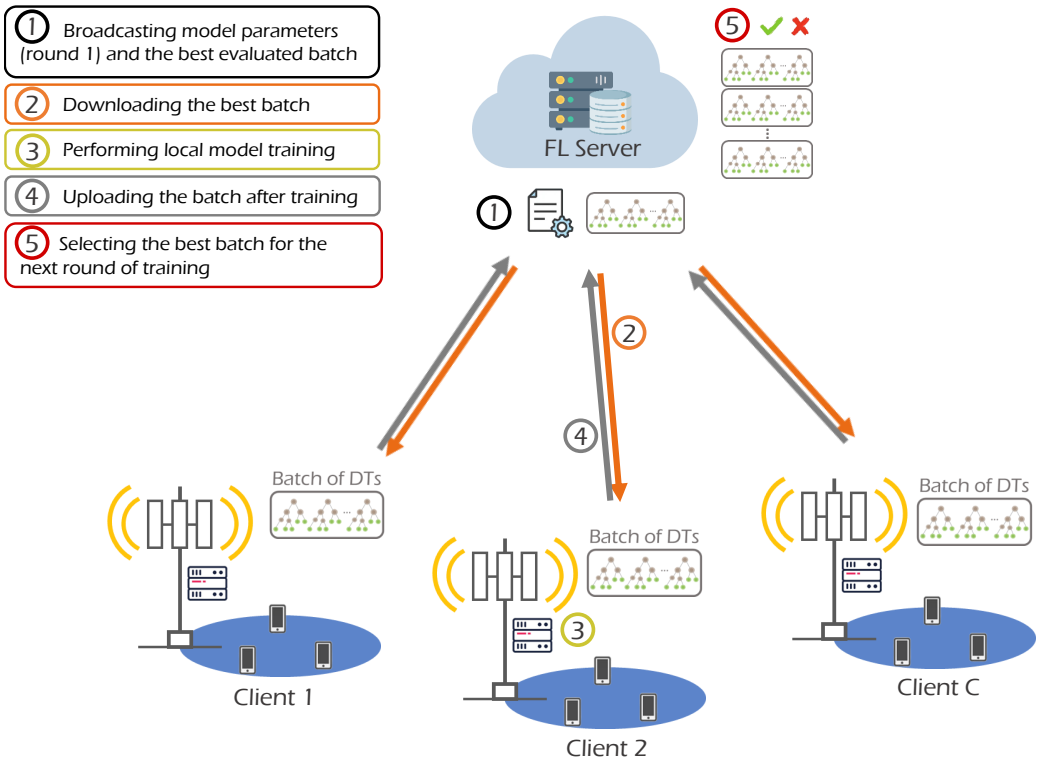


Figure 6.3: FedraTrees sequential operation steps for the energy forecasting task considering C SBSs as clients.

authors construct a federated GBDT model for human mobility prediction, this study introduces FedraTrees. The FedraTrees framework harnesses the capabilities of the LGBM algorithm within the FL environment. What distinguishes FedraTrees from FedVoting is that its simplicity and scalability. Specifically, FedVoting is developed for the cross-silo setup and relies on cross-validations. To determine the optimal model for each training round, each client must validate other clients' trained local models, incurring additional computation cost. Moreover, scalability is not by design of FedVoting since it is designed to be performed on a limited number of clients. By contrast, FedraTrees is developed to fulfil simplicity, efficiency, and scalability, accommodating both cross-device and cross-silo settings. In FedraTrees, complexity is alleviated because there is no need for clients to validate others' models; instead, the central server validates the received models and selects the best one to build upon in upcoming training rounds.

Fig. 6.3 demonstrates the structure of FedraTrees and its detailed training process. FedraTrees is a batch-based aggregation algorithm that determines the number of DTs per batch to be trained in each round and builds on it. In addition to determining the number of trees per batch, constructing FedraTrees begins with choosing the best hyperparameters of the LGBM model, like the number of leaves for each estimator/tree, boosting type, and maximum depth. Once these parameters are determined, the central server broadcasts them to all participating client. Upon receiving the model parameters, each client creates

Algorithm 2 FedraTrees

Require: Communications limit (C_l), Clients count (C_c), Number of trees per batch (G)
Server executes:
for $r = 1, 2, ..., C_l$ **do**
  if $r == 1$ **then**
    Broadcast LGBM parameters values to clients
    $Batch^0 = \{\}$
  Broadcast $Batch^{r-1}$
  for $k = 1, 2, ..., C_c$ **in parallel do**
    Store $Batch^{r-1}$ in memory
    $Batch_k^r \leftarrow \text{ClientUpdate}(k, \sum_{i=0}^{r-1} Batch^i, r)$
  //Find the best Batch
  $Batch^r \leftarrow$ Server validation $\{\sum_{i=0}^{r-1} Batch^i + Batch_k^r\}_{k=1}^{C_c}$

ClientUpdate $(k, \sum_{i=0}^{r-1} Batch^i, r)$**:**
$Model_k^r \leftarrow \sum_{i=0}^{r-1} Batch^i + \sum_{t=1}^{G} DT_t$
for $t = 1, 2, ..., G$ **do**
  Find the optimal split for each split node in the new batch trees and update $Model_k^r$
Return $\sum_{t=1}^{G} DT_t$

the LGBM model accordingly and starts the iterative model training process. At the end of each training round, clients only send the last trained batch of trees to the central server for evaluation. The server updates the global model by appending a single batch of trees at a time, evaluating the model's performance, and repeating this process until all clients' batches have been tested. The batch of trees that achieves the best evaluation performance is retained to build upon in the subsequent communication rounds, while the batches from other clients are discarded. The server broadcasts the newly elected batch of trees to the clients, where it is appended to the previous version of the model, and a new training round begins. This process is repeated until the global LGBM model reaches the desired accuracy or no further improvement can be attained. The complete pseudo-code for FedraTrees is presented in Algorithm 2.

6.3.2 Delta-based FL Stopping

The challenge of training an ML model in the federated setting lies in determining the number of communication rounds. A large number of rounds can result in unnecessary computations and resources wastage, potentially leading to overfitting. Whereas a small number of rounds may lead to a suboptimal model suffering from underfitting. This study also considers this challenge by developing a stopping algorithm that continuously monitors the evaluation performance of the trained model at each iteration and halts the FL training process when no further improvement is expected. This algorithm is inspired by the

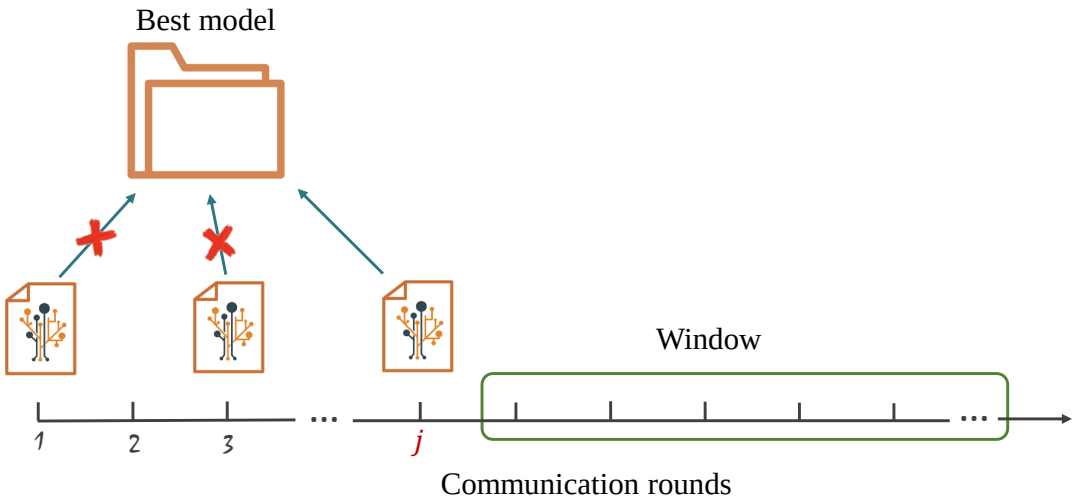


Figure 6.4: Illustration of the delta-based FL stopping algorithm; the current j communication round has a better model that replaces the previous one, emptying the window.

Scikit-Learn stopping criterion, originally developed to stop the training process within the model [165]. Instead of fixing the number of communication rounds, the delta-based FL stopping algorithm sets a threshold (delta) for comparing the currently trained model's performance with the best model from previous rounds, based on the validation dataset, as depicted in Fig. 6.4. If the current model outperforms the previously best-stored model, it replaces it; otherwise, the best model remains unchanged. This comparison runs for several rounds, defined as a window size. If the window is filled in without a better model being found, the training stops and returns the best model. This algorithm has been implemented in both the FedAvg and FedraTrees frameworks after an extensive study to determine the best delta and window size values. More details will be provided in the following section.

6.4 Performance Evaluation and Results

6.4.1 Dataset Pre-processing and Evaluation Methods

Since this study aims to provide a practical framework capable of forecasting power consumption, it is essential to find an excellent dataset to evaluate the performance of FedraTrees versus FedAvg. Notably, there is a noticeable gap in the existing literature concerning the integration of energy forecasting for enhancing energy efficiency in UDNs. Consequently, and taking into account that energy forecasting requires load profiles regardless of the domain or area under investigation, this study examines energy forecasting approaches within the context of smart grids. In this regard, the Tetouan power consumption dataset is selected for this purpose [166]. This dataset was collected

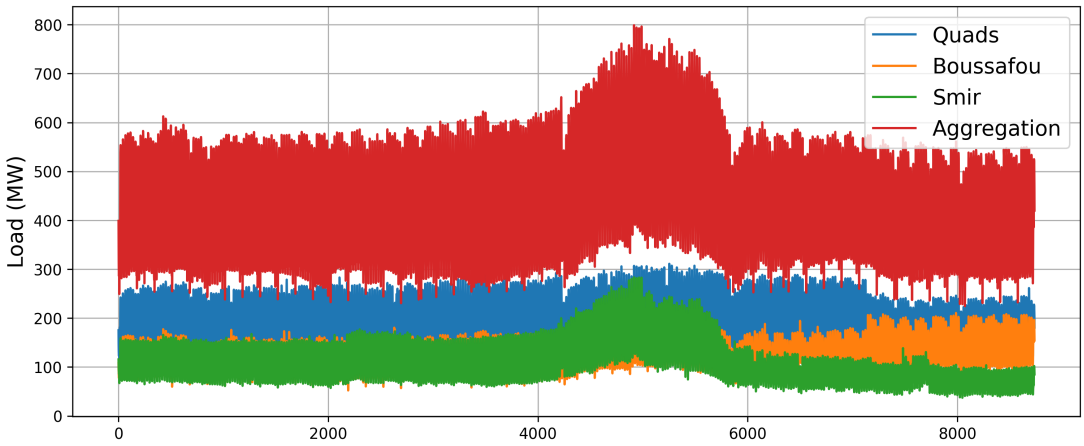


Figure 6.5: Tetouan dataset preparation generated hourly power consumption of the three zones in addition to the aggregated power.

Table 6.1: Tetouan dataset features used for load forecasting

Context	**Features**
Calendar	① Month ② Day ③ Hour
Weather	④ Temperature ⑤ Humidity ⑥ Wind speed ⑦ Diffuse flow ⑧ General diffuse flow
Power	⑨ PrevHourAgg

in 2017 at three different distribution substations from the zones: Quads, Boussafou, and Smir in Tetouan, a city located in northern Morocco. In addition to providing information on power consumption every 10 minutes, the Tetouan dataset offers complementary data on the calendar and weather conditions. To prepare this dataset for this study, the time scale is converted from 10 minutes to 60 minutes because the focus is on predicting short-term loads for the next hour. Furthermore, two new dataset features are created; the aggregation feature that aggregates the power consumption of the three zones for use while performing centralised and distributed learning, and the previous hour aggregation (PrevHourAgg) feature that gives the aggregated feature reading of the previous hour. Fig. 6.5 demonstrates the hourly aggregated load as well as zones-specific load information. Moreover, Table. 6.1 shows the features considered for load forecasting.

Before commencing the training process, the default features scale is normalised using MinMax scaler, bringing it into the range [0,1]. This ensures that all features have an equal opportunity to contribute to model fitting while avoiding bias. For evaluation purposes, the most popular metrics used are MAE and mean absolute percentage error (MAPE), as defined in [167]:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|\,, \tag{6.1}$$

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{6.2}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and m represents the number of data points.

6.4.2 Simulation Setup

To evaluate the effectiveness of the proposed FedraTrees framework in providing accurate energy forecasting, a comparative analysis is conducted against the widely adopted LSTM-based FedAvg algorithm. Furthermore, a comparison is made between both algorithms and the Persistence approach, which predicts the current value to be identical to the preceding actual value. The forecasting problem is converted to a multivariate regression problem leveraging various calendar, weather, and power features to predict the power consumption for the next hour. For this purpose, LSTM and LGBM models are constructed based on the Random search strategy to discover the best combination of hyperparameters values. The LSTM model consists of a single hidden layer with 64 LSTM units that use the rectified linear unit (ReLU) activation function and a dense output layer with one neuron. Also, it uses Adam optimiser for compilation, with the dataset divided into 80/20% train/test split ratio, batch size equals 30, and the number of epochs equals 300. Whereas the LGBM's boosting type, the number of trees[1], max_depth, learning rate, num_leaves, and train/test split ratio are set to DART, 800, 12, 0.078, 30, and 80/20%, respectively. The simulation experiments are based on Python programs installed on a Windows operating system with Intel Xeon CPU E5-2620 @ 2GHz and 16GB RAM.

6.4.3 Numerical Results

The performance of the selected models in the conventional approach of centralised training is initially investigated. Table 6.2 shows the results of the evaluation metrics and the computation time required for each model. The MAE and MAPE for the LSTM model are 0.02 and 3.04%, respectively, while they are improved when using the LGBM model and reached 0.017 and 2.69%, respectively. Notably, the LGBM model's computational efficiency is confirmed in this table, as it converges in just two seconds, while the LSTM model requires more than 97% additional computations compared to the LGBM model.

Furthermore, a study is conducted to explore the impact of each feature in predicting the target value. The LGBM model is equipped with a feature importance tool that can be used to find out what features contribute the most to the prediction of power values. Fig. 6.6 demonstrates the results of the feature importance study, revealing

[1]This hyperparameter is replaced by batch in FedraTrees.

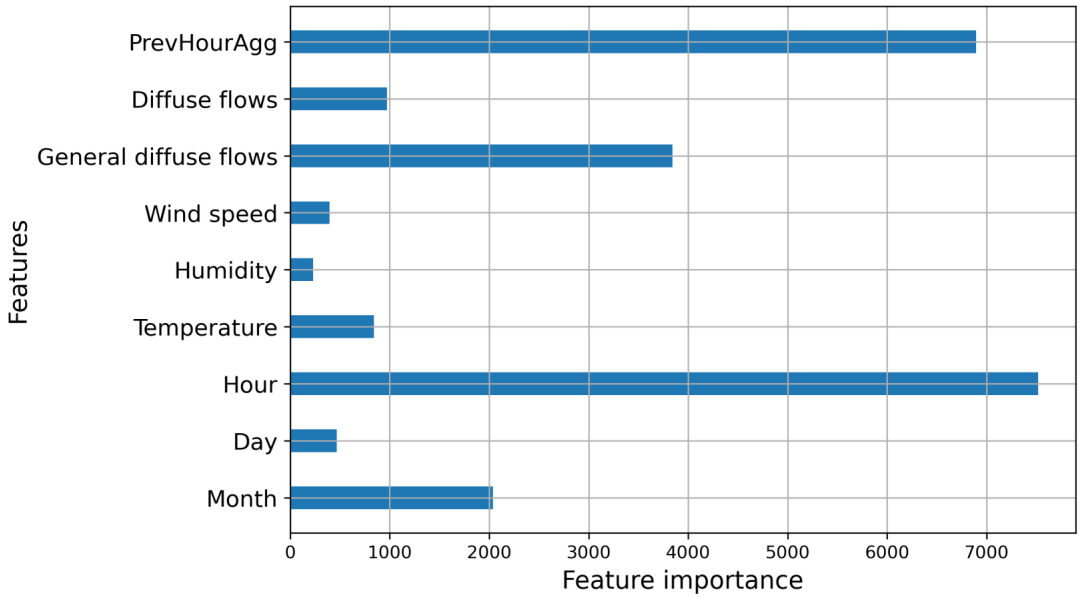


Figure 6.6: The importance of each feature in forecasting power consumption.

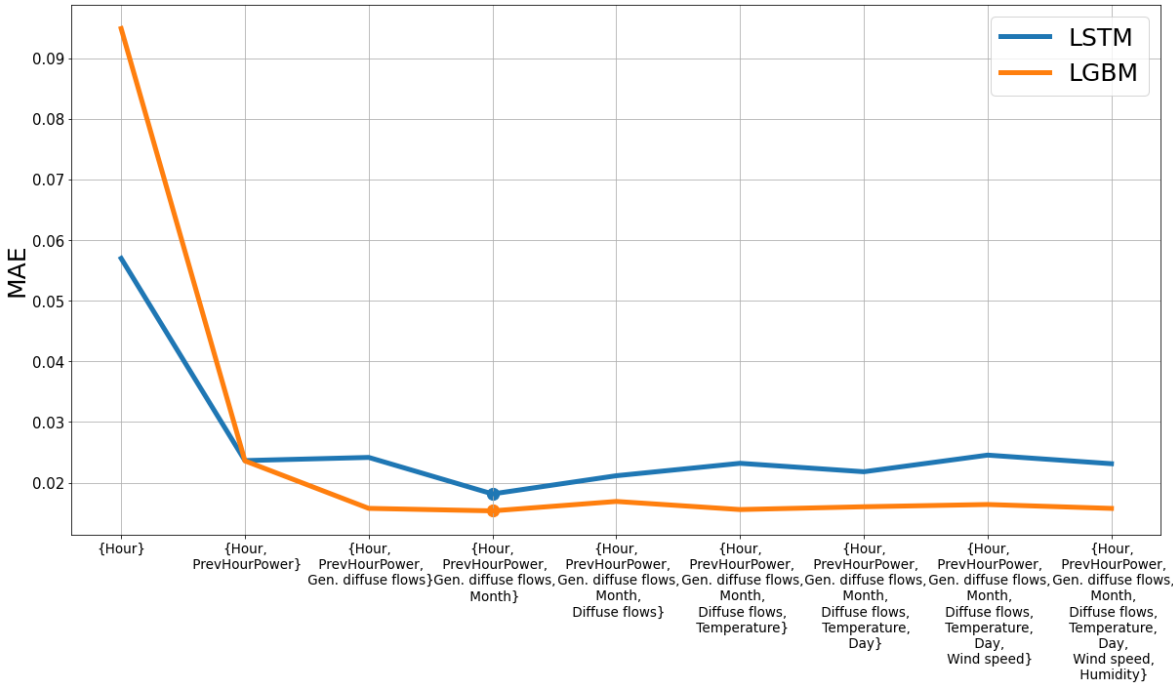


Figure 6.7: MAE as a function of different features that contribute to the prediction process.

Table 6.2: Performance comparison between LSTM and LGBM models when performing centralised model training.

ML model	MAE	MAPE	Computation time
LSTM	0.019	3.04%	77 seconds
LGBM	0.017	2.69%	2 seconds

that Hour, PrevHourAgg, General diffuse flows, and Month have the most significant impact in predicting the power consumption values. Other features also contribute to the prediction, but their impact is less pronounced. Furthermore, another study is conducted to examine the prediction performance when using several numbers of features for the LSTM and LGBM models. Fig. 6.7 shows the best achieved MAE versus using different numbers of features, ordered based on their rank in the feature importance study. This figure indicates that, in general, multivariate prediction gives improved performance over univariate. However, for both models, the best performance is obtained when the four most important features are used.

Moving on to the FL setup, the FedraTrees and FedAvg comprise a central orchestrating server and three clients representing the three zones of the Tetouan dataset. The number of communication rounds is not fixed, as the delta-based FL stopping technique, discussed in Section 6.3.2, is used to find the best round that produces the optimal trained model while reducing the computation and communication costs. An extensive study is carried out to determine the best delta and window size values for the stopping algorithm. The findings of this study are given in Tables 6.3 and 6.4. From Table 6.3, the most computationally efficient and the best MAE/MAPE for the LSTM model are obtained when the delta and window size values are 0.00001 and 55, respectively. Regarding LGBM, Table 6.4 demonstrates that the best values for the delta and window size are 0.00001 and 10, respectively. Table 6.5 summarises the best results obtained for each of the Persistence, FedAvg, and FedraTrees algorithms. The performance of the Persistence model is poor compared to other algorithms, and the FedAvg has the best performance, slightly outperforming FedraTrees. However, FedraTrees excels over the FedAvg algorithm in terms of the communication rounds and the required computations. FedraTrees only requires 65 rounds of communications that result in approximately 26 seconds of computations, while FedAvg requires a much higher number of communication rounds and computation time by a factor of 7.6 and 52.2, respectively, to achieve the same level of performance.

Furthermore, Fig. 6.8 displays the convergence curve of MAE for both algorithms during the training of the global model. In addition, Fig. 6.9 shows the actual and the predicted power consumption of both algorithms and the baseline Persistence model. These figures indicate that FedraTrees converges faster and achieves an outstanding

Table 6.3: A study to determine the best values of delta and window size for LSTM-based FedAvg.

Delta	Window size	MAE	MAPE %	No. of rounds	Computation time (sec)
0.001	45	0.0217	4.62	106	214.2
0.0001	45	0.0214	4.55	133	277.5
0.00001	45	0.0171	3.72	401	1284
0.000001	45	0.0171	3.72	401	1284
0.001	55	0.0217	4.62	106	214.2
0.0001	55	0.0195	4.2	235	787
0.00001	**55**	**0.0157**	**3.43**	**491**	**1356**
0.000001	55	0.0157	3.43	491	1356
0.001	65	0.0217	4.62	106	214.2
0.0001	65	0.0157	3.43	491	1383.6
0.00001	65	0.0157	3.43	491	1383.6
0.000001	65	0.0157	3.43	491	1383.6

Table 6.4: A study to determine the best values of delta and window size for LGBM-based FedraTrees.

Delta	Window size	MAE	MAPE%	No. of rounds	Computation time (sec)
0.001	5	0.0199	4.25	8	2.0
0.0001	5	0.0177	3.79	26	7.6
0.00001	5	0.0175	3.75	35	10.3
0.000001	5	0.0175	3.75	35	10.3
0.001	10	0.0186	3.99	14	4.5
0.0001	10	0.0174	3.7	50	18.8
0.00001	**10**	**0.0173**	**3.69**	**65**	**26.2**
0.000001	10	0.0173	3.69	65	26.2
0.001	15	0.0186	3.99	14	4.5
0.0001	15	0.0174	3.7	50	18.8
0.00001	15	0.0173	3.69	65	27.2
0.000001	15	0.0173	3.69	65	27.2

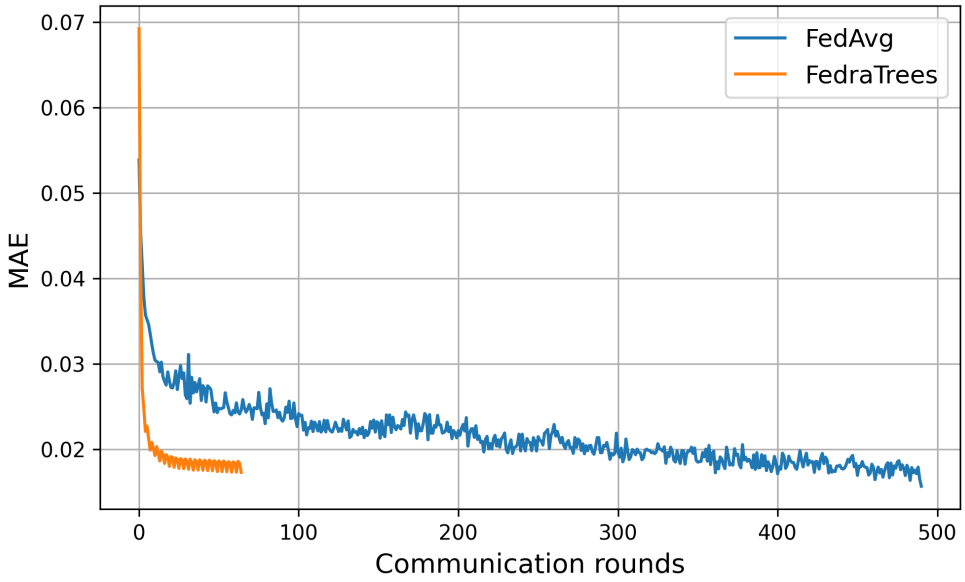


Figure 6.8: MAE as a function of the communication rounds needed to train the global model of FedAvg and FedraTrees.

performance compared to FedAvg.

Another study is conducted, focusing solely on the four most important features to assess the impact of using fewer features on the required number of communication rounds, the computation time, and the model performance. Table 6.6 presents the outcomes of this study and indicates that the performance of FedraTrees is improved when the less important features are removed. However, this is not the case with FedAvg, as this table shows that its performance is slightly degraded compared to using all the features. The number of communication rounds is slightly less than that in the full feature study for both algorithms. However, this study also highlights the outstanding performance of FedraTrees, as it requires significantly fewer rounds of communications and reduced computation costs. Similarly, Figs. 6.10 and 6.11 give the MAE convergence curve and the actual and forecasted power consumption for both algorithms, respectively. Also, these figures ensure the superb overall performance of FedraTrees.

Table 6.5: Performance results of the FedraTrees compared to the FedAvg and the Persistence model.

Algorithm	**MAE**	**MAPE%**	**No. of rounds**	**Computation time**
Persistence	0.08	6.64	N/A	N/A
FedAvg	0.0157	3.43	491	1356 seconds
FedraTrees	0.0173	3.69	65	26.2 seconds

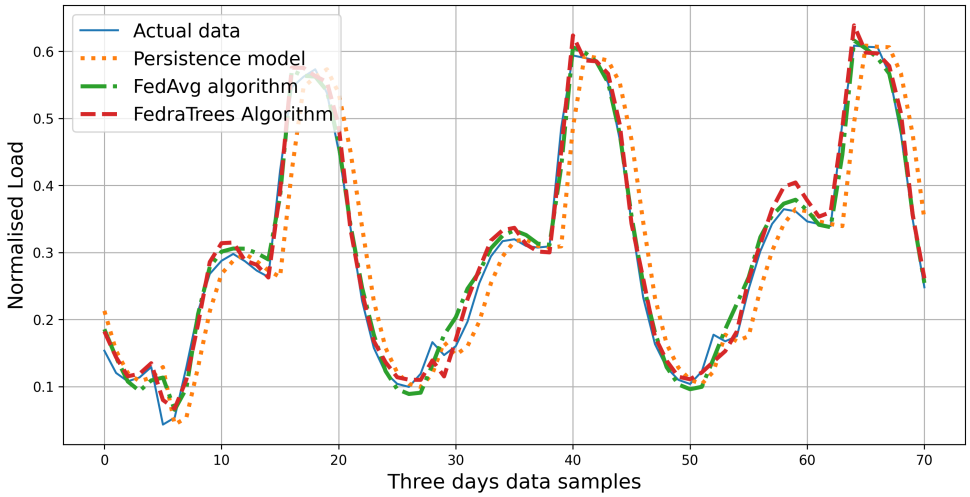


Figure 6.9: Forecasting power consumption for three days.

Table 6.6: Performance results of the FedraTrees compared to the FedAvg and Persistence models when considering only the top four features.

Algorithm	**MAE**	**MAPE%**	**No. of rounds**	**Computation time**
Persistence	0.08	6.64	N/A	N/A
FedAvg	0.0177	3.93	465	1293 seconds
FedraTrees	0.0168	3.54	50	8.8 seconds

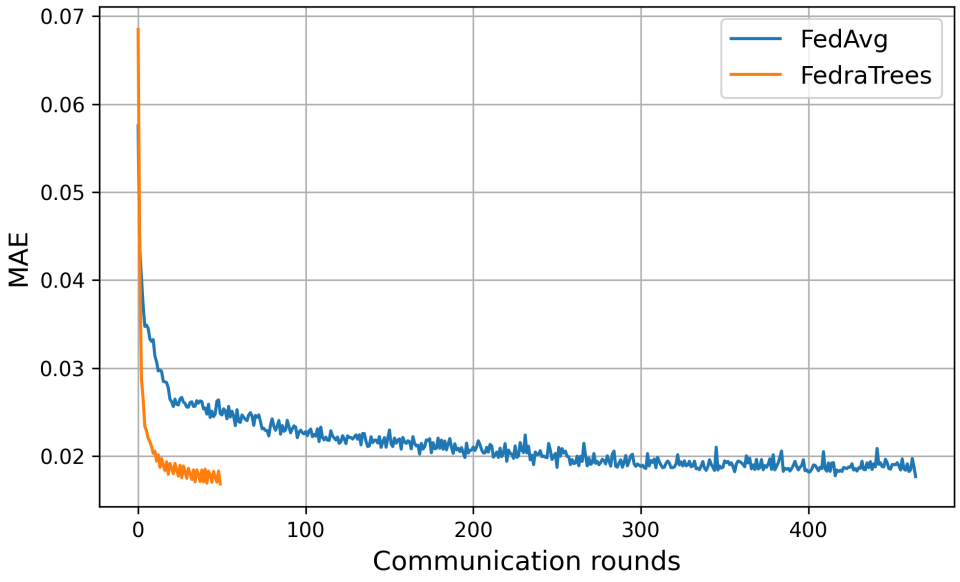


Figure 6.10: MAE as a function of the communication rounds needed to train the global model of FedAvg and FedraTrees when considering only the top four features.

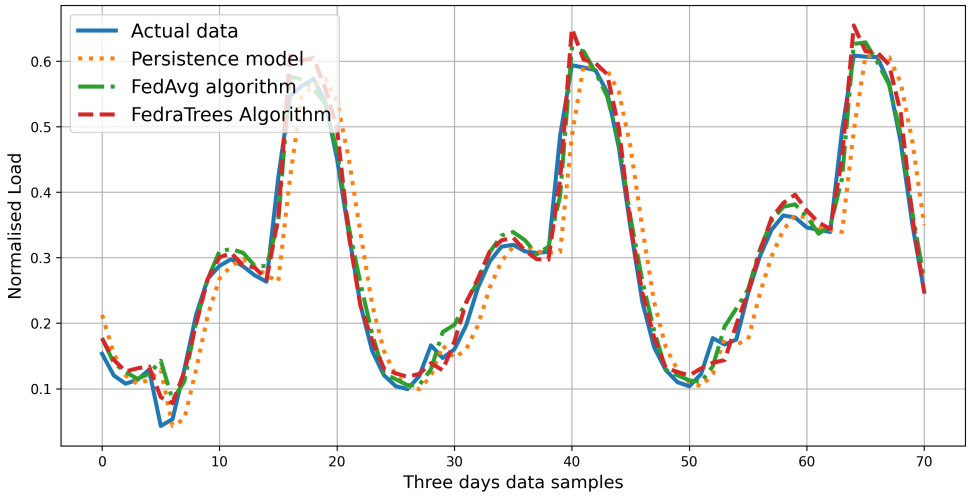


Figure 6.11: Forecasting power consumption for three days when considering only the top four features.

6.5 Summary

This chapter aimed at developing FedraTrees, a novel framework that incorporates ensemble learning, specifically the LGBM model, within the FL settings. Utilising the LGBM model transforms the FL into a highly efficient, fast-processing, and scalable framework. Furthermore, instead of following the conventional fixed number of communication rounds method in FL, a delta-based FL stopping algorithm is developed to monitor and stop the FL training process when no further enhancement is possible, thus ensuring achieving the desired training accuracy with the minimal use of computation and communication resources. FedraTrees is employed for the energy demand forecasting problem and benchmarked against LSTM-based FedAvg and Persistence model. The simulation results demonstrated that FedraTrees has a remarkable performance in predicting short-term energy patterns and requires much less computation and communication compared to FedAvg, with only 2% and 13%, respectively.

Chapter 7

Conclusions and Future Works

The conclusions drawn from each part of this thesis are summarised in this chapter. It also highlights potential avenues for further exploration and expansion of this research.

7.1 Concluding Remarks

In Chapter 3, novel techniques have been explored to address the challenges associated with beam blockages and frequent HOs in next-generation wireless networks, specifically within the mmWave and THz bands. Within this chapter, a CV-assisted PHO framework has been introduced, harnessing visual information to augment network awareness and predict blocking events (BLOCKs) stemming from obstacles and users' movements. Through the combination of object detection and multivariate regression models, along with performing a centralised training method, precise predictions have been shown for obstacle and user locations, in addition to estimating the remaining time until users reach obstructed areas. Furthermore, the chapter has undertaken an analysis aimed at prioritising QoE by examining the optimal location and timing for performing HO, effectively mitigating QoE degradation. Simulation results have demonstrated a significant improvement in sustaining user connectivity and QoE, thus underscoring the effectiveness of the framework in addressing these pivotal challenges. This research aligns with the vision of enabling low-latency and time-sensitive applications in B5G and 6G networks, making future UDNs more dynamic and responsive to their surroundings.

In Chapter 4, the research has also delved into the use of CV and ML to enhance the reliability and latency within more complex high-frequency communication systems, characterised by numerous dynamic users and obstacles. A novel framework, aided by CV technology, has been developed to proactively identify potential blockage scenarios and trigger PHOs at the optimal time and distance from the blocked area. This framework leverages the ODL algorithm along with a NN model to accurately predict blockages and estimate the time until users reach obstructed areas. Furthermore, the adoption

of FL ensures decentralised and collaborative model training, safeguarding data privacy and conserving BW resources. Simulation results have demonstrated the effectiveness of this approach, achieving an impressive PHO success rate and outperforming reactive-HO benchmarks in terms of latency, all while maintaining a high QoE for users. This research has shown promising solutions for mitigating blockages in multi-user mmWave/THz networks, paving the way for more reliable and efficient high-frequency communications.

In Chapter 5, a novel RaDaR framework has been developed to address the challenges of frequent beam blockages in high-frequency outdoor networks. The framework employs radar sensors to monitor object movement and generate range-angle and range-velocity maps for scene analysis and predictions. Radar measurements and FL have been utilised to train a dual-output NN model capable of predicting blockage status and time simultaneously. This predictive capability allows for PHO or beam switching, reducing latency and ensuring high-quality user experiences. FL brings advantages such as privacy protection, resource efficiency, scalability, and knowledge sharing. Evaluation results, based on a comprehensive real-world dataset, demonstrate RaDaR's substantial improvement in network reliability, achieving a higher average PHO success rate compared to reactive HO procedures. RaDaR has been shown to maintain a superior quality of experience by sustaining high throughput levels and minimising PHO latency, presenting a promising solution for mmWave and THz network dependability in dynamic outdoor environments.

In Chapter 6, a novel framework called FedraTrees has been developed to achieve the energy efficiency requirement of next-generation wireless networks. FedraTrees leverages ensemble learning, particularly the LGBM model, within the FL framework, resulting in an efficient, fast-processing, and scalable solution. The delta-based FL stopping algorithm has been introduced to monitor and optimise the training process, reducing computational and communication costs while ensuring accurate energy forecasting. Additionally, a feature importance evaluation study has been conducted to enhance forecasting performance further. FedraTrees has been benchmarked against LSTM-based FedAvg and the Persistence model, demonstrating superior performance with significantly reduced computational and communication resource requirements, as shown in the simulation results.

In conclusion, this thesis has focused on addressing two critical aspects of next-generation high-frequency wireless networks: mobility and energy management. The primary focus was on mitigating the beam blockage problem prevalent in these networks, for which various frameworks were proposed. These frameworks leveraged sensing-aided wireless communications, integrating multimodal information such as visual, radar fingerprint, and wireless data to effectively address beam blockages. Moreover, the use of ML models, trained through both centralised and decentralised mechanisms, provided

valuable insights into the strengths and limitations of each training mechanism. Simulation results underscored the efficacy of these frameworks in enhancing network reliability, thereby enabling the support of latency-sensitive applications. Additionally, the thesis delved into energy management aspects, proposing a lightweight FL aggregation algorithm aimed at achieving the energy efficiency requirements of these networks. By predicting power consumption and enabling proactive measures to avoid energy wastage, the proposed algorithm contributes to the optimisation of network energy usage. Overall, the research presented in this thesis contributes significantly to advancing the capabilities of next-generation high-frequency wireless networks, making them more reliable, efficient, and conducive to supporting a wide range of applications.

7.2 Future Work

The fields of mobility and energy management within UDNs offer extensive research opportunities, with unexplored potential applications. The contributions outlined in this thesis represent the first steps towards multiple research avenues aimed at enhancing UDNs. In this section, a few interesting research topics are highlighted, all of which have the potential to improve the performance of the various optimisation frameworks proposed in this thesis.

7.2.1 CV-aided PHO Frameworks Under Low-quality Imaging.

The proposed PHO frameworks, which rely on visual information to augment the network's situational awareness, have demonstrated exceptional capabilities in detecting a wide range of objects, including users and blockages. These frameworks effectively notify the network of the potential need for a PHO following the detection of BLOCK events. However, an implicit assumption in these frameworks is that vision cameras remain unaffected by changes in ambient lighting conditions, including transitions between day and night, as well as variations in weather conditions. This assumption is primarily based on the ViWi datasets, which were predominantly generated under favorable lighting and weather conditions. The vision-assisted frameworks presented in this thesis rely on YOLO object detection and localisation models, and several studies have shown that such models struggle to detect objects accurately under low-light and adverse weather [168, 169]. Low-light and adverse weather conditions pose imaging difficulties, resulting in increased image noise. Fortunately, the literature offers several insightful solutions to address this issue. For instance, the study in [169] achieved a remarkable improvement in image quality by introducing a novel image processing pipeline, featuring a specially designed CNN model to address the challenges of low-light photography. Additionally, the study in [170] presented a patch-base image restoration algorithm based on diffusion probabilistic models.

Incorporating such advanced image denoising approaches into the CV-assisted PHO frameworks proposed in this thesis holds the potential to strengthen their resilience under low-light and adverse weather. Therefore, a natural extension of this research involves investigating the integration of such solutions into the proposed frameworks. This exploration should also encompass an evaluation of their impact on framework timing, the determination of the optimal PHO execution time, and their potential influence on the overall PHO success rate and the QoE.

7.2.2 Exploring RIS and UAV Technologies as Alternatives for PHO

In the proposed sensing-aided PHO frameworks, the main focus is on accurately predicting BLOCK events to alert the network of the necessity for the performing PHO for wireless users, transitioning them from the serving base station (S-BS) to the target base station (T-BS). During HO, the T-BS might receive requests from multiple devices to access the network simultaneously, leading to users competing for available resources. Consequently, the proposed frameworks take into account the worst-case scenario of contention-based random access, which demands approximately 70ms more execution time compared to contention-free random access [114]. While the solutions presented in this thesis have shown remarkable performance in supporting real-time applications and maintaining high QoE levels, the latency associated with preparing for HO to another SBS can be significantly reduced by leveraging RIS and UAV technologies. Instead of transferring the user to another SBS, it would be much latency-conscious to retain the user under the same S-BS while exploring novel methodologies to circumvent blockages and service disconnection.

In light of this, adopting RIS and UAV technologies is anticipated to improve the proposed frameworks. For instance, one or more RIS elements could be strategically deployed within UDNs based on investigations to eliminate blocked regions by creating virtual LoS links. When a blockage is detected, instead of switching the user to another SBS, the S-BS could leverage the employed RIS(s) to establish a virtual LoS and continue serving the user. Similarly, another research direction could involve the utilisation of UAVs as relays to form virtual links and serve the user when reaching the shadowed area. However, it is essential to carefully investigate the framework's timing and how these methodologies impact the success rate of switching to virtual links and, subsequently, how they affect the QoE.

7.2.3 Enhancing FedraTrees Through Adaptive Feature Selection

An exciting research direction for improving FedraTrees framework is the development of adaptive feature selection techniques. The nature of load profiles in energy forecasting can exhibit temporal variations, causing certain features to become more or less relevant over time. In this context, future research can explore the implementation of dynamic feature selection mechanisms that automatically identify and select the most informative features during the FL training process. This adaptability paves the way for more efficient models that can seamlessly respond to changing data patterns, ensuring accurate forecasts even in dynamic energy consumption scenarios. Furthermore, considering the computational constraints frequently encountered on edge devices in FL setups, research can delve into innovative feature engineering techniques designed to reduce the dimensionality of data while preserving critical information. This extension to adaptive feature selection not only promises to enhance the overall performance of FedraTrees but also holds the potential to significantly reduce training times. By tackling the challenges posed by evolving load profiles, adaptive feature selection emerges as a pivotal strategy for more precise and agile energy forecasting.

Bibliography

[1] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 1201–1221, Apr. 2017.

[2] D. Gil *et al.*, "Internet of things: A review of surveys based on context aware intelligent services," *Sensors*, vol. 16, no. 7, p. 1069, July 2016.

[3] Z. Rehena, "Internet of things: challenges and its applications," in *Interoperability in IoT for Smart Syst.* CRC Press, Dec. 2020, pp. 1–14.

[4] X. You *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inform. Sci.*, vol. 64, no. 1, pp. 1–74, Jan. 2021.

[5] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surv. Tuts.*, vol. 18, no. 4, pp. 2522–2545, May 2016.

[6] H. N. Dai *et al.*, "Big data analytics for large scale wireless networks: Challenges and opportunities," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, Sept. 2019.

[7] M. Obschonka and D. B. Audretsch, "Artificial intelligence and big data in entrepreneurship: a new era has begun," *Small Business Economics*, vol. 55, pp. 529–539, Oct. 2020.

[8] P. Li *et al.*, "Multi-key privacy-preserving deep learning in cloud computing," *Future Generation Comput. Syst.*, vol. 74, pp. 76–85, Sept. 2017.

[9] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network NR, "Radio Resource Control (RRC) protocol specification," *in 3GPP TS 38.331 version 16.6.0 Release 16*, Sept. 2021.

[10] G. Yu *et al.*, "A hierarchical SDN architecture for ultra-dense millimeter-wave cellular networks," *IEEE Commun. Maga.*, vol. 56, no. 6, pp. 79–85, June 2018.

[11] S. Collonge, G. Zaharia, and G. E. Zein, "Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel," *IEEE Trans. Wireless Commun.*, vol. 3, no. 6, pp. 2396–2406, Nov. 2004.

[12] A. Yamamoto *et al.*, "Path-loss prediction models for intervehicle communication at 60 GHz," *IEEE Trans. veh. technol.*, vol. 57, no. 1, pp. 65–78, Jan. 2008.

[13] A. Mughees *et al.*, "Energy-efficient ultra-dense 5G networks: recent advances, taxonomy and future research directions," *IEEE Access*, vol. 9, pp. 147 692–147 716, Oct. 2021.

[14] S. Henry, A. Alsohaily, and E. S. Sousa, "5G is real: Evaluating the compliance of the 3GPP 5G new radio system with the ITU IMT-2020 requirements," *IEEE Access*, vol. 8, pp. 42 828–42 840, Mar. 2020.

[15] J. Liu, W. Xiao, A. C. Soong, A. Anpalagan, M. Bennis, and R. Vannithamby, "Dense networks of small cells." 2015.

[16] S. H. A. Shah, M. Sharma, and S. Rangan, "LSTM-based multi-link prediction for mmWave and sub-THz wireless systems," in *Proc. IEEE Int. Conf. Commun. (ICC), Dublin, Ireland*. IEEE, June 2020, pp. 1–6.

[17] A. E. Kalør, O. Simeone, and P. Popovski, "Prediction of mmWave/THz link blockages through meta-learning and recurrent neural networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 12, pp. 2815–2819, Oct. 2021.

[18] S. Moon *et al.*, "Deep neural network for beam and blockage prediction in 3GPP-based indoor hotspot environments," *Wireless Pers. Commun.*, vol. 124, no. 4, pp. 3287–3306, June 2022.

[19] A. Bonfante *et al.*, "Performance of predictive indoor mmWave networks with dynamic blockers," *IEEE Trans. Cogn. Commun. Netwo.*, vol. 8, no. 2, pp. 812–827, Oct. 2021.

[20] C. Jia *et al.*, "Machine learning empowered beam management for intelligent reflecting surface assisted MmWave networks," *China Commun.*, vol. 17, no. 10, pp. 100–114, Oct. 2020.

[21] A. Aldalbahi, F. Shahabi, and M. Jasim, "Instantaneous beam prediction scheme against link blockage in mmWave communications," *Appl. Sci.*, vol. 11, no. 12, p. 5601, June 2021.

[22] B. N. Alsunbuli, W. Ismail, and N. M. Mahyuddin, "Convolutional neural network and Kalman filter-based accurate CSI prediction for hybrid beamforming under a minimized blockage effect in millimeter-wave 5G network," *Appl. Nanoscience*, pp. 1–22, Sept. 2021.

[23] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, June 2020.

[24] X. Li *et al.*, "Diffraction Characteristics Aided Blockage and Beam Prediction for mmWave Communications," in *IEEE 95th Veh. Technol. Conf. (VTC2022-Spring), Helsinki, Finland*. IEEE, June 2022, pp. 1–5.

[25] S. Ohta *et al.*, "Point Cloud-based Proactive Link Quality Prediction for Millimeter-wave Communications," *arXiv preprint arXiv:2301.00752*, Jan. 2023. [Online]. Available: http://arxiv.org/abs/2301.00752

[26] S. Wu, C. Chakrabarti, and A. Alkhateeb, "Lidar-aided mobile blockage prediction in real-world millimeter wave systems," in *IEEE Wireless Commun. Netw. Conf. (WCNC), Austin, TX, USA*. IEEE, Apr. 2022, pp. 2631–2636.

[27] Y. Yang *et al.*, "Environment Semantics Aided Wireless Communications: A Case Study of mmWave Beam Prediction and Blockage Prediction," *arXiv preprint arXiv:2301.05837*, Jan. 2023. [Online]. Available: http://arxiv.org/abs/2301.05837

[28] 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, Aug. 2019.

[29] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided dynamic blockage prediction for 6G wireless communication networks," in *IEEE Int. Conf. Commun. Workshops (ICC Workshops), Montreal, QC, Canada*. IEEE, June 2021, pp. 1–6.

[30] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *Proc. IEEE 91st Vehic. Technol. Conf. (VTC2020-Spring), Antwerp, Belgium*, May 2020, pp. 1–5.

[31] Z. Xiong *et al.*, "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," *IEEE Vehic. Technol. Mag.*, vol. 14, no. 2, pp. 44–52, Apr. 2019.

[32] L. Zhang *et al.*, "Energy efficient resource optimization in user-centric UDNs with NOMA and beamforming," in *6th Int. Conf. Comput. Commun. (ICCC), Chengdu, China*. IEEE, Dec. 2020, pp. 115–121.

[33] M. Amine, A. Kobbane, and J. Ben-Othman, "New network slicing scheme for UE association solution in 5G ultra dense HetNets," in *ICC Int. Conf. Commun. (ICC), Dublin, Ireland*. IEEE, June 2020, pp. 1–6.

[34] G. Ciulla and A. D'Amico, "Building energy performance forecasting: A multiple linear regression approach," *Appl. Energy*, vol. 253, p. 113500, Nov. 2019.

[35] S. Saab, E. Badr, and G. Nasr, "Univariate modeling and forecasting of energy consumption: the case of electricity in Lebanon," *Energy*, vol. 26, no. 1, pp. 1–14, Jan. 2001.

[36] B. Singh and D. Pozo, "A guide to solar power forecasting using ARMA models," in *Proc. IEEE PES Innov. Smart Grid Technol. Eur. (ISGT-Europe), Bucharest, Romania*, Sept. 2019, pp. 1–4.

[37] G. A. Darbellay and M. Slama, "Forecasting the short-term demand for electricity: Do neural networks stand a better chance?" *Int.l J. Forecasting*, vol. 16, no. 1, pp. 71–83, Jan. 2000.

[38] J. A. Zancanaro *et al.*, "Energy consumption forecasting using SARIMA and NARNET: An actual case study at university campus," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf.-Latin America (ISGT Latin America), Gramado, Brazil*. IEEE, Sept. 2019, pp. 1–6.

[39] F.-Y. Ju and W.-C. Hong, "Application of seasonal SVR with chaotic gravitational search algorithm in electricity forecasting," *Appl. Math. Model.*, vol. 37, no. 23, pp. 9643–9651, Dec. 2013.

[40] I. Koprinska, D. Wu, and Z. Wang, "Convolutional neural networks for energy time series forecasting," in *Proc. int. joint conf. neural netw. (IJCNN), Rio de Janeiro, Brazil*, July 2018, pp. 1–8.

[41] P.-H. Kuo and C.-J. Huang, "A high precision artificial neural networks model for short-term energy load forecasting," *Energies*, vol. 11, no. 1, p. 213, Jan. 2018.

[42] J. Q. Wang, Y. Du, and J. Wang, "LSTM based long-term energy consumption prediction with periodicity," *Energy*, vol. 197, p. 117197, Apr. 2020.

[43] K. Yan *et al.*, "A hybrid LSTM neural network for energy consumption forecasting of individual households," *IEEE Access*, vol. 7, pp. 157 633–157 642, Oct. 2019.

[44] G. K. Tso and K. K. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, Sept. 2007.

[45] Z. Yu *et al.*, "A decision tree method for building energy demand modeling," *Energy and Buildings*, vol. 42, no. 10, pp. 1637–1646, Oct. 2010.

[46] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and buildings*, vol. 147, pp. 77–89, July 2017.

[47] S. Liu *et al.*, "Short-term load forecasting based on GBDT combinatorial optimization," in *Proc. IEEE Conf. Energy Internet and Energy Syst. Integr. (EI2), Beijing, China*, Oct. 2018, pp. 1–5.

[48] R. A. Abbasi *et al.*, "Short term load forecasting using XGBoost," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.* Springer, Mar. 2019, pp. 1120–1131.

[49] A. Taïk and S. Cherkaoui, "Electrical load forecasting using edge computing and federated learning," in *Proc. IEEE Int. Conf. Commun. (ICC), Dublin, Ireland*, June 2020, pp. 1–6.

[50] M. Savi and F. Olivadese, "Short-term energy consumption forecasting at the edge: A federated learning approach," *IEEE Access*, vol. 9, pp. 95 949–95 969, July 2021.

[51] C. Briggs, Z. Fan, and P. Andras, "Federated learning for short-term residential energy demand forecasting," *arXiv preprint arXiv:2105.13325*, May 2021. [Online]. Available: http://arxiv.org/abs/2105.13325

[52] N. Gholizadeh and P. Musilek, "Federated learning with hyperparameter-based clustering for electrical load forecasting," *Internet Things*, vol. 17, p. 100470, Mar. 2022.

[53] M. N. Fekri, K. Grolinger, and S. Mir, "Distributed load forecasting using smart meter data: Federated learning with recurrent neural networks," *Int. J. Elect. Power & Energy Syst.*, vol. 137, p. 107669, May 2022.

[54] M. Hao *et al.*, "Efficient and privacy-enhanced federated learning for industrial artificial intelligence," *IEEE Trans. on Ind. Informat.*, vol. 16, no. 10, pp. 6532–6542, Oct. 2019.

[55] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.* PMLR, Apr. 2017, pp. 1273–1282.

[56] M. Salehi *et al.*, "Hierarchical federated learning across heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP), Barcelona, Spain*, May 2020, pp. 8866–8870.

[57] Q. Li *et al.*, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *arXiv:1907.09693*, Jan. 2021. [Online]. Available: http://arxiv.org/abs/1907.09693

[58] H. B. McMahan *et al.*, "Federated learning of deep networks using model averaging," *arXiv:1602.05629*, vol. 2, Feb. 2016. [Online]. Available: http://arxiv.org/abs/1602.05629

[59] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," *arXiv:1902.01046*, Feb. 2019. [Online]. Available: https://arxiv.org/abs/1902.01046

[60] F. Lai *et al.*, "Oort: Informed participant selection for scalable federated learning," *arXiv preprint arXiv:2010.06081*, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.06081

[61] M. R. Sprague *et al.*, "Asynchronous federated learning for geospatial applications," in *Proc. Conf. Mach. Learn. Knowl. Discov. Databases.* Springer, Nov. 2018, pp. 21–28.

[62] J. Nguyen *et al.*, "Federated learning with buffered asynchronous aggregation," in *Int. Conf. Artif. Intell. Stat.* PMLR, May 2022, pp. 3581–3607.

[63] M. Chen, B. Mao, and T. Ma, "FedSA: A staleness-aware asynchronous federated learning algorithm with non-IID data," *Future Generation Computer Systems*, vol. 120, pp. 1–12, July 2021.

[64] W. Wu *et al.*, "SAFA: A semi-asynchronous protocol for fast federated learning with low overhead," *IEEE Trans. Computers*, vol. 70, no. 5, pp. 655–668, May 2020.

[65] T. Li *et al.*, "Federated optimization in heterogeneous networks," *in Proc. Mach. learn. syst.*, vol. 2, pp. 429–450, Mar. 2020.

[66] R. Pathak and M. J. Wainwright, "FedSplit: An algorithmic framework for fast federated optimization," *arXiv preprint arXiv:2005.05238*, May 2020. [Online]. Available: https://arxiv.org/abs/2005.05238

[67] S. Reddi *et al.*, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, Feb. 2020. [Online]. Available: https://arxiv.org/abs/2003.00295

[68] D. Basu *et al.*, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[69] J. Hamer, M. Mohri, and A. T. Suresh, "Fedboost: A communication-efficient algorithm for federated learning," in *Proc. Int. Conf. Mach. Learn.* PMLR, Nov. 2020, pp. 3973–3983.

[70] M. Al-Quraan, A. Khan, A. Centeno, A. Zoha, M. A. Imran, and L. Mohjazi, "FedraTrees: A novel computation-communication efficient federated learning framework investigated in smart grids," *Eng. Appl. Artif. Intell.*, vol. 124, p. 106654, Sept. 2023.

[71] A. Aissioui *et al.*, "On enabling 5G automotive systems using follow me edge-cloud concept," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5302–5316, Feb. 2018.

[72] Y. Lu *et al.*, "Blockchain and federated learning for 5G beyond," *IEEE Netw.*, vol. 35, no. 1, pp. 219–225, Feb. 2021.

[73] T. Subramanya and R. Riggio, "Centralized and federated learning for predictive VNF autoscaling in multi-domain 5G networks and beyond," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 1, pp. 63–78, Mar. 2021.

[74] Y. Wei *et al.*, "Federated learning empowered end-edge-cloud cooperation for 5G HetNet security," *IEEE Netw.*, vol. 35, no. 2, pp. 88–94, Apr. 2021.

[75] S. Jere *et al.*, "Federated learning in mobile edge computing: An edge-learning perspective for beyond 5G," *arXiv:2007.08030*, July 2020. [Online]. Available: https://arxiv.org/abs/2007.08030

[76] A. M. Elbir and S. Coleri, "Federated learning for channel estimation in conventional and RIS-assisted massive MIMO," *IEEE Trans. Wireless Commun.*, Nov. 2021.

[77] T. Vu *et al.*, "Energy-efficient massive MIMO for serving multiple federated learning groups," in *Proc. Global Commun. Conf. (GLOBECOM), Madrid, Spain.* IEEE, Dec. 2021, pp. 1–6.

[78] D. Yu *et al.*, "Optimizing over-the-air computation in IRS-aided C-RAN systems," in *21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Atlanta, GA, USA*. IEEE, May 2020, pp. 1–5.

[79] S. Manzoor *et al.*, "Federated learning empowered mobility-aware proactive content offloading framework for fog radio access networks," *Future Gener. Comput. Syst.*, vol. 133, pp. 307–319, Aug. 2022.

[80] Z. Ji and Z. Qin, "Federated learning for distributed energy-efficient resource allocation," *arXiv preprint arXiv:2204.09602*, Apr. 2022. [Online]. Available: http://arxiv.org/abs/2204.09602

[81] A. Hammoud *et al.*, "AI, blockchain, and vehicular edge computing for smart and secure IoV: Challenges and directions," *IEEE Internet Things Mag.*, vol. 3, no. 2, pp. 68–73, June 2020.

[82] S. R. Pokhrel and J. Choi, "Federated learning with blockchain for autonomous vehicles: Analysis and design challenges," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4734–4746, Aug. 2020.

[83] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146–1159, Nov. 2019.

[84] D. Ye *et al.*, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23 920–23 935, Jan. 2020.

[85] Y. M. Saputra *et al.*, "Federated learning meets contract theory: Economic-efficiency framework for electric vehicle networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2803–2817, Dec. 2020.

[86] B. Qolomany *et al.*, "Particle swarm optimized federated learning for industrial IoT and smart city services," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[87] A. Rovira-Sugranes *et al.*, "A review of AI-enabled routing protocols for UAV networks: Trends, challenges, and future outlook," *Ad Hoc Netw.*, vol. 130, p. 102790, May 2022.

[88] Y. Wang *et al.*, "Learning in the air: Secure federated learning for UAV-assisted crowdsensing," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–1, Aug. 2021.

[89] Y. Liu *et al.*, "Federated learning in the sky: Aerial-ground air quality sensing framework with UAV swarms," *IEEE Internet Things J.*, Sept. 2020.

[90] S. Tang *et al.*, "Battery-constrained federated edge learning in UAV-enabled IoT for B5G/6G networks," *arXiv:2101.12472*, Jan. 2021. [Online]. Available: https://arxiv.org/abs/2101.12472

[91] H. Zhang and L. Hanzo, "Federated learning assisted multi-UAV networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 14 104–14 109, Nov. 2020.

[92] W. Ni *et al.*, "Optimal transmission control and learning-based trajectory design for UAV-assisted detection and communication," in *Proc. IEEE 31st Ann. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC), London, UK*, Oct. 2020, pp. 1–6.

[93] H. A. B. Salameh *et al.*, "Jamming-aware simultaneous multi-channel decisions for opportunistic access in delay-critical IoT-based sensor networks," *IEEE Sensors J.*, vol. 22, no. 3, pp. 2889–2898, Dec. 2021.

[94] N. I. Mowla *et al.*, "Federated learning-based cognitive detection of jamming attack in flying Ad-Hoc network," *IEEE Access*, vol. 8, pp. 4338–4350, Dec. 2019.

[95] E. Basar *et al.*, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[96] K. Yang *et al.*, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Netw.*, vol. 34, no. 5, pp. 16–22, Sept. 2020.

[97] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach," *IEEE Trans. Wireless Commun.*, June 2021.

[98] H. Liu, X. Yuan, and Y. A. Zhang, "CSIT-Free federated edge learning via reconfigurable intelligent surface," *arXiv:2102.10749*, Feb. 2021. [Online]. Available: https://arxiv.org/abs/2102.10749

[99] L. Li *et al.*, "Enhanced reconfigurable intelligent surface assisted mmWave communication: A federated learning approach," *Chin. Commun.*, vol. 17, no. 10, pp. 115–128, Oct. 2020.

[100] Y. Shaikh *et al.*, "Survey of smart healthcare systems using internet of things IoT," in *Proc. IEEE Int. Conf. Commun. Comput. Internet Things (IC3IoT), Chennai, India*, Feb. 2018, pp. 508–513.

[101] L. Romeo *et al.*, "Internet of robotic things in smart domains: Applications and challenges," *Sensors*, vol. 20, no. 12, p. 3355, Jan. 2020.

[102] I. Mohammed *et al.*, "Budgeted online selection of candidate IoT clients to participate in federated learning," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5938 – 5952, Apr. 2020.

[103] Y. Liu *et al.*, "Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6348 – 6358, Apr. 2020.

[104] Y. Chen *et al.*, "Fedhealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Apr. 2020.

[105] B. Liu *et al.*, "Experiments of federated learning for COVID-19 chest X-ray images," *arXiv:2007.05592*, July 2020. [Online]. Available: https://arxiv.org/abs/2007.05592

[106] W. Yang *et al.*, "FFD: A federated learning based method for credit card fraud detection," in *Proc. Big Data 8th Int. Congr., Services Conf. Federation, (SCF),*

San Diego, CA, USA, K. Chen, S. Seshadri, and L. Zhang, Eds., vol. 11514, June 2019, pp. 18–32.

[107] J. Bao, T. Shu, and H. Li, "Handover prediction based on geometry method in mmWave communications-a sensing approach," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops), Kansas City, MO, USA*, May 2018, pp. 1–6.

[108] L. Yu *et al.*, "Long-range blockage prediction based on diffraction fringe characteristics for mmWave communications," *IEEE Commun. Lett.*, Apr. 2022.

[109] K. Qi *et al.*, "Dual connectivity-aided proactive handover and resource reservation for mobile users," *IEEE Access*, vol. 9, pp. 36 100–36 113, Feb. 2021.

[110] C. Wang *et al.*, "Deep learning-based intelligent dual connectivity for mobility management in dense network," in *Proc. IEEE 88th Vehic. Technol. Conf. (VTC-Fall), Chicago, IL, USA*, Aug. 2018, pp. 1–5.

[111] A. R. Khan *et al.*, "When federated learning meets vision: An outlook on opportunities and challenges," in *Proc. EAI Int. Conf. Body Area Netw., Glasgow, UK*. Springer, Dec. 2021, pp. 308–319.

[112] X. Cheng, M. Wang, and S. Li, "Compressive sensing-based beamforming for millimeter-wave OFDM systems," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 371–386, Oct. 2016.

[113] R. W. Heath *et al.*, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE j. sel. topics sig. process.*, vol. 10, no. 3, pp. 436–453, Feb. 2016.

[114] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided 6G wireless communications: Blockage prediction and proactive handoff," *IEEE Trans. Vehic. Technol.*, vol. 70, no. 10, pp. 10 193–10 208, Aug. 2021.

[115] M. M. Ahamed and S. Faruque, "5G backhaul: requirements, challenges, and emerging technologies," *Broadband Communications Networks: Recent Advances and Lessons from Practice*, vol. 43, Nov. 2018.

[116] Z. L. Fazliu *et al.*, "Mmwave beam management in urban vehicular networks," *IEEE Syst. J.*, vol. 15, no. 2, pp. 2798–2809, Jan. 2020.

[117] Y. Jian *et al.*, "WiMove: Toward infrastructure mobility in mmWave WiFi," in *Proc. 18th ACM Symp. Mobility Manage. Wireless Access*, Nov. 2020, pp. 11–20.

[118] S. Wan, S. Ding, and C. Chen, "Edge computing enabled video segmentation for real-time traffic monitoring in internet of vehicles," *Pattern Recognit.*, vol. 121, p. 108146, Jan. 2022.

[119] R. Nawaratne *et al.*, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Trans. Ind. Inform.*, vol. 16, no. 1, pp. 393–402, Aug. 2019.

[120] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[121] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sept. 2021.

[122] M. Al-Quraan, A. R. Khan, L. Mohjazi, A. Centeno, A. Zoha, and M. A. Imra, "A hybrid data manipulation approach for energy and latency-efficient vision-aided UDNs," in *IEEE Eighth Int. Conf. Softw. Defined Syst. (SDS), Gandia, Spain*, Dec. 2021, pp. 1–7.

[123] Y. Huang *et al.*, "When deep learning meets edge computing," in *Proc. IEEE 25th international conference on network protocols (ICNP), Toronto, ON, Canada*, Oct. 2017, pp. 1–2.

[124] 3rd Generation Partnership Project (3GPP) Technical Specification Group Radio Access Network NR, "UE conformance specification; Part 1: Protocol," *in 3GPP TS 38.523-1 version 15.2.0 Release 15*, Apr. 2019.

[125] J. Thota and A. Aijaz, "On performance evaluation of random access enhancements for 5G uRLLC," in *Proc. Wireless Commun. Netw. Conf. (WCNC), Marrakesh, Morocco.* IEEE, Apr. 2019, pp. 1–7.

[126] M. Alrabeiah *et al.*, "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proc. IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium*, May 2020, pp. 1–5.

[127] G. Cermak, M. Pinson, and S. Wolf, "The relationship among video quality, screen resolution, and bit rate," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 258–262, Mar. 2011.

[128] I.-H. Mkwawa, E. Jammeh, and L. Sun, "Mapping of received signal strength indicator to QoE in VOIP applications over WLAN," in *Proc. 2012 Fourth Int. Workshop on Qual. Multimedia Experience*, July 2012, pp. 156–157.

[129] Q. C. Li, G. Wu, and T. S. Rappaport, "Channel model for millimeter-wave communications based on geometry statistics," in *Proc. IEEE Globecom Workshops (GC Wkshps), Austin, TX*, 2014, pp. 427–432.

[130] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, Apr. 2018. [Online]. Available: http://arxiv.org/abs/1804.02767

[131] D. Dardari, P. Closas, and P. M. Djurić, "Indoor tracking: Theory, methods, and technologies," *IEEE Trans. Vehic. Technol.*, vol. 64, no. 4, pp. 1263–1278, Feb. 2015.

[132] Rastorgueva-Foi *et al.*, "User positioning in mmW 5G networks using beam-RSRP measurements and Kalman filtering," in *Proc. IEEE 21st Int. Conf. Inf. Fusion (FUSION), Cambridge, UK*, July 2018, pp. 1–7.

[133] M. Koivisto *et al.*, "Joint device positioning and clock synchronization in 5G ultra-dense networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2866–2881, Mar. 2017.

[134] M. Al-Quraan, A. Khan, L. Mohjazi, A. Centeno, A. Zoha, and M. A. Imran, "Intelligent beam blockage prediction for seamless connectivity in vision-aided next-generation wireless networks," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 2, pp. 1937–1948, Oct. 2023.

[135] M. Alrabeiah *et al.*, "ViWi: A deep learning dataset framework for vision-aided wireless communications," in *Proc. IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium*, May 2020, pp. 1–5.

[136] L. Simić *et al.*, "RadMAC: Radar-enabled link obstruction avoidance for agile mmWave beamsteering," in *Proc. 3rd Workshop Hot Topics Wireless*, Oct. 2016, pp. 61–65.

[137] U. Demirhan and A. Alkhateeb, "Radar aided proactive blockage prediction in real-world millimeter wave systems," in *IEEE Int. Conf. Commun. (ICC), Seoul, Republic of Korea*. IEEE, May 2022, pp. 4547–4552.

[138] A. Alkhateeb *et al.*, "DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset," *arXiv preprint arXiv:2211.09769*, Nov. 2022.

[139] A. El-Mowafy *et al.*, "Real-time determination of orthometric heights accurate to the centimeter level using a single GPS receiver: Case study," *J. surv. eng.*, vol. 132, no. 1, pp. 1–6, Feb. 2006.

[140] R. W. Heath *et al.*, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE j. sel. topics sig. process.*, vol. 10, no. 3, pp. 436–453, Feb. 2016.

[141] R. Mundlamuri *et al.*, "Sensing aided channel estimation in wideband millimeter-wave MIMO systems," *arXiv preprint arXiv:2302.02065*, Feb. 2023. [Online]. Available: http://arxiv.org/abs/2302.02065

[142] A. Thornburg, T. Bai, and R. W. Heath, "Performance analysis of outdoor mmWave ad hoc networks," *IEEE Trans. Signal Process.*, vol. 64, no. 15, pp. 4065–4079, Apr. 2016.

[143] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 1100–1114, Oct. 2014.

[144] X. Yu *et al.*, "Coverage analysis for millimeter wave networks: The impact of directional antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1498–1512, Apr. 2017.

[145] M. Alizadeh *et al.*, "Remote monitoring of human vital signs using mm-Wave FMCW radar," *IEEE Access*, vol. 7, pp. 54 958–54 968, Apr. 2019.

[146] E. Grafarend, "The optimal universal transverse mercator projection," in *Geodetic Theory Today: Third Hotine-Marussi Symposium on Mathematical Geodesy L'Aquila, Italy, May 30–June 3, 1994*. Springer, June 1995, pp. 51–51.

[147] H. Jiang *et al.*, "Vehicle classification applying many-to-one input network architecture in 77-GHz FMCW radar," *IET Radar, Sonar & Navigation*, vol. 16, no. 2, pp. 267–277, Feb. 2022.

[148] I. Urazghildiiev *et al.*, "Vehicle classification based on the radar measurement of height profiles," *IEEE Trans. intell. transp. syst.*, vol. 8, no. 2, pp. 245–253, June 2007.

[149] S. Wu, C. Chakrabarti, and A. Alkhateeb, "Proactively predicting dynamic 6G link blockages using LiDAR and In-Band signatures," *IEEE Open J. Commun. Society*, vol. 4, pp. 392–412, Jan. 2023.

[150] Y. Rao *et al.*, "Global filter networks for image classification," *Adv. neural inf. proc. syst.*, vol. 34, pp. 980–993, Dec. 2021.

[151] W. Niu *et al.*, "26ms inference time for resnet-50: Towards real-time execution of all DNNs on smartphone," *arXiv preprint arXiv:1905.00571*, May 2019.

[152] 3GPP, "NR; Radio Resource Control (RRC) protocol specification," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.331, 03 2023, version 17.4.0.

[153] M. Al-Quraan, A. Centeno, A. Zoha, M. A. Imran, and L. Mohjazi, "Federated learning for reliable mmwave systems: Vision-aided dynamic blockages prediction," in *IEEE Wireless Commun. Netw. Conf. (WCNC), Glasgow, UK*, Mar. 2023, pp. 1–6.

[154] K. T. Chui, M. D. Lytras, and A. Visvizi, "Energy sustainability in smart cities: Artificial intelligence, smart monitoring, and optimization of energy consumption," *Energies*, vol. 11, no. 11, p. 2869, Nov. 2018.

[155] M. Waqas *et al.*, "The role of artificial intelligence and machine learning in wireless networks security: principle, practice and challenges," *Artif. Intell. Rev.*, pp. 1–47, Feb. 2022.

[156] Y. Himeur *et al.*, "Building power consumption datasets: Survey, taxonomy and future directions," *Energy and Buildings*, vol. 227, p. 110404, Nov. 2020.

[157] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. neural info. process. syst.*, vol. 30, 2017.

[158] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, Mar. 2020.

[159] M. H. D. M. Ribeiro and L. dos Santos Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Appl. Soft Comput.*, vol. 86, p. 105837, Jan. 2020.

[160] M. Al-Quraan, L. Mohjazi, L. Bariah, A. Centeno, A. Zoha, K. Arshad, K. Assaleh, S. Muhaidat, M. Debbah, and M. A. Imran, "Edge-native intelligence for 6G communications driven by federated learning: A survey of trends and challenges," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 3, pp. 957–979, Mar. 2023.

[161] H. B. McMahan *et al.*, "Federated learning of deep networks using model averaging," *arXiv preprint arXiv:1602.05629*, Feb. 2016. [Online]. Available: http://arxiv.org/abs/1602.05629

[162] E. Sannara *et al.*, "A federated learning aggregation algorithm for pervasive computing: Evaluation and comparison," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom), Kassel, Germany*, Mar. 2021, pp. 1–10.

[163] L. Zhao *et al.*, "Inprivate digging: Enabling tree-based distributed data mining with differential privacy," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2018, pp. 2087–2095.

[164] Y. Liu *et al.*, "Fedvoting: A cross-silo boosting tree construction method for privacy-preserving long-term human mobility prediction," *Sensors*, vol. 21, no. 24, p. 8282, Jan. 2021.

[165] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[166] A. Salam and A. El Hibaoui, "Comparison of machine learning algorithms for the power consumption prediction:case study of Tetouan city," in *Proc. 2018 6th Int. Renew. Sustain. Energy Conf. (IRSEC), Rabat, Morocco.* IEEE, Dec. 2018, pp. 1–5.

[167] T. Ahmad and H. Chen, "Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches," *Energy and Buildings*, vol. 166, pp. 460–476, May 2018.

[168] C. Tung *et al.*, "Large-scale object detection of images from network cameras in variable ambient lighting conditions," in *Conf. Multimedia Inf. Process. Retrieval (MIPR), San Jose, CA, USA.* IEEE, Mar. 2019, pp. 393–398.

[169] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. conf. comput. vision pattern recognit.* IEEE, 2018, pp. 3291–3300.

[170] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Analys. Mach. Intell.*, Aug. 2023.