UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MOBILITY MANAGEMENT IN MULTI-RAT MULTI-BAND HETEROGENEOUS NETWORKS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL & COMPUTER ENGINEERING

BY

SYED MUHAMMAD ASAD ZAIDI
Norman, Oklahoma
2021

MOBILITY MANAGEMENT IN MULTI-RAT MULTI-BAND HETEROGENEOUS NETWORKS

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL & COMPUTER ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Ali Imran, Chair

Dr. James Sluss

Dr. Samuel Cheng

Dr. Timothy Ford

Acknowledgments

I would like to thank my esteemed supervisor – Dr. Ali Imran for his invaluable supervision, support, and tutelage during the course of my Ph.D. degree. His expertise, guidance, support, and patience added considerably to my graduate research experience. The door to Dr. Ali's office was always open whenever I ran into a trouble spot or had a question about my research or writing. He steered me in the right direction whenever he thought I needed it.

I thank the committee for agreeing to be part of my Ph.D. committee, and for reviewing my work. Special thanks to the late Dr. Runolfsson for the support he provided me during my stay here in the University of Oklahoma.

My heartfelt gratitude to my fellow colleagues in AI4Networks Lab for their support and providing a research conducive work environment.

Here I would like to mention Krista Pettersen, Renee Wagenblatt, and Denise Davis who took care of all administrative tasks and were of immense help throughout my four years of study at OU-Tulsa.

Finally, I must express my very profound gratitude to my mother, my wife, and to my daughter for providing me with unfailing support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without them.

Contents

List of Figures

List of Tables

Abstract

Support for user mobility is the *raison d'etre* of mobile cellular networks. However, mounting pressure for more capacity is leading to adaption of multi-band multi-RAT ultra-dense network design, particularly with the increased use of mmWave based small cells. While such design for emerging cellular networks is expected to offer manyfold more capacity, it gives rise to a new set of challenges in user mobility management. Among others, frequent handovers (HO) and thus higher impact of poor mobility management on quality of user experience (QoE) as well as link capacity, lack of an intelligent solution to manage dual connectivity (of user with both 4G and 5G cells) activation/deactivation, and mmWave cell discovery are the most critical challenges. In this dissertation, I propose and evaluate a set of solutions to address the aforementioned challenges.

The beginning outcome of our investigations into the aforementioned problems is the first ever taxonomy of mobility related 3GPP defined network parameters and Key Performance Indicators (KPIs) followed by a tutorial on 3GPP-based 5G mobility management procedures. The first major contribution of the thesis here is a novel framework to characterize the relationship between the 28 critical mobility-related network parameters and 8 most vital KPIs.

A critical hurdle in addressing all mobility related challenges in emerging networks is the complexity of modeling realistic mobility and HO process. Mathematical models are not suitable here as they cannot capture the dynamics as well as the myriad parameters and KPIs involved. Existing simulators also mostly either omit or overly abstract the HO and user mobility, chiefly because the problems caused by poor HO management had relatively less impact on overall performance in legacy networks as they were not multi-RAT multi-band and therefore incurred much smaller number of HOs compared to emerging networks. The second key contribution of this dissertation is development of a first of its kind system level simulator, called SyntheticNET

that can help the research community in overcoming the hurdle of realistic mobility and HO process modeling. SyntheticNET is the very first python-based simulator that fully conforms to 3GPP Release 15 5G standard. Compared to the existing simulators, SyntheticNET includes a modular structure, flexible propagation modeling, adaptive numerology, realistic mobility patterns, and detailed HO evaluation criteria. SyntheticNET's python-based platform allows the effective application of Artificial Intelligence (AI) to various network functionalities.

Another key challenge in emerging multi-RAT technologies is the lack of an intelligent solution to manage dual connectivity with 4G as well 5G cell needed by a user to access 5G infrastructure. The 3rd contribution of this thesis is a solution to address this challenge. I present a QoE-aware E-UTRAN New Radio-Dual Connectivity (EN-DC) activation scheme where AI is leveraged to develop a model that can accurately predict radio link failure (RLF) and voice muting using the low-level measurements collected from a real network. The insights from the AI based RLF and mute prediction models are then leveraged to configure sets of 3GPP parameters to maximize EN-DC activation while keeping the QoE-affecting RLF and mute anomalies to minimum.

The last contribution of this dissertation is a novel solution to address mmWave cell discovery problem. This problem stems from the highly directional nature of mmWave transmission. The proposed mmWave cell discovery scheme builds upon a joint search method where mmWave cells exploit an overlay coverage layer from macro cells sharing the UE location to the mmWave cell. The proposed scheme is made more practical by investigating and developing solutions for the data sparsity issue in model training. Ability to work with sparse data makes the proposed scheme feasible in realistic scenarios where user density is often not high enough to provide coverage reports from each bin of the coverage area. Simulation results show that the proposed scheme, efficiently activates EN-DC to a nearby mmWave 5G cell and thus substantially reduces the mmWave cell discovery failures compared to the state of the art cell discovery methods.

List of Key Symbols

HO	Handover
U	Set of all UEs
u	Any user u $\in U$
U_c	Set of UEs with EN-DC configuration
U_a	Set of EN-DC activated UEs
δ^u_{5R}	5G RSRP of u
θ_{B1}	5G RSRP threshold
δ^u_{4R}	4G RSRP of u
θ_{4R}	4G RSRP threshold
δ^u_{5S}	5G SINR of u
θ_{5S}	5G SINR threshold
δ^u_{4S}	4G SINR of u
θ_{4S}	4G SINR threshold
$\boldsymbol{\Delta}^u$	$[\delta^u_{5R}, \delta^u_{4R}, \delta^u_{5S}, \delta^u_{4S}]$
$\boldsymbol{\Theta}$	$[\theta_{B1}, \theta_{4R}, \theta_{5S}, \theta_{4S}]$
$\boldsymbol{\Delta}^{u,4}$	$[\delta^u_{4R}, \delta^u_{4S}]$
$\boldsymbol{\Delta}^{u,5}$	$[\delta^u_{5R}, \delta^u_{5S}]$
α	EN-DC Activation function
ζ	Potential RLF AI-Model
β	RLF function
η	Potential Muting AI-Model
γ	Muting function
κ	Cell-range
UE	User Equipment

$mmWave$	millimeter Wave
BS	Base Station
LTE	Long Term Evolution
$4G$	Forth Generation Mobile Network
$5G$	Fifth Generation Mobile Network
NR	5G New Radio
LoS	Line of Sight
$NLoS$	Non Line of Sight
$CAPEX$	Capital Expenditure
$OPEX$	Operational Expenditure
RRC	Radio Resource Control
$3GPP$	Third Generation Partnership Project
MME	Mobility Management Entity
SGW	Serving Gateway

CHAPTER 1

Background

The exponential rise in mobile traffic originating from mobile devices highlights the need for making mobility management in future networks even more efficient and seamless than ever before. Ultra-Dense Cellular Network vision consisting of cells of varying sizes with conventional and mmWave bands is being perceived as the panacea for the eminent capacity crunch. However, mobility challenges in an ultra-dense heterogeneous network with a motley of high frequency and mmWave band cells will be unprecedented due to plurality of handover instances, and the resulting signaling overhead and data interruptions for miscellany of devices. Similarly, issues like user tracking and cell discovery for mmWave with narrow beams need to be addressed before the ambitious gains of emerging mobile networks can be realized. Mobility challenges are further highlighted when considering the 5G deliverables of multi-Gbps wireless connectivity, <1ms latency, and support for devices moving at the maximum speed of 500km/h, to name a few.

This dissertation is the first to provide a comprehensive survey on the panorama of mobility challenges in the emerging ultra-dense mobile networks. This dissertation not only presents a detailed tutorial on 5G mobility approaches and highlight key mobility risks of legacy networks, but also review key findings from recent studies and highlight the technical challenges and potential opportunities related to mobility from the perspective of emerging ultra-dense cellular networks.

Mobility management is a complex process with myriad network parameters, and mathematical models become intractable. Similarly, existing network simulators do not incorporate detailed mobility criteria as specified in 3GPP standards. This dissertation overcomes this challenge by explaining the development and some key features of the SyntheticNET simulator.

This dissertation also presents the first framework to quantify and optimize the trade-off between utilization of 5G network and the degradation in QoE due to potential RLF or potential muting, by leveraging real network data measurements. Finally, I present a joint search-based mmWave cell discovery approach that can help networks configure mmWave camping to mobile users while keeping into account the LoS conditions and maximum allowable distance.

3GPP Mobility criteria is a complicated process and is, therefore, necessary to explain the 3GPP mobility management process in 5G networks. The following section explicates both the intra-frequency and inter-frequency handover and reselection criteria in idle and connected mode users respectively.

1.1 Understanding Mobility in Cellular Networks

Mobility in cellular networks plays a pivotal role ensuring an optimal experience to the subscribers. It guarantees that mobile users won't just be able to maintain connectivity but attain the best available connection to the network as they move towards the destination. Seamless and timely HO and cell reselection has always been a major challenge in any wireless communication systems including 5G. Mobility has been categorized as Idle and Connected Mode Mobility in 5G. Note that the mobility procedure in LTE (4G) is very similar in 5G New Radio (NR) using events A1, A2, A3, A4, A5 and A6 to trigger HOs. Event A2 and A1 are triggered when RF condition of the UE falls below and exceeds the configured threshold respectively and are used to start and stop inter-frequency neighbor search. Intra-frequency HO is initiated by event A3 where the neighbor RF condition becomes higher than serving RF condition by a configured threshold. Event A4 and A5 are typically used for inter-frequency HO where target inter-frequency cell has to be higher than an absolute threshold for the event A4 to be triggered. On the contrary, event A5 in addition to event A4 condition, requires serving cell RF condition to be below a certain threshold. Finally, event A6 is similar to event

A3 but is used for intra-frequency HO of the secondary frequency the UE is camped onto. Event A4 and A5 can also be used for conditional HO management for e.g. for load balancing. In addition to the events described above, event B1 and B2 (A4 and A5 alike) are also used for inter-technology HO, and for dual-connectivity, but they are not discussed here to keep the focus of this chapter confined to basic mobility procedures and the associated challenges. The only difference between 5G and 4G mobility criteria is in the idle mode where respective idle mode reselection parameters in 5G NR are present in different SIB# than in LTE. Moreover, the idle mode parameter names and functionalities in 5G are similar as in 4G. Comprehensive explanation of 5G mobility procedure while keeping in view the 5G network architecture and interfaces is presented in the following subsections.

1.1.1 Idle Mode Mobility

UE is in idle mode when it is neither running any active communication service nor is connected to any cell. UE in idle mode is constantly trying to search and maintain services such as Public Land Mobile Network selection, cell selection and reselection, location registration, and reception of system information. By maintaining an idle mode connection, UE can readily establish a Radio Resource Connection (RRC) for signaling or data transfer as well as be able to receive any possible incoming connections. UE always powers ON in idle mode and selects the cell with the maximum signal strength through a process known as cell selection. However, this initially selected cell will not always be the best to serve especially when UE moves from one place to another. Therefore, to maintain the quality of signal, UE has to camp on another optimal cell, a process known as cell reselection.

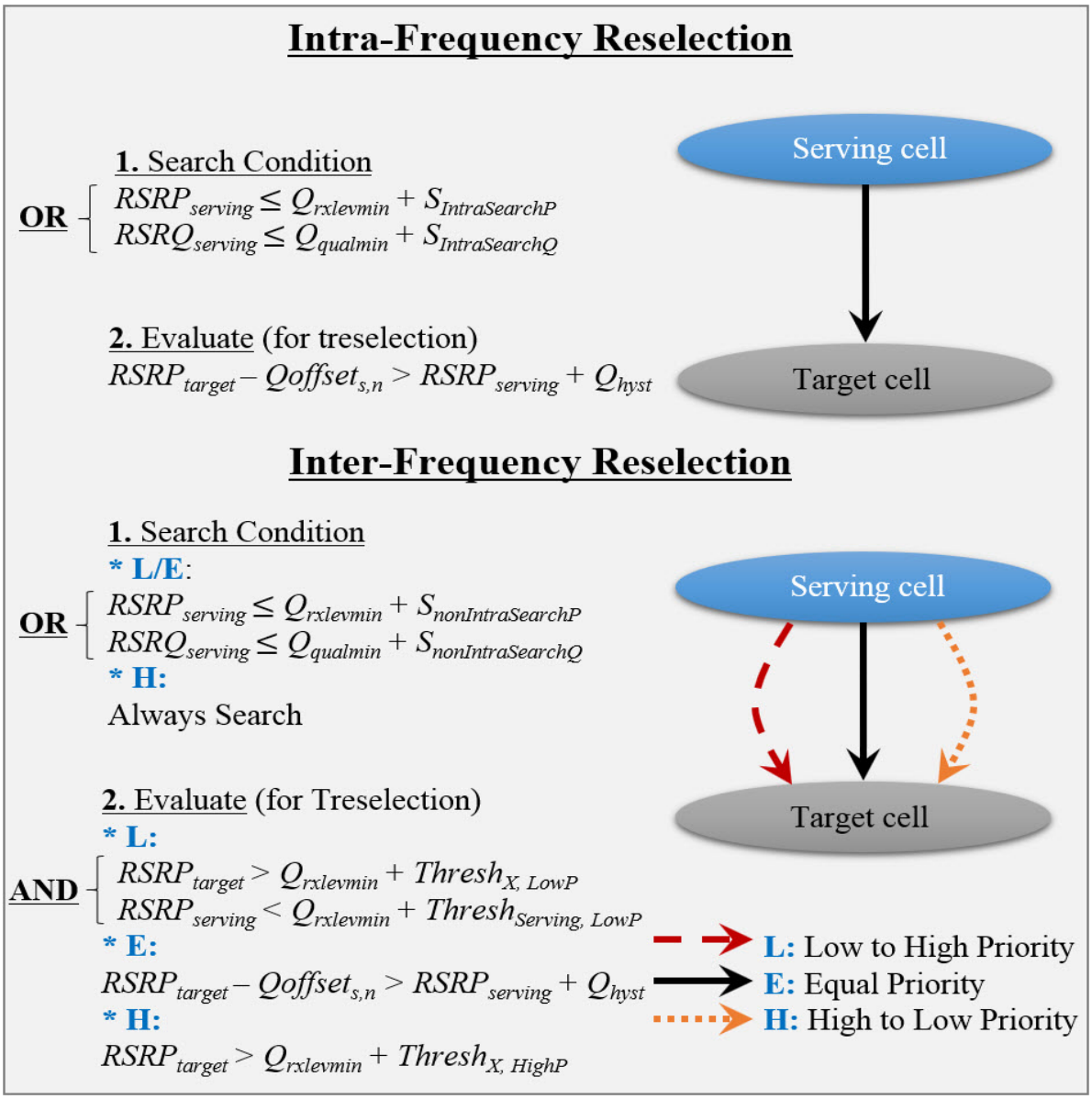

Figure 1.1: 3GPP [1] cell reselection criteria based on SIB2 and SIB4 parameter for intra-frequency and inter-frequency reselection respectively.

Table 1.1: 3GPP [1] Intra/Inter-Frequency Reselection Parameters

Parameter	SIB#	Description
$Q_{rxlevmin}$	SIB1	Minimum RSRP threshold required to camp in idle mode
$Q_{rxlevmin}$	SIB2	RSRP$_{serving}$ threshold required to compute intra-frequency reselection conditions
$Qoffset_{s,n}$	SIB2	Positive or negative bias required to promote or avoid intra-frequency cell reselection to target cell. * Idle Mode Cell Individual Offset
Q_{hyst}	SIB2	RSRP$_{target}$ – RSRP$_{s}erving$ required to satisfy intra-frequency reselection condition.
treselection	SIB2	Time needed to satisfy intra-frequency reselection condition before actual reselection to the optimal cell
$S_{IntraSearchP/Q}$	SIB2	RSRP/RSRQ threshold below which user searches for intra-frequency target cell.
$Q_{rxlevmin}$	SIB4	RSRP$_{s}erving$ threshold required to compute inter-frequency reselection conditions
$Q_{qualmin}$	SIB4	RSRQ$_{s}erving$ threshold required to compute inter-frequency reselection condition
$Qoffset_{s,n}$	SIB4	Positive or negative bias required to promote or avoid inter-frequency cell reselection to equal priority target cell. * Idle Mode Cell Individual Offset
Q_{hyst}	SIB4	RSRP$_{t}arget$ – RSRP$_{s}erving$ required to satisfy reselection condition to equal priority cell
treselection	SIB4	Time needed to satisfy inter-frequency reselection condition before actual reselection to the optimal cell
$S_{NonIntraSearchP/Q}$	SIB4	RSRP$_{s}erving$ / RSRQ$_{s}erving$ threshold below which user searches for inter-frequency target cell
$Thresh_{X,LowP}$	SIB4	RSRP$_{t}arget$ threshold required to trigger inter-frequency reselection to lower priority target cell
$Thresh_{ServingLowP}$	SIB4	RSRP$_{s}erving$ threshold required to trigger inter-frequency reselection to lower priority target cell
$Thresh_{X,HighP}$	SIB4	RSRP$_{t}arget$ threshold required to trigger inter-frequency reselection to higher priority target cell

Cell Reselection Criteria

In 5G, BS broadcasts nine System Information Block (SIB) messages for the UE as defined in 3GPP [1]. Out of those messages, SIB 1, 2, 3 and 4 contain critical parameters to execute idle mode cell reselection to the optimal 5G cell. SIB1 has the serving cell parameters as well as the cell selection parameters, while SIB2 has the common parameters used for intra-frequency and inter-frequency reselection. SIB3 is dedicated to intra-frequency reselection parameters, however, operators can broadcast the related parameters in SIB2 instead, and thus SIB3 is not broadcasted. SIB4 contains inter-frequency reselection through target frequency priority and the associated parameters. Fig. 1.2 illustrates a pictorial demonstration of the reselection conditions and evaluation in 5G as described by 3GPP. Description of the related reselection parameter, and the respective location (SIB#) can be found in Table 1.1. LTE uses the same reselection procedure with the only difference that the contents of SIB2, SIB3 and SIB4 in 5G are found in SIB3, SIB4 and SIB5 of LTE instead.

User Tracking

The idle mode mobility of the UE is the responsibility of Access and Mobility Function (AMF) at the Tracking Area (TA) level for RRC idle mode users and at the RAN Notification Area (RNA) for RRC inactive mode users. Here I only talk about the idle mode users as the mobility procedure in 5G is similar for RRC idle mode and RRC inactive mode users. Note that unlike the connected mode, network is unaware of cell-level UE location in idle mode. After powering ON, UE acquires the Tracking Area List (TAL) composed of a list of TA codes through the periodic SIB1 broadcast from the cell. As UE traverses through the network while performing cell reselection procedure, it compares the TA code of the new cell with its own TAL. If the TA code of a newly visited cell does not match with its own TAL, it initiates TA Update (TAU) process to request AMF for location update as seen in the Fig. 1.2(a). TAU helps to track the UE

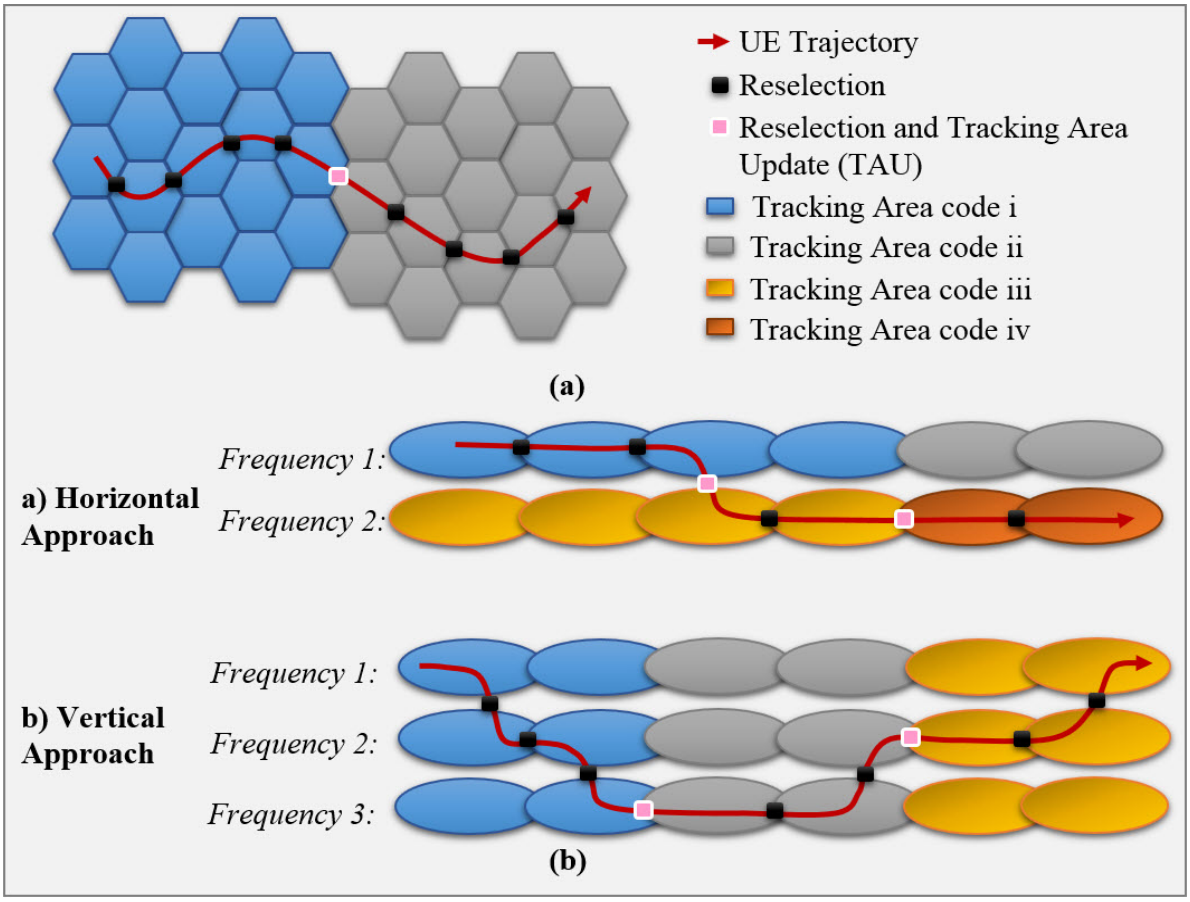

Figure 1.2: (a) Tracking Area Update (TAU) procedure in LTE networks, (b) Common Tracking Area (TA) planning approaches.

in case of any incoming call. Bigger TA size reduces signaling overhead of TAU at the expense of larger paging domain, ultimately resulting in higher paging-based downlink signaling load at network level.

Common Idle Mode Mobility Risks

In this subsection, I discuss about the common idle mode mobility risks in the existing LTE network. But since the mobility process is similar in 5G networks, 5G capable UEs are expected to face similar challenges. In idle mode, data transmission does not take place, therefore reliability and QoS are not the issues of concern. However, reselection procedure can incur accessibility and user tracking issues in rare occasions. During the network attach procedure, idle mode UE first sends connection request and awaits connection setup message from the BS. If UE does not receive any message from the BS within a predefined time (t300 timer known to UE via SIB2 'SIB1 in 5G [2]'), it restarts the accessibility procedure. Under special circumstances, if UE sends a connection request to the serving cell followed by reselection to a neighboring cell, it cannot receive the connection grant simultaneously. The new serving cell in this case does not become aware that the UE which just moved under its coverage needs to access the network. Thus, UE has to wait for a time defined in t300 before re-initiating the access procedure in the new serving cell. During this time, UE experiences latency and can have serious impact on the applications requiring ultra-low latency. The delay can be suppressed by having smaller t300 timer, but at the cost of increased signaling load due to the increase in redundant connection requests and replies. Moreover, smaller t300 also negatively impact UE energy consumption (due to recurrent Random-Access Channel 'RACH' attempts). Repeated RACH attempts might result in higher Central Processing Unit (CPU) load of serving cell, especially at busy hour. Similar accessibility delay at TA border can result in paging failure, since the network can be unaware of the accurate UE location unless TAU followed by a successful accessibility is performed. TA planning is

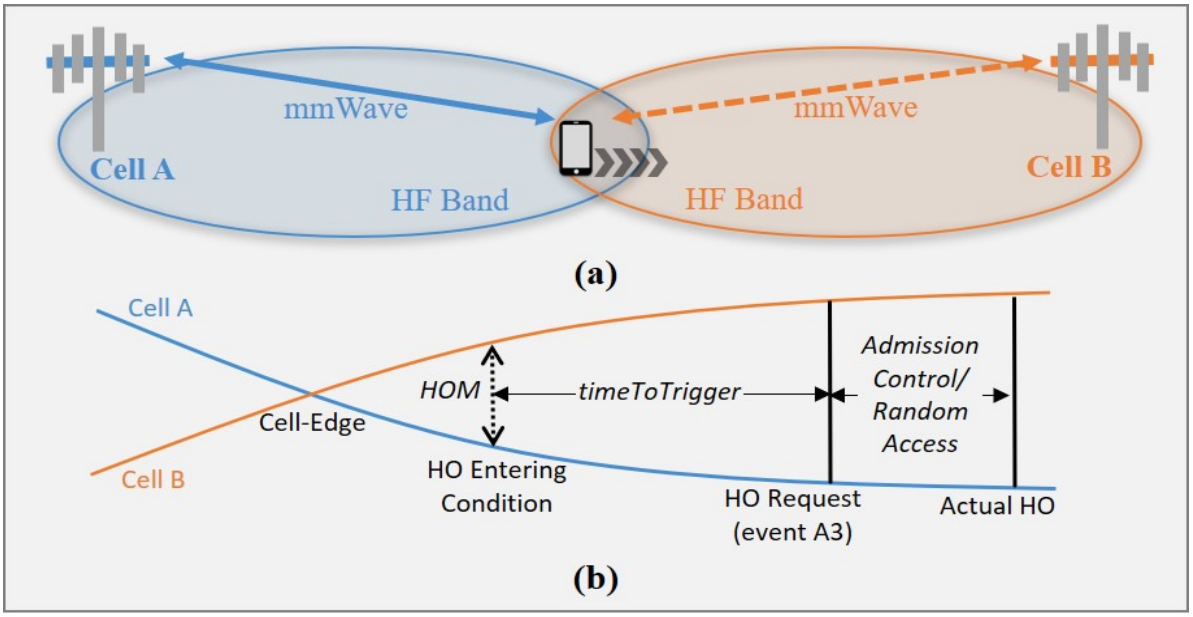

Figure 1.3: General HO procedure. (a) UE performs HO from cell A to cell B at cell-edge as it moves closer to the cell B. Scenario 1 and 2 represents HF coverage and mmWave narrow beams, (b) 3GPP [18] based intra-frequency HO process.

a crucial task and two approaches are used in existing networks: a) horizontal approach, b) vertical approach, as shown in Fig. 1.2(b). TAU procedure initiates for every inter-frequency reselection in horizontal approach, thus it is deployed where radio condition is good, and user is least expected to make recurrent inter-frequency reselection. On the contrary, poor radio condition area should have vertical approach to minimize TAU for inter-frequency reselection instances. Horizontal approach is favorable for high speed traffic like train lines or highways. One approach to address this issue in the existing cellular network is the use of adaptive TA codes, where users are configured with a list of TA codes to prevent ping-pong TAUs. However, determining the optimal number of TA codes in a list and the cumulative TA size still remain an open research problem.

1.1.2 Connected Mode Mobility

UE is said to be in connected mode when it has established a connection with its peer Radio Resource Control (RRC) layer at the serving BS and the network can transmit and/or receive data to/from the UE. As there is an exchange of data between the UE and the BS, uninterrupted data transfer needs to take place for a seamless continuity of

Table 1.2: 3GPP [2] HO Parameters Conveyed to UE in RRC Reconfiguration Layer 3 Message

Parameter	Descriptions
s-Measure	RSRP threshold below which user searches for optimal intra-frequency target cel
Ofn	Frequency offset for target cell
Ofp	Frequency offset for serving cell
Ocn	Target cell offset * Commonly known as Cell Individual Offset 'CIO'
Ocp	Serving cell offset
$Hys*$	Hysteresis to prevent ping-pong HOs
$A3-Off*$	RSRP$_target$ – RSRP$_serving$ offset required to satisfy A3 condition
$A2-Thr*$	Event A2 RSRP$_serving$ threshold
$A1-Thr*$	Event A1 RSRP$_serving$ threshold
$A4-Thr*$	Event A4 RSRP$_serving$ threshold
$A5-Thr1*$	Event A5 RSRP$_serving$ threshold
$A5-Thr2*$	Event A5 RSRP$_target$ threshold
timeToTrigger (TTT)	Time for which Event (A1-A5) condition need to be satisfied

service when a UE moves from one BS to another BS. This ideally seamless mobility in connected mode is termed as handover (HO).

UE Side Mobility Trigger

UE triggers an intra-frequency HO request to the next optimal cell by sending A3-Measurement Report (MR) to its serving cell as shown in Fig. 1.3. The serving cell then decides whether to entertain the request and perform the HO, by communicating with the target cell and serving AMF. An intra-frequency HO is the first preference in cellular networks; however, there are instances in which an inter-frequency HO is the preferred choice. For example: a) when there is a coverage hole in the serving frequency, b) when the current serving cell does not support the requested service e.g. Voice over NR, and c) when load balancing is needed to avoid congestion in the serving frequency. In Fig. 1.4, I illustrate the 3GPP [2] defined inter-frequency HO criteria. For a description of each HO parameter, refer to Table 1.2.

Network Side Mobility Trigger

HOs are undoubtedly more complicated than cell reselection. Aside from the source and target cell, core entities which include Access and Mobility Function (AMF), Session

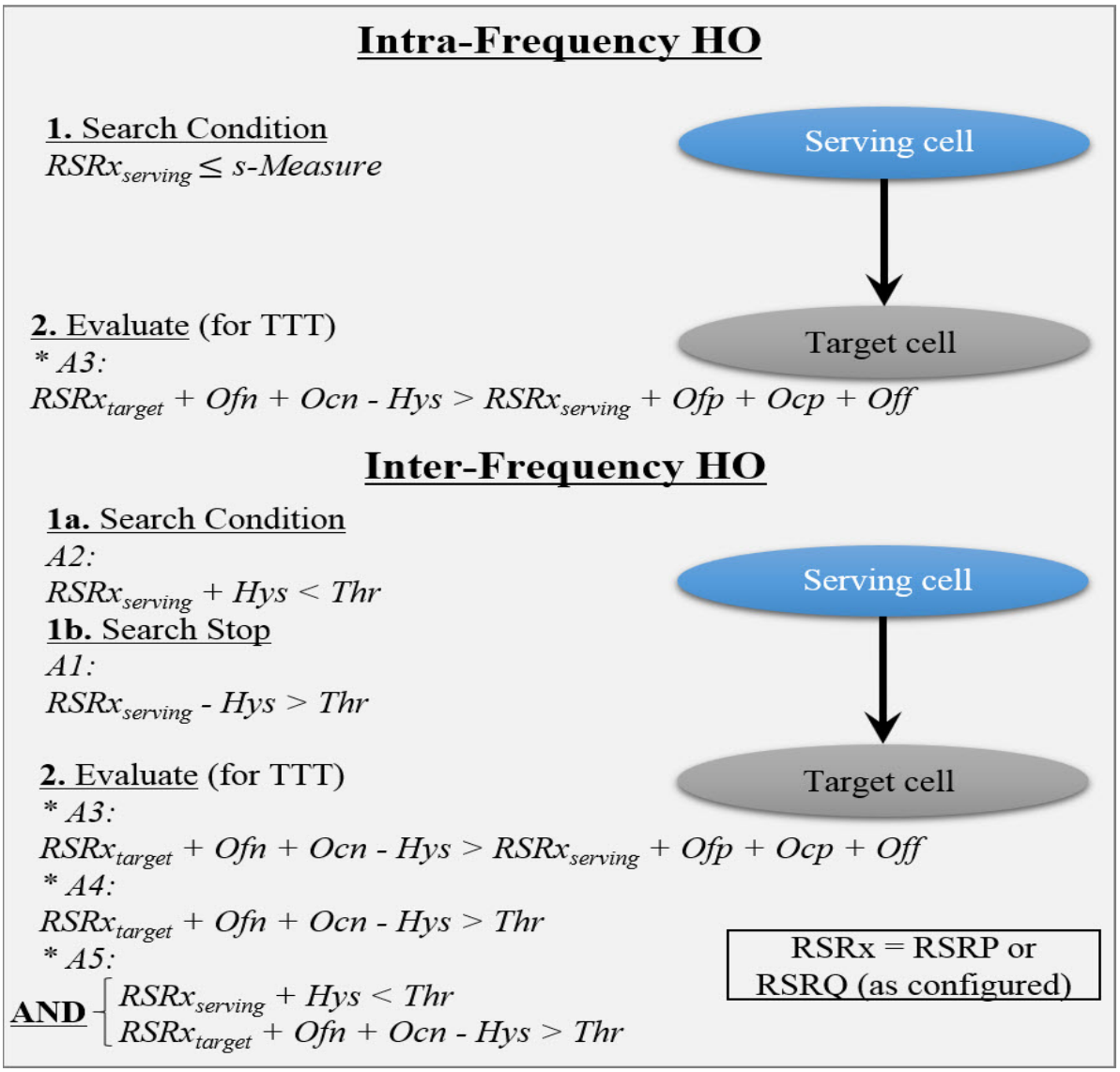

Figure 1.4: 3GPP [2] intra-frequency and inter-frequency handover criteria in LTE networks.

Management Function (SMF) and User Plane Function (UPF) need to be updated as well. Depending on the scenario, data transfer and handling could pose several challenges. In normal cases, when AMF, SMF and UPF do not change during the HO, signaling is reasonable and it is termed Xn based HO. Here, the Xn interface is used for the preparation phase of the HO. However, when the Xn interface does not exist between the participating cells, an N2 based HO is performed where cells use a longer path for communication. Signaling flow for the Xn based HO is illustrated in Fig. 1.5. 3GPP [2] named Xn as the interface used to connect 5G BSs directly, and N2 interface is the logical interface between two 5G BSs connected through the core network (AMF).

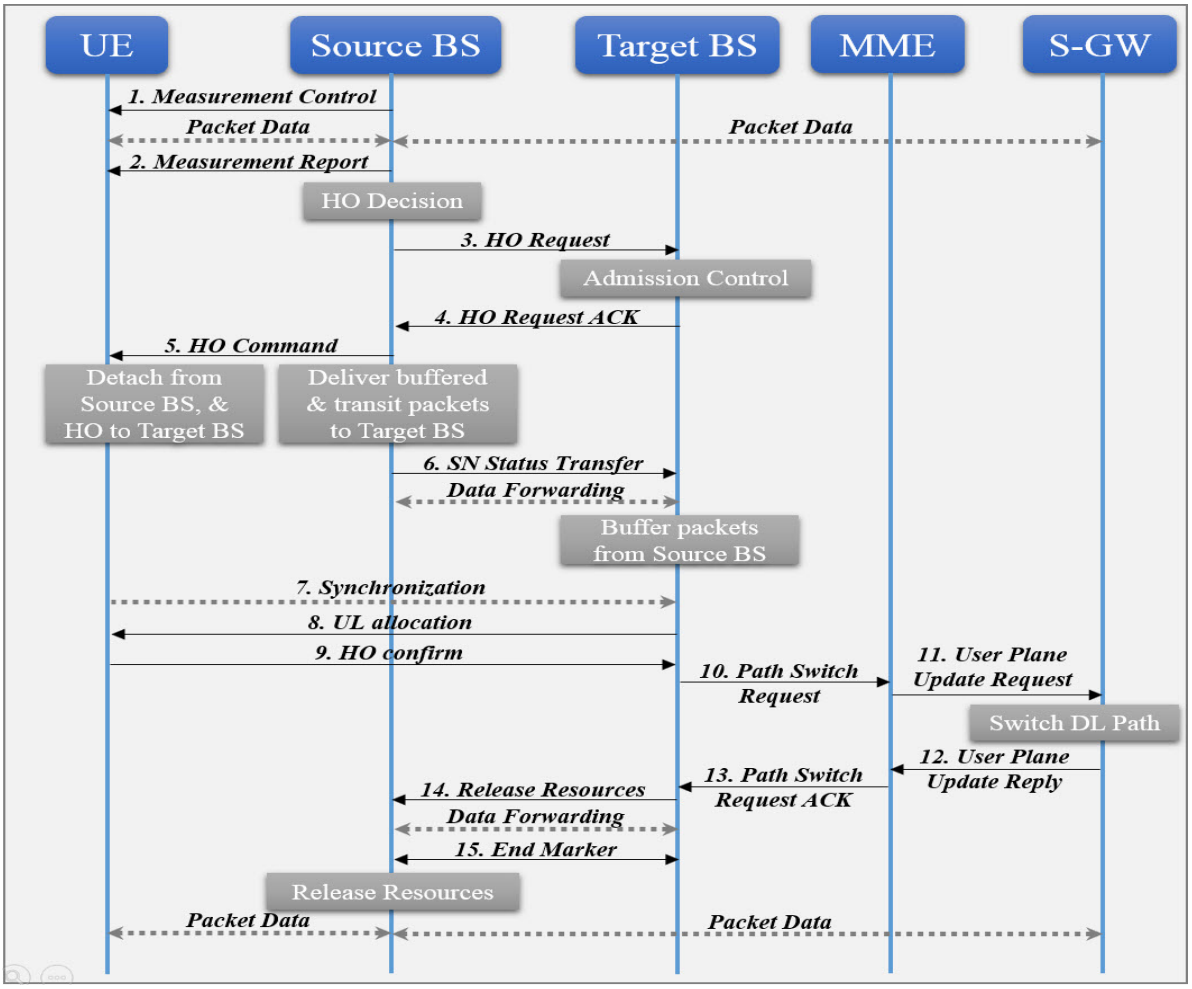

Figure 1.5: Xn based handover without UPF re-allocation in 5G networks.

N2 interface is used if the direct Xn interface between the neighboring BSs do not exists.

Common connected Mode Mobility Risks

Apart from the fast fading effect due to Doppler shift in physical layer, the mobile UE has to cope with several Layer 3 issues as well, which can be eluded primarily by a timely HO and an optimal selection of the target BS. Some of the issues mobile UE experiences during inter-site mobility are presented in Fig. 1.6, with possible solution(s) in Table 1.3.

Table 1.3: Common HO Issues and Their Solutions

HO Issue	Parameter Optimization Solution	Possible Cons
Late Intra HO	i. Lower A3 offset, shorter TTT ii. Positive CIO towards target cell	Prone to unwanted HO's to non-target cells Potential Ping-Pong between source and target especially for static users.
Late Inter HO	Higher A2, Accelerate A3/A4/A5, shorter TTT	Prone to unwanted HO's to non-target cells/layers.
Wrong Intra HO	i. Higher A3 offset, longer TTT ii. Negative CIO towards wrong-target cell	May cause delayed HO to target cell Stationary users might experience poor signal quality.
Wrong Inter HO	Lower A2, Delay A3/A4/A5, shorter TTT	May cause delayed HO to target cell.
Early Intra HO	i. Higher A3, longer TTT ii. Negative CIO towards target cell	May cause HO delay to target cell
Early Inter HO	Lower A2, Delay A3/A4/A5, shorter TTT	

1.2 Motivation for Artificial Intelligence (AI) Enabled Mobility Management

Network operators optimize their network by tuning a set of mobility related parameters, and then by observing the HO attempt, HO success and few other QoE KPIs affected by those modified network parameters. This approach will soon be impractical due to the large number of parameters per cell. Moreover, the ultra-densification of heterogeneous networks having not only multiple frequencies per RAT (Random Access Technology), but also different RATs (2G, 3G, 4G, 5G) operating in parallel to each other, will make the traditional approach of hit-and-trial totally useless and unmanageable. The only answer to the complex optimization requirements of emerging networks can be given by Artificial intelligence (AI) based approaches, some of which have been described in this dissertation.

This section explains the complex interplay between mobility related network parameters and Key Performance Indicators (KPIs) deemed essential to maintain reliable and high-speed network services to the UEs. The complex interplay between parameters and KPIs will further clarify why the traditional approach of hit-and-trial based optimization methods will not suffice the ambitious QoE requirements expected from emerging networks. Few of the vital mobility related KPIs are outlined below:

- User tracking KPI indicates the paging hit rate when users served under the TA are notified by an incoming call. The idle mode mobile user must update its location (via TAU) to the core network when it moves into the neighboring TA. By doing

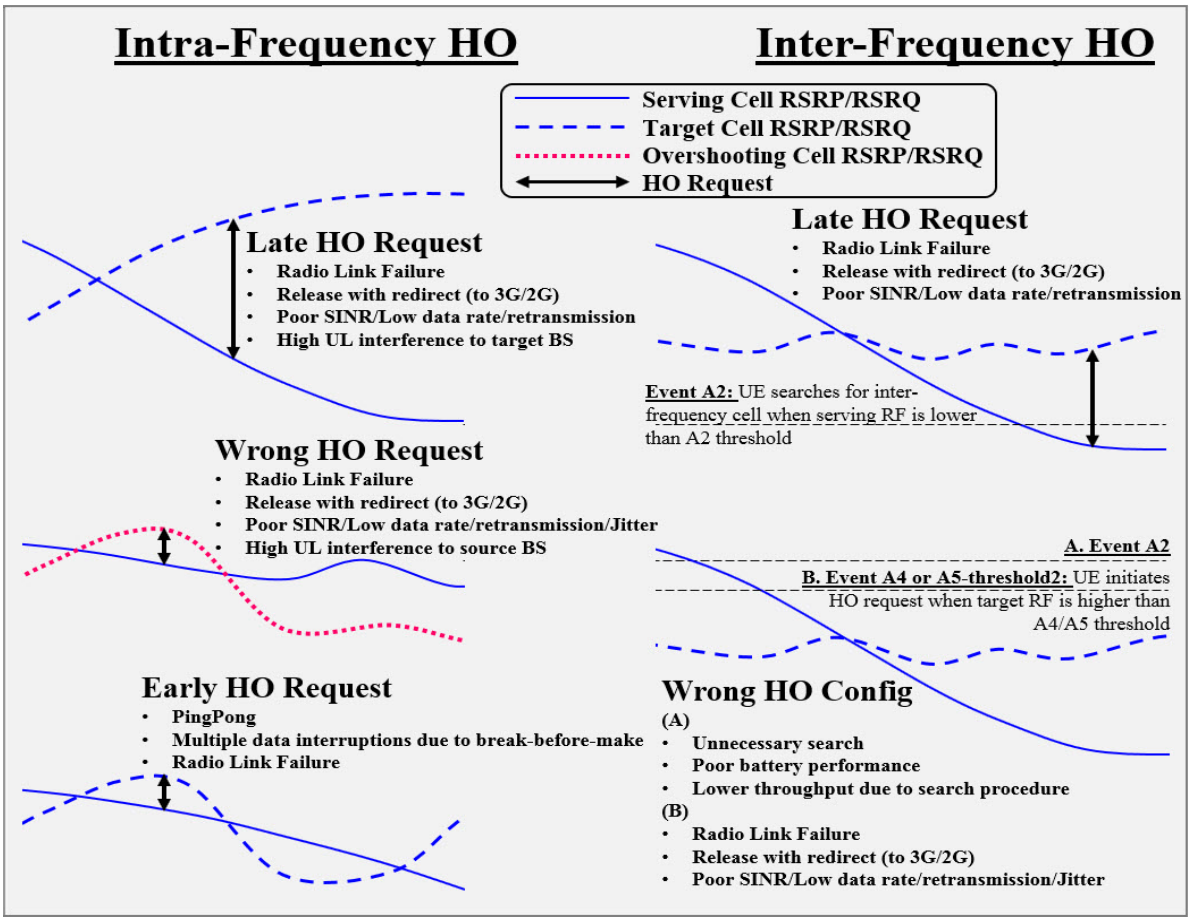

Figure 1.6: Common Mobility Related Risks in 4G/5G networks.

so, the respective TA is broadcasted with paging attempt messages in case of any incoming call. A delay in TAU can result in paging failure and reattempts.

- Mobility oriented HO process or TAU trigger results in the control plane messages being sent in the air interface and in the core network. The percentage of network resources used by control plane are measured by signaling data KPI.

- User terminal energy consumption e.g. during data delivery and location update, can be measured by the UE battery KPI.

- Reliability (or retainability) KPI indicates the percentage of users that dropped the connection with their participating cells during the HO procedure. Majority of the HO failure instances are observed due to late HO attempts.

- Ping-pong HO KPI point out the early HO occasions in a cell. UE undergoing

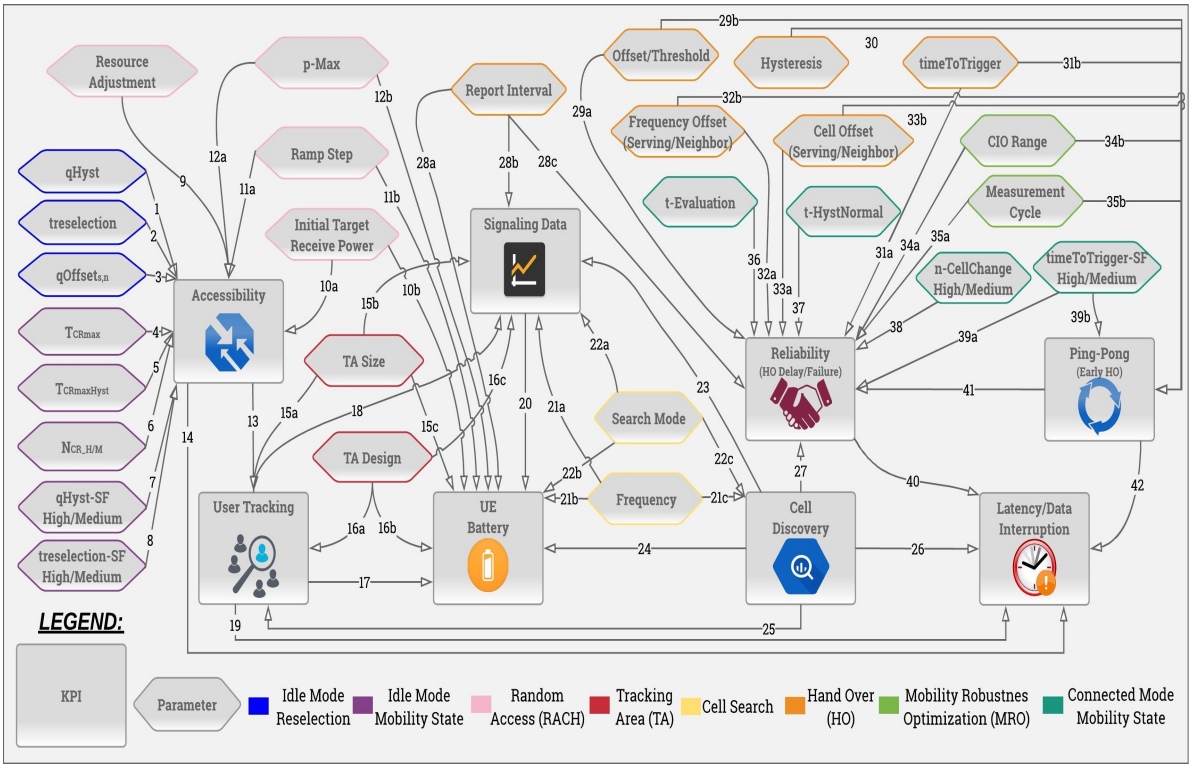

Figure 1.7: Relationship diagram for mobility related KPIs and their interplay with the associated network parameters (grouped in different colors)
Source: [1, 2].

ping-pong HOs leads to back-and-forth HOs between the participating cells and can lead to higher signaling load and sometimes even low retainability KPI.

- Cell discovery KPI measure the small cell camping rate each time a UE is configured with a cell search process. Timely cell discovery can result in more offloading opportunities, and hence, efficient utilization of the available resources.

- Latency or data interruption KPI represents the delay UE observe during HO execution, paging attempt to success duration, accessibility etc.

- Accessibility KPI for a given time interval represents the percentage of idle mode UEs that were able to successfully acquire network access. Accessibility KPI indirectly impacts latency and user tracking KPI under rare circumstances for mobile users.

In most cases the KPI-parameter dependency is multi-pronged and leads to complex

and often conflicting interplay between the KPIs and parameters. This interplay in the mobility KPI and the associated key parameters is summarized in Fig 1.7. The key challenges that arise from the convolved association between the mobility KPI and parameter [1, 2] are briefly described below:

1: Smaller qHyst value accelerates reselection, as soon as the target cell RSRP becomes greater than serving cell RSRP. As a result, accessibility issues related to idle mode mobility (as discussed earlier in the section) can be addressed. However, too low of a qHyst can result in unnecessary reselection (for instance, to an over-shooting cell).

2: Shorter Treselection will improve the accessibility KPI at the cell boundary due to timely reselection. However, too short Treselection will result in ping-pong reselection especially for stationary users (i.e. due to shadowing).

3: Idle mode Cell Individual Offset (CIO) to accelerate or decelerate reselection towards a neighboring cell. (configuring a positive CIO towards a particular neighbor can accelerate reselection, and vice versa).

4: Time window to evaluate mobility State [1] of a UE. Number of reselections made within this time window will dictate mobility state (normal, medium, or high) of a UE. Reselection criteria is typically eased as mobility state changes from normal to medium or high.

5: Specify additional time period before UE can enter back to its normal mobility state with default reselection parameters. Recurrent mobility state change can be avoided by tuning this parameter.

6: Number of cell change needed (ignoring similar cells) within 'parameter #4' before UE changes mobility state from normal to medium or high respectively.

7: Scaling factor by which the default qHyst (parameter #1) is decreased when the mobility state is changed to medium or high.

8: Scaling factor by which the default treselection (parameter #2) is decreased when

the mobility state is changed to medium or high.

9: Amount and location of RACH resources to ensure RACH success (providing adequate RACH resources, and avoiding RACH resource conflict between neighboring cells).

10: Higher target power can increase chances of RACH success at first attempt (better accessibility KPI) at the cost of a) higher battery consumption and b) chances of increased uplink interference for neighboring cells. An optimal target receive power is vital for better network operations.

11: Increase in the transmission power every time a RACH attempt fails. Higher step size can increase RACH success but with more battery consumption and vice versa.

12: Maximum allowable UE RACH power - Increasing maximum allowable UE transmission improves RACH success probability but with high energy consumption.

13: Improved accessibility to achieve a faster TAU can ensure accurate user tracking and prevent paging failure instances for mobile users.

14: Reduce latency through faster accessibility for mobile users (e.g. fast reselection to best signal cell and appropriate power for RACH success).

15: Smaller TA size will improve UE location estimate and will decrease the core network signaling due to smaller paging area. However, frequent TAU by mobile users will add radio access side signaling.

16: Suitable TA design (horizontal/vertical assignment) based on coverage conditions and type of traffic (e.g. high speed UEs) to ensure accurate user tracking and minimize TAU and hence, conserve UE battery and network signaling load.

17: Reducing TAU attempts for mobile users to conserve UE battery.

18: Reducing TAU attempts for mobile users to lessen signaling load.

19: Fast and efficient user tracking to reduce latency in accessing the network.

20: Minimizing signaling helps avoid unnecessary transmission and the UE battery can be conserved.

21: Higher cell search frequency will be beneficial to offload users to other cells. However, more battery will be consumed while searching. In addition, signaling load will increase every time a UE is configured with cell search procedure.

22: Periodic search mode will reduce signaling data generation as search configuration will be transferred to UE just once. However, small periodicity will waste the UE battery, and a large periodicity might miss a suitable offloading opportunity. On the contrary, a smart aperiodic search mode (e.g. location triggered) will be efficient and will save battery but signaling will be generated with each search configuration.

23: Signaling data generated for cell discovery purposes should be minimized.

24: UE consumes battery during cell search, hence, cell discovery should be minimized with high hit rate.

25: Timely cell discovery (intra-frequency) will prevent out-of-service (unreachable UE) occasions and Radio Link Failure (RLF) can be prevented.

26: Timely cell discovery (intra-frequency) will prevent recurrent re-transmissions and ultimately lead to Radio Link Failure at the cell edge.

27: Timely cell discovery (intra-frequency) will ensure HO success especially for mmWave cells and the UE will not observe Radio Link Failure.

28: Smaller report interval (HO requests) will have more signaling data and battery utilization. However, the reliability KPI will improve as there will be more chances of BS being able to successfully receive and decode the HO request.

29: HO offset/threshold can be tuned to achieve timely HO.

30: Suitable hysteresis parameter will minimize chances of ping-pong HOs.

31: Small timeToTrigger can result in ping-pong HOs (e.g. for non-mobile users), while

long timeToTrigger can avoid the HO resulting in low reliability/retainability KPI (e.g. to overshooting cells). Similarly, high speed users should be configured with lower timeToTrigger to accelerate HO to cell with best RSRP.

32: Frequency based CIO to accelerate or decelerate inter-frequency HOs to all neighboring cell(s). Optimal CIO can prevent late and/or early HO.

33: Relation based CIO to accelerate or decelerate intra/inter-frequency HOs toward the configured neighboring cell(s). Optimal CIO can prevent late and/or early HO.

34: Configuring a large CIO range can avoid the chances MRO assigns a large CIO (a large CIO is not recommended as it can have negative consequences especially for static users)

35: Shorter MRO cycle can recommend suitable CIO configuration based on changing traffic conditions. However, too short of a cycle should be prevented as it can have sub-optimal recommendations due to inadequate statistical data required to configure optimal CIO.

36: Similar to 'parameter #4' but for connected mode.

37: Similar to 'parameter #5' but for connected mode.

38: Similar to 'parameter #6' but for connected mode.

39: Similar to 'parameter #8' but for connected mode.

40: HO failure results in higher latency and more data interruption occasions.

41: Frequent HOs increases the risk of HO failure both for static and mobile users.

42: Latency and data interruption are intrinsic to break-before-make HOs, hence ping-pong HOs should be avoided.

Fig. 1.7 illustrates the simplest representation of the complex interaction between various KPIs and mobility related network parameters. It can act as a foundation, with the help of which, AI researchers can devise an ideal mobility management scheme that

aims to minimize the negative impact on KPIs indirectly affected by tuning mobility related network parameters.

1.3 Research Objectives

In light of the above discussion in section 1.2, the research presented in this dissertation provides answers to the following questions.

1. Can we characterize the relationship between mobility related network parameters and the vital network Key Performance indicators (KPIs)?

2. Are existing simulators capable to implement and validate mobility related research proposals?

3. Dual 4G and 5G connectivity enables UE to access the key 5G features, however, UE needs to maintain strong connection with both 4G and 5G network. How can UE make reliable and effective dual connectivity - E-UTRAN New-Radio Dual Connectivity (EN-DC)?

4. mmWave cells require beamforming to deliver good signal strength but the pencil like beams incur a challenging mmWave cell discovery procedure. Is there a way to achieve an efficient mmWave cell discovery that minimizes cell discovery failures, and the delay in cell search procedure?

This dissertation addresses the aforementioned research questions. Real mobile network data is collected, and synthetic data is generated from a 3GPP compliant SyntheticNET simulator. 3GPP-compliant rigorous simulation studies are carried out to find and validate the answers to the above questions. The key contributions of the dissertation are outlined in the following section.

1.4 Contributions

The contributions of this dissertation can be summarized as follows:

- The heterogeneous multi-band multi-RAT ultra-dense network deployment to increase the area spectral efficiency may hamper the ambitious QoE goals if the optimal mobility management approaches are not employed. To date, the identification of mobility related network parameters and KPIs remain implicit in literature. The dissertation not only presents a detailed taxonomy of the key mobility related 3GPP defined network parameters and KPIs, but also establishes a framework to characterize the relationship between the vital 28 mobility parameters and 8 related KPIs. The first major contribution of the thesis here is a novel framework to characterize the relationship between the 28 critical mobility-related network parameters and 8 most vital KPIs. The dissertation also lays down the first comprehensive tutorial on 3GPP-based 5G mobility management procedures for both a) idle/inactive mode, and b) connected mode mobile users. This tutorial acts as a base to correlate all mobility management related network parameters with all mobility management related KPIs.

- Mathematical models to incorporate the realistic mobility and HO process becomes intractable due to the myriad mobility related network parameters and KPIs involved. Moreover, existing network simulators do not support comprehensive mobility criteria conditions due to the complexity, and the resource hungry requirements required to integrate user mobility. To overcome the hurdle of realistic mobility and HO process modeling in multi-band multi-RAT ultra-dense networks, this dissertation discusses the development of SyntheticNET - the very first python-based simulator that fully conforms to 3GPP Release 15 5G standard and is upgradable to future releases. The key distinguishing features of SyntheticNET compared to existing simulators include: 1) a modular structure to facilitate

cross validation and upgrading to future releases; 2) flexible propagation modeling using empirical model based, measurement based, ray tracing based, or AI-based propagation modeling; 3) ability to import data sheet based on realistic vendor specific base station features such as antenna and energy consumption pattern; 4) support for 5G standard adaptive numerology; 5) realistic and user-specific mobility patterns that are yielded from actual geographical maps; 6) detailed handover (HO) process implementation; and 7) incorporation of database-aided edge computing. Another key feature of the SyntheticNET is the ease with which it can be used to test AI-based network automation solutions. Being the first python-based 5G simulator, this facilitates the SyntheticNET's built-in capability to process and analyze large data sets and integrated access to Machine Learning libraries.

- A key challenge in emerging multi-RAT technologies is the lack of an intelligent solution to manage dual connectivity with 4G as well as 5G cell needed by a user to access 5G services. This dissertation presents a framework to quantify and optimize the trade-off between 5G network utilization and QoE degradation due to potential radio link failures (RLF) or potential muting during EN-DC activation leveraging real network data measurements. The framework leverages a two-stage AI model capable of accurately detecting potential RLF and muting instances to tune the parameters used to activate EN-DC.

- Futuristic mobile networks face an unprecedented challenge of mmWave cell discovery accentuated by the highly directional nature of mmWave transmission crucial to compensate the severe propagation losses. This dissertation presents a novel mmWave cell discovery approach in which AI is leveraged to build an optimal mmWave coverage map build using realistic mmWave network data of RLFs, coverage holes, and serving mmWave cell identifiers. This dissertation also demonstrates a case study in which existing network operators can facilitate EN-DC activation using the proposed joint search based mmWave cell discovery approach.

Results when compared to state-of-the-art cell discovery approaches, quantify the gains in terms of mmWave cell discovery failure avoidance and the increase in number of EN-DC activations due to successful mmWave cell discovery to optimal cell.

1.5 Dissemination and Publications

Throughout the course of preparation for this dissertation, several dissemination activities were carried out. These activities have resulted in the following presentations and (accepted or pending) peer reviewed articles.

1.5.1 Journals

1. M. Manalastas, H. Farooq, S. M. Asad Zaidi, A. Abu-Dayya, and A. Imran, "AI-Based Handover Failure Prediction Model for Handover Success Rate Improvement in 5G," IEEE Global Communications Conference (GLOBECOM), 2021 (under review).

1.6 Organization

The dissertation is structured as follows. Chapter 2 presents a detailed literature review and the state-of-the-art work done in mobility management in emerging cellular networks. Chapter 3 presents a system level simulator in Python platform, named SyntheticNET. This chapter discuss the reasons behind the development of SyntheticNET simulator, and its attributes that makes it feasible platform to implement and validate the mobility management proposals. Chapter 4 discusses the QoE aware 4G and 5G dual connectivity activation criteria after taking into account the radio link failures and voice muting anomalies. A mmWave discovery approach keeping in view the mmWave

signal blockage condition is presented in chapter 5. Finally, chapter 6 discusses the conclusions and future work, and it thus concludes the dissertation. In this chapter we also outline some possible directions for future work that can be built on the work presented in this dissertation.

CHAPTER 2

Literature Survey

2.1 Introduction

The unprecedented rise in the Internet traffic volume seen in recent years is attributed to high speed internet, and the advent of smart phone technology. It is anticipated that the number of 5G subscriptions will be 2.8 billion by the year 2025 [4]. Furthermore, the insatiable demand for new bandwidth-hungry applications will lead to an avalanche of traffic volume growth. Mobile data traffic will increase from 10.7 exabytes/month in 2016 to 83.6 exabytes/month by 2021 [5], and that number will further increase exponentially in the years to follow. The emerging cellular networks including 5G mobile network standard as the next revolution of mobile cellular technology needs to support the ever-increasing mobile users, provide adequate data rate for the bandwidth hungry applications, address the QoS issues of delay tolerant applications and realize the concept of Internet-of-Things (IoT) [6, 7]. 5G promises to deliver "more" of everything [8]: a) top speeds of up to 1 Gbps, b) 100 Mbps data rate per end user even at the cell edge, c) RTT (Round-Trip-Time) latencies in the millisecond range, d) higher connection densities (1 million connections per km2 [9]), and e) support for mobile devices at the speed of up to 500 km/h. Currently, Signal to Interference and Noise Ratio (SINR) is considered as the primary metric for planning, dimensioning and optimization of the existing cellular networks [6]. However, for a few exceptions like fixed IoT services, an additional network planning/design criterion in the future may be the mobility related QoE. This is likely the outlook in the backdrop of the following observations:

1. Coverage and SINR provisioning will become a relatively easy challenge given the anticipated higher Base Station (BS) density in emerging cellular networks,

along with the sophisticated interference management schemes and massive MIMO assisted beamforming.

2. However, the very same advances in the network design i.e. densification, beamforming, massive MIMO make the mobility management a more challenging problem. The challenges stem not only from the increased number of handovers (HOs) but also, beam management to maintain the expected QoE. Challenges related to beam management includes focusing narrow beams on the mobile users, cell discovery in narrow beam cells, and large signaling overheads when the user moves from one massive MIMO cell to another cell.

3. With the advent of mmWave, narrow beams of mmWave bands will have limited overlap with each other, making HO a challenging problem (see Fig.4 for observing the difference in HO scenarios in low frequencies and mmWave frequencies).

The growing demand for mobile services in public transport, highways, open-air gatherings etc. [10] will be critical to customer experience. Providing a satisfactory Quality of Experience (QoE) to a relatively large number of mobile users and a miscellany of the devices including phones, tablets, sensors etc. at the speed up to 500km/h imposes extreme challenges to the future mobile networks. Mobility requirements in emerging cellular networks require high efficiency of the HO mechanism, which makes the cell-change seamless for the users. Unlike the legacy technologies (i.e. 3G and 4G) that do not give primary importance to high mobility, future mobile networks will treat mobility as an integral part of the communication standard. Moreover, the mobility management schemes in Long Term Evolution (LTE) systems (also known as 4G system) and to a certain extent, even in the latest 5G New Radio (NR) standard are not well adapted to the typical deployment of the futuristic mobile networks due to multiple factors, few of which are highlighted below:

- The legacy LTE architecture makes use of a centralized network control entity

Table 2.1: Comparison of LTE latency with 5G expected goals

Parameter	LTE Requirement	5G Goal
Control Plane Latency (Accessibility)	100ms	10ms
User Plane Latency	20ms	1ms
HO Execution	49.5ms	0ms

called MME (Mobility Management Entity) located in the core network. The emerging cellular networks are expected to have 10-folds higher density [11], with a larger fraction of mobile users. Thus, without a mobility centric redesign of the architecture, future networks should have 10 times more MME's just to achieve a similar QoS as in LTE.

- To achieve the logistic feasibility for high density deployment, BS placement in future mobile networks are likely to be impromptu or much less planned [11]. This will increase mobility related signaling load that is bound to complicate the core network management and planning.

- HO decision in existing networks is made by participating BSs without considering the deployment of the BSs and backhaul limitations. In futuristic mobile networks with flexible BS deployment, the chances of User Equipment (UE) in selecting the optimal target BS may become smaller.

- While the capacity crunch will be addressed by small cells (SC), a large number of inter-SC HOs will take place leading to frequent session interruptions during HO.

- With smaller inter-site-distance as in SCs, the performance of the existing mobile network reduces sharply owing to the risk of HO failures due to high radio link variability as shown in [12].

- In existing mobile networks, UE context has to move from one BS to another for every HO. This will impose unprecedented signaling overhead in the future ultra-dense network architecture. While signaling is already growing 50% faster than data traffic [13], network efficiency will drop by many folds using the current HO

approaches.

- HOs in 4G networks are based on the broadcast signal called Reference Signal (RS). The mmWaves with narrow beams cannot have RS broadcast to the whole coverage area within the cell range. Hence, cell discovery, especially for mobile UEs is another key mobility challenge in emerging cellular networks not faced by the traditional mobile networks.

- With SON stepping up the automatization of network configuration and optimization in LTE, myriad of mobility management parameters associated with the large number of closely deployed 5G BSs need to be well managed. For that, the existing SON solutions will not be sufficient.

- 5G applications with Ultra Reliable Low Latency Communications (URLLC) e.g. self-driven cars demand very low latency requirements as shown in Table 2.1 [14].

- When UE perform HO to a better cell, it experiences a latency and data interruption period. HO management in the future mobile networks should ensure a seamless and latency-free transition from the source to the target cell.

- With mobile phone traffic on the rise, and with the advent of self-driven cars and drones needing robust connectivity, seamless and reliable mobility management has become more significant than ever. The adaptation of ultra-dense cellular networks and mmWave BSs makes the mobility management even more complex challenge requiring significant research effort.

In light of the above discussion, I can conclude that mobility management will have much stronger impact on the design and architecture of upcoming cellular networks, than it had on the legacy networks. The futuristic networks will incorporate high mobility requirements as an integral part, and appreciable efforts are required to attain ubiquitous top-notch QoE. Majority of mobility oriented surveys in the literature target adhoc

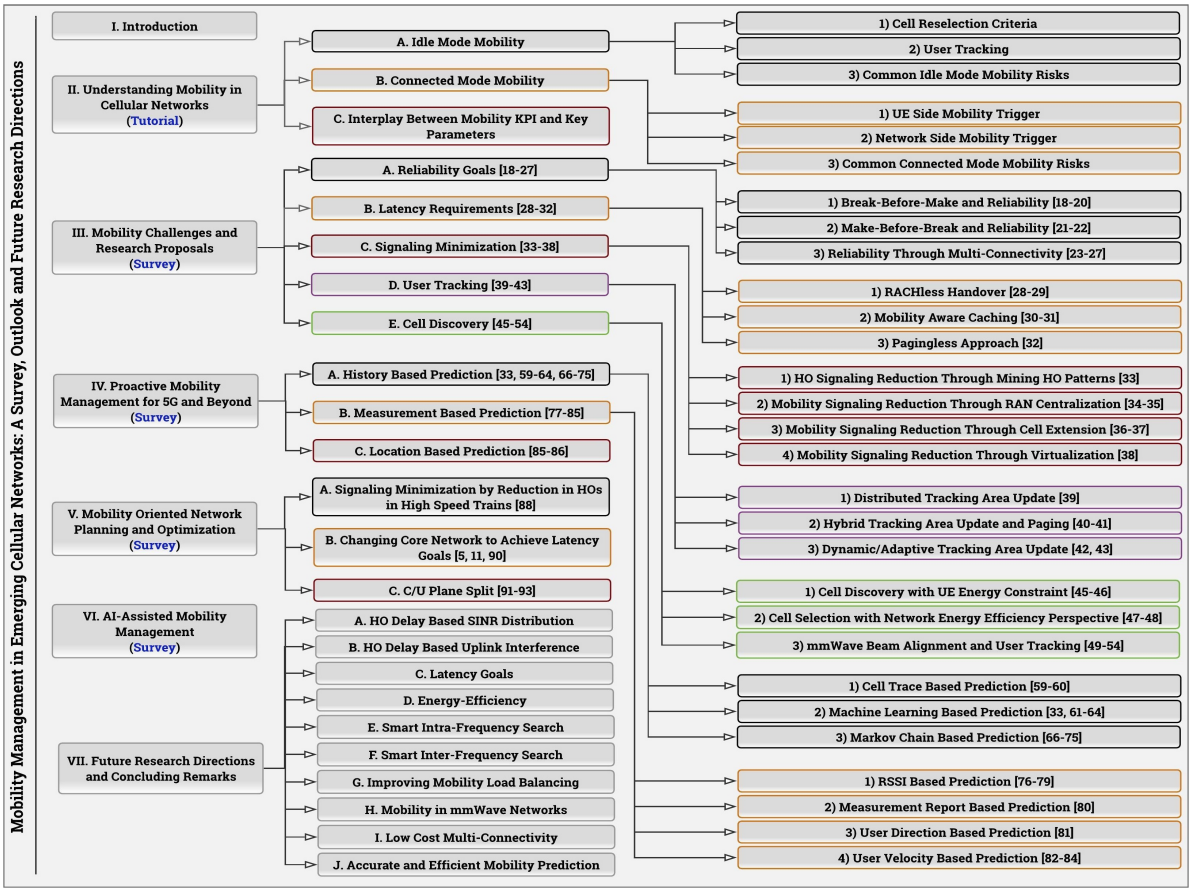

Figure 2.1: Layout of the contents and outline of this chapter on mobility survey.

networks [15, 16, 17]. Mobility surveys on cellular networks do exists e.g. Xenakis et al. [18] presented survey on HO decision algorithms for the femtocells in LTE-Advance. Another survey on high mobility wireless communication has recently been presented in [19], however, the attributes and intricacies of the 5G architecture have not been addressed. To the best of the authors' knowledge, this survey is the first to address the novel contributions by research community targeting mobility in emerging ultra-dense mobile networks. The contributions in this chapter and its organization are as follows:

- This chapter gives the first comprehensive tutorial on 3GPP based 5G mobility management procedures for both a) idle/inactive mode, and b) connected mode mobile users.
- Mobility related surveys do exist in the literature (e.g. [15, 16, 17] on adhoc networks), but none of the aforementioned surveys addresses the futuristic mobile

networks. This chapter presents a single go-to manuscript where future researchers not only understand the 3GPP mobility procedure and the existing mobility related literature but also assist them in finding the research directions they might undertake.

- It presents a first of its kind framework to correlate all mobility management related parameters with all mobility management related KPIs. To facilitate easy understanding, this framework is presented in the form of a flow chart shown in Fig. 1.7.

- It presents a comprehensive and taxonomized review of the literature on mobility management.

- It identifies the need for a new paradigm for mobility management deemed essential to meet the quality of experience (QoE) requirements of the emerging applications and use-cases.

- It proposes a novel proactive mobility management framework to meet the requirements of the emerging mobile networks. Since the challenges of 5G networks (e.g. low latency, less overhead and high quality of experience) cannot be addressed by the current reactive mobility management techniques, I discussed the proactive mobility management in section 2.4.

- It highlights the need to come up with Mobility oriented Network planning and dimensioning

- It provides a collection of the latest AI-based techniques to smartly address mobility related challenges.

- It identifies the future research direction and few open research problems to achieve this paradigm shift.

Table 2.2: List of Acronyms

Acronyms	Descriptions	Acronyms	Descriptions
3GPP	Third Generation Partnership Project	4G	Fourth Generation
5G NR	Fifth Generation New Radio	AMF	Access & Mobility Function
BS	Base Station	CDR	Call Detail Record
CIO	Cell Individual Offset	CoMP	Co-Ordinated Multi Point
CQI	Channel Quality Indicator	CSI	Channel State Identifier
gNB	5G Base Station (Next Generation NodeB)	HF	High Frequency
HO	Hand Over	HOM	Hand Over Margin
IMMCI	Idle Mode Mobility Control Information	ICIC	Inter Cell Interference Coordination
IoT	Internet of Things	KPI	Key Performance Indicator
LB	Load Balancing	LoS	Line of Sight
LTE	Long Term Evolution (4G)	MLB	Mobility Load Balancing
MME	Mobility Management Entity	MR	Measurement Report
MRO	Mobility Robustness Optimization	MIMO	Multiple Input Multiple Output
MDT	Minimization of Drive Test	NLoS	Non-Line of Sight
PCI	Physical Cell Identifier	P-GW	PDN Gateway
QoE	Quality of Experience	RAT	Random Access Technology
RRC	Radio Resource Control	RTT	Round Trip Time
RS	Reference Signal	RSRP	Reference Signal Receive Power
RSRQ	Reference Signal Receive Quality	RSSI	Receive Signal Strength Indicator
RwR	Release with Redirect	RLF	Radio Link Failure
SC	Small Cell	SINR	Signal to Interference plus Noise Ratio
S-GW	Serving Gateway	SON	Self-Organizing Networks
SDN	Software Defined Networking	SIB	System Information Base
TA	Tracking Area	TAL	Tracking Area List
TAU	Tracking Area Update	UPF	User Plane Function
UE	User Equipment	URLLC	Ultra-Reliable Low Latency Communication

Fig. 2.1 outlines the structure of the chapter. It also provides a taxonomy of the literature on mobility.

2.2 Mobility Challenges and Research Proposals

Seamless mobility experience at a very high-speed is considered as one of the major use cases for 5G networks, particularly in wake of advent of autonomous cars, low altitude drones, and emerging high-speed commute systems. The mobility characteristics of the emerging networks, such as densification and adaptation of mmWave narrow beam cells (discussed in section 2.1), combined with the intrinsic complexity of the mobility management process (discussed in section 1.1) means that the mobility management in

5G and beyond requires significant research efforts by wider community. In this section, I review the recent contributions made by the research community to address 5G and beyond mobility challenges, by categorizing them in six sections as shown earlier in Fig. 2.1. Studies focused on reliability goals that involve achieving seamless and timely HO while preventing HO failures and ping-pong HOs are discussed in the first sub-section. Studies focused on achieving mobility while maintaining small delay are discussed in the Latency Requirements sub-section. Signaling Minimization approaches are presented in the next sub-section, followed by User Tracking in futuristic ultra-dense networks. Subsequent sub-section covers studies on cell discovery including the goal to perform timely offloading from macro-cells to small cells in order to prevent network congestion and efficiently utilize network resources. Finally, research work focused on lessening energy consumption are presented in the last sub-section.

2.2.1 Reliability Goals

Mobility casts a serious threat to reliability especially when HO is being performed from one cell to another. Now I will discuss different research work on different HO types and the respective reliability goals.

Break-Before-Make and Reliability

5G NR employs break-before-make (hard) HO approach [2] where UE breaks the connection with the serving BS before resuming the new connection with the target BS, and this process makes the mobile UE prone to undesirable service interruption. Repetition of this type of HO under ping-pong scenario makes it even more susceptible to call drops. An effort to deal with the frequent HO case has been presented in [20]. This chapter focuses on the multi-objective learning-based mobility management strategy where a learning model is described to obtain a comprehensive network information. Then a multi-objective mobility management method is proposed taking into consideration

user QoE and number of HOs. Results are compared with 3GPP based HO scheme, and the authors show that number of HOs are reduced by more than 5 times. As a future step, simulations can be presented by using a stochastic network model. Much of the reliability concerns are studied while keeping in view the UE downlink performance only. Authors in [21] studied reliability for uplink channel of multi-user MIMO channel. Authors employed Quadrature Spatial Modulation (QSM) to lower the uplink Bit Error Rate (BER) from 10-1 (when using spatial multiplex) to the order of 10-3. As a future work, BER results can be shown with different user velocity to evaluate the efficacy of the proposed approach for a realistic scenario of mobile users.

Make-Before-Break and Reliability

Unlike 5G NR and LTE, 3G uses an alternative of break-before-make HO, i.e. make-before-break vis-a-vis soft HO. 3G UE apply macro diversity where it can establish simultaneous connection to more than one cell, and the set of participating cells are referred to as Active Set (AS). Authors in [22] propose a 3G like soft HO approach where multiple serving cells are represented by AS. The results show that fixed AS window can prevent RLF to a great extent. However, throughput degradation is observed as radio resources of the weaker cells are unnecessarily wasted by the user. To counter this problem, the authors propose a dynamic AS window where add/remove parameters are adapted based on the slope of the linear curve that creates the dependency between the add/remove offset and the size of AS. AS based approach will result in more signaling, computation and energy requirements in maintaining and updating the connectivity to different cells in the AS. One drawback of make-before-break HO scheme is the complexity at UE side to process multiple RF chains. Note that the advent of narrow mmWave beams in 5G that is likely to lower the source link reliability for the mobile users, further undermines the perceived advantages of make-before-break HO. Authors in [23] analyzed the pros and cons of make-before-break HO in more detail and concluded

that they are unsuitable for 5G networks. For similar reasons, 3GPP RAN WG2 during its meeting #94 decided to discard make-before-break like procedures from the scope. For the above-mentioned reasons and to achieve higher reliability and retainability goals, the 5G networks have employed hard HO process requiring successful break-before-make procedures. Reliability goals in literature are usually addressed through multi-connectivity approaches.

Reliability Through Multi-Connectivity

Multi-Connectivity (MC) can be employed in conjunction with break-before-make HO approach to mitigate interference through coordination. MC can attain ultra-reliability, low latency, and interruption-free communication by preparing the target cell before the transmission is broken. Furthermore, it tackles connection failures by using a coordinated transmission among the serving cells. As a result, HO failures and RLFs are drastically suppressed. However, drawback of MC includes added complexity in adding/removing MC participant cells. A study by Tesema et al. [24] on intra-frequency MC shows that the RLFs can be avoided while enhancing throughput through joint transmission of BSs. The authors in [24] then extended their idea in [25] to inter-frequency MC and prove availability benefits in that scenario. However, stationary users were considered with focus on modeling of the best server association. Their study did not incorporate reliability for mobile users. In a separate study [26], the same group of authors deal with mobility concerns and evaluated reliability performance through different intra/inter frequency cells. For intra frequency, Dynamic Single Frequency Network (DSFN) is proposed to dynamically add BSs to the coordination set. This in turn helps to achieve reliability and low latency of less than 1ms. For inter-frequency on the other hand, redundant transmissions are performed on the different frequency layers, such that the UE selects the best transmission, i.e., selection combining is applied. The proposed approach can avoid poor SINR of <-6dB (marked as RLF) and achieve higher

reliability of 99.999% or greater. Tesema et al. further enhanced their work in [27] by proposing a novel multi-connectivity scheme that uses fast selection of serving cell from a set of prepared cells similar to Co-ordinated Multi-Point Transmission (CoMP). Fig. 2.2 shows different types of CoMP. Control plane in CoMP is served by a primary cell only, and if radio condition of the respective control channel degrades, then user plane data may not be guaranteed even if radio condition of user plane cell is better. On the contrary, Fast Cell Select (FCS) is proposed in which the selected cell from the set of pre-arranged cells is used for transmission of both data and control signals. The presented work provides gain in the quality of the control and data signals, which ultimately solves RLF problem and improve throughput of cell-edge user. CoMP, although beneficial, has an intrinsic conflict with the hard-HO methods used in 5G networks, as connection with source cell terminates before setting up a connection to the target cell. In [28], authors addressed this conflict by introducing a new HO mechanism based on CoMP joint transmission scheme in order to minimize inter-cell-interference (ICI) level between the adjacent cells during the HO execution. Their algorithm consists of Coordination set (CS) and Transmission set (TS) of BSs. CS selection is assisted by the UE through sending periodic measurement report which contains UE velocity and RF condition. Velocity metric is used to avoid small cells for high velocity UEs, and RF condition is used to determine TS. Performance evaluation results show that ICI is reduced considerably leading to a better average throughput per user during the HO procedure. Benefits are achieved at the cost of higher complexity and increase in signaling data. A study on optimal TS size to improve reliability, and throughput, taking into consideration the processing complexity and the magnitude of the control data would be a good research contribution.

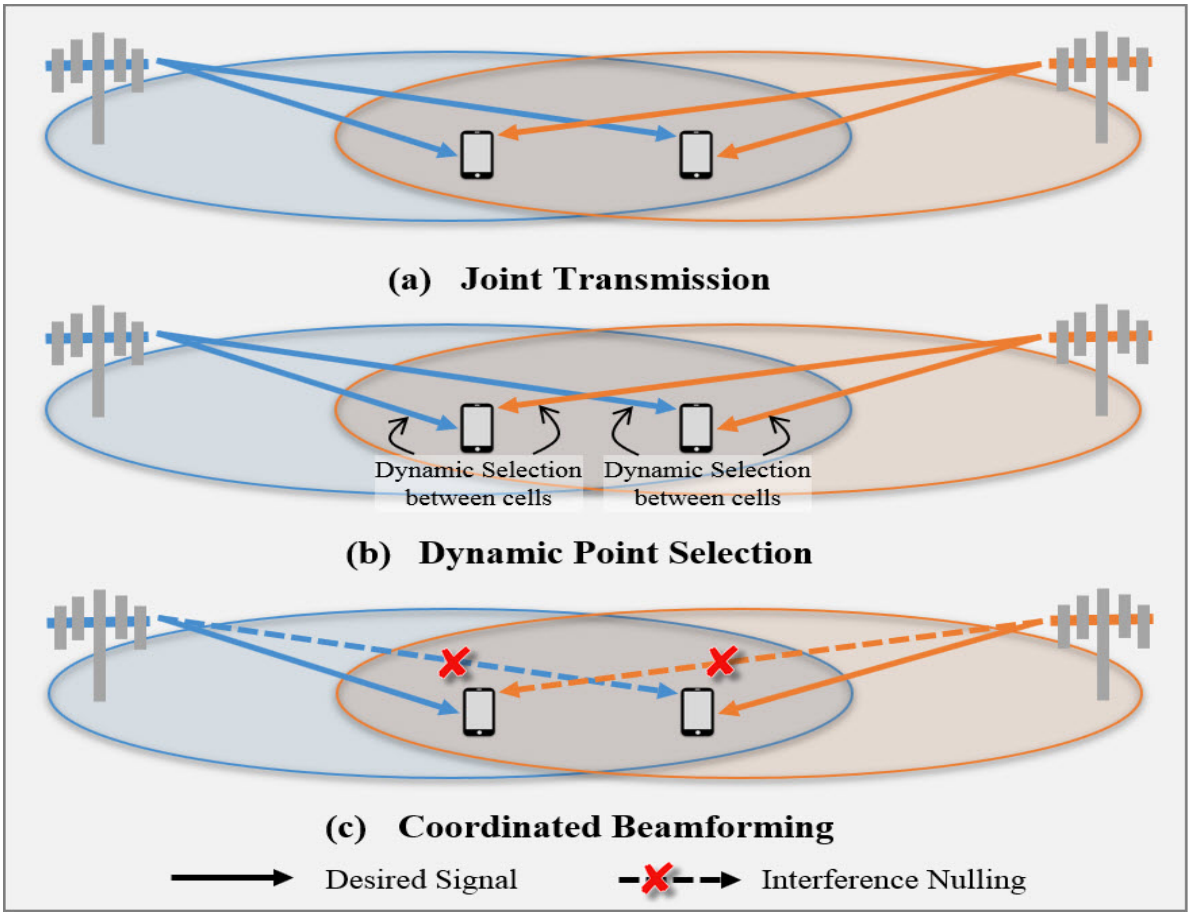

Figure 2.2: Types of Downlink CoMP.

2.2.2 Latency Requirements

Besides reliability, another mobility management objective of paramount importance is to minimize the length of transmission disruption during the HO process. In this subsection, I review the studies and research efforts aimed to minimize HO delay.

RACH-less Handover

Authors in [23] identified that RACH takes about 8.5ms out of 50ms interval required to accomplish HO task in LTE. Based on this assumption, they proposed a RACH-less HO technique to improve the latency by 17%. Authors suggest alternate means to perform the same functionalities as of RACH. For instance, RACH helps target BS to compute Timing Advance, though with lower accuracy. In the proposed RACH-less HO, UE can estimate timing advance from the time difference between the source and target

cell signals. Accuracy evaluation of the proposed approach will help gain confidence to the researchers. Such timing advance estimation method has been further evaluated in [29]. Alternatively, target BS can also compute timing advance through Sounding Reference Signals (SRS) which is used in LTE for uplink channel estimation as shown in [30]. However, this process might result in the timing advance estimation delay as it requires UE to be configured with SRS first. Initial uplink power, Physical Uplink Shared Channel (PUSCH) in LTE, normally known after successful RACH procedure, can be determined through source BS prior to HO initiation. Eliminating RACH is a novel proposal. However, UE in turn has to do more processing to compute timing advance that may lead to decreased battery life in a dense network. While RACH-less HO has its merits, the aforementioned challenges call for alternative approaches to reduce HO latency. One example of such approach is mobility aware caching.

Mobility Aware Caching

From the mobile users' perspective, more data rate alone is not enough to ensure better user experience. Any bottleneck in the distribution network between RAN and content servers can result in a prolonged Round-Trip-Time (RTT). During a HO, the chances of such bottleneck increase as momentarily the UE's QoE becomes dependent on two cells instead of one. This makes caching in the BS a useful tool to help accelerate the data delivery to the intended user. However, mobility degrades cache efficiency when UE moves to another BS. A study in [31] proposes to incorporate caching and computing ability deep into the base stations. The authors in [31] proposed a seamless RAN-cache HO framework based on mobility prediction algorithm (MPA). In the proposed scheme, the target BS is predicted for a UE with unfinished transmission during HO. This prediction is then used to pre-trigger the source RAN cache. This notifies the target RAN cache associated with the target BS to prepare for serving the UE and ultimately reducing latency. As a result, false probability of RAN-cache HO pre-trigger through

MPA though recorded to be less than 1.36% show an 8% increase in the maximal RAN-cache HO processing time. Researchers should benefit from the history of user mobility to come up with an improved algorithm. Mobility aware caching has been investigated in [32] to maximize the cache hit ratio that is defined as the number of requests delivered by the cache server, divided by the total number of requests. Compared to [31], authors in [32] considered both macro-cells and small cells. The first priority is given to the local cache followed by small-cell. However, if data is not received within the set deadline, macro-cell is then accessed to acquire data. Results assert that the proposed caching strategy outperforms prior caching strategies. The proposed cache scheme has a better cache hit ratio and low latency requirement for 5G networks.

paging-less Approach

Authors in [33] presented a novel frame structure with sub-millisecond subframe duration operating in Time Division Duplex (TDD) mode aimed for 5G networks. The frame structure carries UL beacon resources to enable a pagingless system for idle mode users. For connected mode users, UL beacons provide channel state information (CSI) for improved frequency selective scheduling. However, a caveat of this approach is that it can lead to an excessive amount of uplink messages. This in turn, may cause accelerated UE battery drainage and thus smaller battery life which is contradictory to one of the major 5G requirements.

2.2.3 Signaling Minimization

In both LTE and 5G NR, the processing unit is shifted to the edge, i.e., BS, primarily to reduce latency. However, this comes at the expense of increased signaling generated as the UE context is shifted from one cell to another during the HO procedure. This issue aggravates with the ultra-dense BS deployment. High signaling not only chokes the CPU of BSs, but also results in lower effective spectrum efficiency by consuming a substantial

amount of resources in the air interface. Too much signaling between neighboring BSs and BS-Core can result in potential congestion in the backhaul for the 5G networks with ultra-dense BS deployment. Reason being the expected myriad of mobile UEs, ultra-dense BS deployment, and added features that require high coordination e.g. multi-connectivity, carrier aggregation, and interference mitigation techniques. Thus, there is a possibility of network being paralyzed especially in busy hours due to the avalanche of signaling traffic. Signaling avalanche is an eminent threat in future ultra-dense networks. The research efforts by the research community to minimize the mobility signaling load can be loosely categorized in the following four sub-categories.

HO Signaling Reduction Through Mining HO Patterns

One basic but effective way to reduce HO signaling is to characterize HO behavior among cells to identify cells with an unusually large number of HOs or otherwise abnormal HO pattern e.g. ping-pong. Authors in [34] study the HO behavior of cells and propose a clustering model using K-means, to group cells with similar HO behavior. Further evaluation was done using actual HO attempt and HO success KPI of nearly two thousand WCDMA cells. The idea is to forecast the number of HOs and detect abnormal HO behavior among cell pairs using linear regression and neural network techniques. The detection is then used to perform targeted optimization of HO parameters in respective cells to minimize HO signaling. Adding a temporal component to training data can further increase the accuracy of the prediction.

Mobility Signaling Reduction Through RAN Centralization

Another method to reduce mobility signaling is to leverage the centralization of RAN e.g. using Cloud-RAN (C-RAN). Uladzamir et al. [35] recently proposed mobility aware hierarchical clustering approach (HIER) to group Virtual Base Stations (VBSs). Clustering based on the location of Radio Resource Heads (RRH) aims to reduce costly

HOs and thus, minimize signaling data. They also proposed location aware packing algorithm (LA) where inter-cluster mobility statistics are obtained by keeping track of UE movement, UE history to predict the traffic intensity between BSs. In addition, the history of inter-RRH HOs is considered as well. The proposed scheme when compared with affinity propagation clustering [36] can reduce up to 34.8% HOs, but at the cost of much higher requirement of RRHs. The approach can be beneficial for urban areas, but for less dense sub urban and rural areas, network deployment at this scale won't be feasible.

Mobility Signaling Reduction Through Cell Extension

An Extended Cell (EC) concept is proposed in [37] to dynamically form groups of several adjacent cells. HO performance improvement is rendered by increasing the overlapping area between two adjacent cells in the Radio over Fiber (RoF) indoor networks. The proposed approach reduces the number of HOs and the call drop probability during the HO by 70%. Although proven effective, it lacks the dynamic procedures to define ECs to optimize network resources. Shortcomings were addressed by authors in [38] by extending the idea and coming up with a proposal on the Moving Extended Cell (MEC). Here, each mobile UE is covered by 7-cell EC where each EC transmits the same user data at every instance. This in turn, reduces HO latency through early preparation. Evaluation results show the proposed architecture can totally avoid call drop and packet loss for UE's with a velocity of up to 40 m/s. The authors in [38] suggested that MEC is very efficient in tackling HO for mmWave cells but is vulnerable to throughput inefficiency as all seven cells in the cluster transmit for a single user.

Mobility Signaling Reduction Through Virtualization

Virtual Cell (VC) has been proposed as a solution by Hossain et al. in [39] to reduce mobility signaling while increasing the throughput efficiency of 60 GHz RoF network.

VC is a central part of an actual cell, and the remaining boundary area is divided into numbered tiles. Wireless Sensor Network keeps track of the UE location and periodically sends report to a centralized controller. Multiple Antenna Terminals (AT) cover a single cell, and only a single AT is activated at an instant. When the UE steps on one of the boundary-located tiles, the controller activates respective neighbor AT to transmit similar data. In the VC scheme proposed in [39], maximum of only two ATs can be activated for HO preparation in contrast to 6 in MEC [38]. End results of using VC concept show an increase of 33% throughput efficiency in comparison to MEC. Drawback of the proposal involves management of a wireless sensor network to track and report UE location. And if the UE velocity is high, the low powered sensors may not be able to timely report or even identify the presence of a high-speed user.

2.2.4 User Tracking

Location management, sometimes referred to as mobility tracking or user tracking, is defined as the set of procedures that determines UE location at any instance. User tracking is inevitable in cellular networks, so that incoming data from the core network can be delivered to the user. Densification of both cells and users, as well as increased mobility focused use cases such as Intelligent Transportation Systems (ITS)/Unmanned Aerial Vehicles (UAV) etc. bring new challenges to user tracking in 5G environment. The recent attempts to address these challenges can be loosely categorized into following three subcategories:

Distributed Tracking Area Update

A framework to minimize conflicting metrics, Tracking Area Update (TAU) and paging, is presented in [40] by distribution of Tracking Area (TA) into Tracking Area Lists (TAL) in two phases. First phase is offline, which is responsible to assign TAs to TALs using three different approaches. The first two favors paging overhead and TAU respec-

tively, while the third one uses Nash bargaining game to ensure fairness between paging overhead and TAU. Second phase is online which controls the probabilistic distribution of TALs on UEs by taking into account their behavior, incoming transmission frequency and mobility patterns. Numerical results were shown for the three approaches of the first phase, where the third solution provides a fair tradeoff between paging overhead and TAU. As a future step, results should be compared with prior schemes. No research work focusing on the horizontal or vertical deployment of TAs is present, therefore researchers can come up with smarter and more effective ways for operators to define Tracking Areas.

Hybrid Tracking Area Update and Paging

5G network will have large range of UEs and dense network deployment as discussed earlier. Hence, a huge amount of paging especially for millions of IoT devices is expected. As a result, signaling associated with paging may become enormous if currently available approach is used. To address this problem, authors in [41] propose a hybrid scheme in which either RAN or core network can initiate paging. RAN based paging with Tracking Area (TA) of just one BS is proposed for the RRC inactive [42] UEs to have low latency at the expense of high buffering capacity to transfer the content to the neighboring BS in case of user mobility. Meanwhile, core network-based paging is recommended to be used for idle UEs. Authors also proposed a hierarchical paging and location tracking scheme to minimize signaling load by assigning an anchor BS for location management. They conclude that RAN based paging is not efficient for high mobility UEs as TA is limited to a single BS. For hierarchical approach on the other hand, there should be more data management and processing for every user at anchor BS which becomes another single point of failure. Processor overload or X2 (inter-cell communication link in LTE) congestion, as a result, can disrupt the paging process.

Dynamic/Adaptive Tracking Area Update

Authors in [43] proposed an adaptive method that employs smart TAs to reduce the frequencies of TAUs and the sizes of paging areas. The proposed scheme uses the interacting multiple model (IMM) algorithm [43] to determine the estimated location of a UE at the time of the latest registration and provide a predicted location after a certain time frame. An experimental evaluation with an artificial trajectory showed that this approach cuts half of the extra location registrations compared with non-adaptive methods. Aside from that, this method also determines TA adaptively to significantly reduce the average paging sizes resulting in to lesser signaling for each paging attempts. As a future step, comparison results can be added for different types of mobile users at different speeds and trajectories to prove the effectiveness of their approach. Authors in [44] employed Apriori algorithm [45] for dynamic Location Area planning using call logs of several mobile users. Apriori algorithm finds frequent itemset using an iterative level-wise search procedure. By taking minimum support of 100%, Apriori algorithm can highlight those cells which serve mobile users every day. Based on this approach, authors in [44] suggested to create a dynamic TA based on more than 80% minimum support. Authors in [44] categorized mobile users into predictable, expected, and random groups based on the minimum support value. For each category, the authors propose to minimize location management cost by employing a suitable algorithm. However, the exact algorithms needed to minimize location updates, in this scheme, remain to be investigated as future work.

2.2.5 Cell Discovery

Traditional networks with High Frequency (HF) bands broadcast the reference signals (pilot symbols) for cell discovery as mandated by 3GPP. Majority solutions proposed in literature for cell discovery involve periodic scanning by the UE of these broadcast signals. The higher frequency of this periodic scanning ensures timely cell discovery but

results in increased battery consumption leading to trade-off between energy efficiency on UE side, network side, QoE, overall capacity and load distribution. In the following I discuss studies that have investigated these trade-offs and proposal solution to optimize one KPI or other.

Cell Discovery with UE Energy Constraints

5G networks will have heterogeneity of BSs with a motely of macro-cells and small-cells. A mobile UE connected to a macro-cell must scan for potential small cells to benefit from the high data rate and traffic offloading opportunity. If a mobile UE uses high scanning periodicity, it is likely to discover small cells in a more timely fashion. Thus, it may avail better offloading opportunities, but at the cost of reduced battery life due to increased amount of energy consumed by the scanning process, and vice versa. The investigation of this tradeoff is interesting and yet a challenging research problem as the optimal scanning periodicity, if exists, might be dependent on the cell density and user speed among several other factors. Authors in [46] use a rigorous approach that leverages stochastic geometry-based modelling of the network and empirical modeling of UE mobility. Analytical expressions have been derived to characterize and quantify the dependency of the UE energy efficiency on the cell density, cell discovery periodicity and the user velocity. Through analytical as well as Monte Carlo simulation results, it's been shown in [46] that UE battery life reduces significantly with increased cell discovery rate, while the UE throughput increases and vice versa. The key finding of this analysis is that, there exists an optimal cell discovery frequency for a given cell density and user speed statistics. This optimal cell discovery frequency maximizes the UE energy efficiency (EE) by achieving a Pareto optimal point between the capacity lost by missing cells with low cell discovery frequency and energy saved at UE in doing so and vice versa. Daniel et al. [47] proposed an energy efficient small cell discovery technique using radio fingerprints. In this proposed solution, network configures UE

with several radio fingerprints which are lists of cell-IDs and RSRP strength at different intervals. As a normal procedure, users served by the macro-cell performs the neighbor cell measurement as it moves around and compares those to the configured radio fingerprints. Upon a successful match, macro-cell is reported back which in return configures the corresponding small-cell. Authors show that energy efficiency of 70-80% is achieved on UE side by avoiding unnecessary small cell discovery measurements, and up to 45% on network side by small cell activation/deactivation. Practical use of this approach will be limited to shadowing since RSRP at a given point changes with time and the effect of environmental changes like rain/snow also affects the standard deviation of shadowing. Moreover, MDT will reveal better results as the location of the UE with respect to the small cell location can be known, followed by the successful small cell association.

Cell Selection with Network Energy Efficiency Perspective

The Information and Communications Technology (ICT) sector contributes around 2-3% to world's carbon emissions and is doubling every four years [48]. Since mobility is closely coupled with uneven and dynamic user distribution, the mobility patterns can be exploited to turn OFF/ON cells for enhancing energy efficiency. A solution to conserve network energy using such mobility leveraging approach is proposed in [48]. Decision of powering OFF the BSs is made using the UE velocity, receive power, BS load and energy consumption. In addition, HO to the small cell can be made only if the UE velocity and the cell load is lower than the respective thresholds. As a result, the low load cells can be powered OFF. However, the paper does not address when and how to turn ON the cell, as the powered OFF cell in the presence of the candidate UEs can have negative impacts on the capacity, efficiency, and user satisfaction. Random way point mobility models and the stochastic geometry theory are utilized in [49] to evaluate the energy efficiency of 5G networks. The network capacity and energy efficiency are evaluated for Ultra-Dense Cellular Networks (UDN) considering the user mobility. Results were

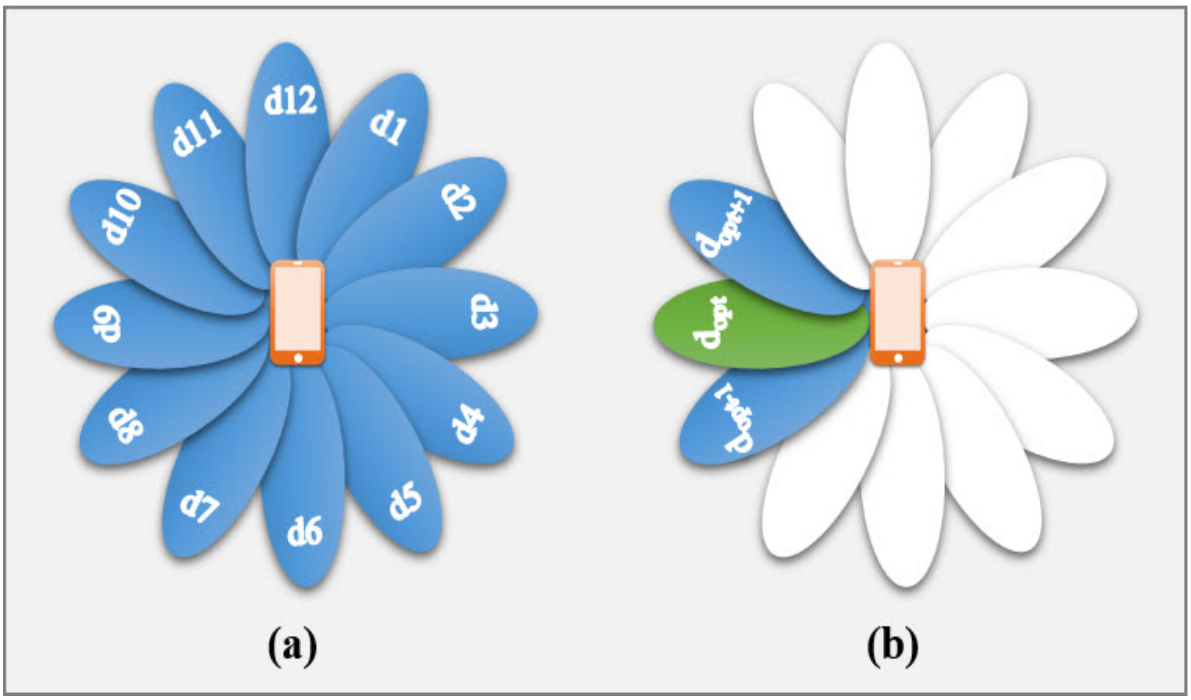

Figure 2.3: mmWave tracking. (a) Refresh procedure through 12 directions, (b) Refinement procedure through 2 directions.

demonstrated using Monte Carlo scheme where a user will keep stationary for a certain time, and then start moving to a random direction with variable but bounded velocity range. Results indicate that the energy efficiency decreases exponentially with increase in the small cell density. Energy efficiency decreases from 160bits/J to 155bits/J and 144bits/J when small cell density was increased from 10 cell/km2 to 15 cell/km2 and 20 cell/km2 respectively.

mmWave Beam Alignment and Tracking

The studies discussed in the last two subsections do not consider the several idiosyncrasies arising from the advent of mmWaves cells, as discussed in the following. mmWave band cell discovery becomes far more complex compared to the high frequency (HF) cells because of the high penetration loss and narrow beams [50]. Directional path in mmWave can deteriorate sharply due to rapid changes in the environment which calls for an intense tracking and alignment. The situation can be aggravated when considering mobile users. To address these issues, authors in [51] proposed two innovative

schemes by which UE can alternately scan the whole angular space exhaustively and select the beam with the best SINR. They propose the mmWave BS to send pilots in the configured finite directions at regular intervals, one at a time. The UE then scans for the mmWave-cell beam using two mechanisms: a) periodic refresh (PR) – The UE scans in all directions one at a time and the direction with the maximum SINR is selected; b) periodic refinement and refresh (PRaR) – The first optimal beam with the maximum SINR is selected as per the PR, and then the UE performs a refinement procedure by scanning the neighboring direction to adapt according to the changing condition or due to the UE mobility. This mmWave tracking approach is depicted in Fig. 2.3. Comparison between both schemes were done using the real-world measurement data collected in New York city on carrier frequency of 28GHz. As expected, PRaR is less energy efficient than PR because of the much frequent refinement procedure. However, they did not compare their schemes with the broadcasting approach or direct alignment schemes. Also, the scenario might arise where both the mmWave BS (in sending pilots) and the UE (in scanning pilots) are not synchronized with each other in terms of direction. Such a scenario is likely to lead to the tracking and alignment delay. Alignment process is done by scanning the adjacent beams only and can give sub-optimal results for the high-speed users. Esmaiel et al. [52] proposed a novel mmWave multi-level beamforming approach. mmWave link is established after multi-level beam search is conducted using a compressive sensing-based channel estimation. The estimated UE location is used to determine the number of beams and the bandwidth required for constructing the sensing matrix used in each beam searching level. Results show an increase in the spectral efficiency by 40% under good radio conditions. Authors in [52] also proposed a novel concept [53] of two-level control and user data (2CU/U) planes splitting, where the LTE BS and the WiFi access point provides control over the distributed sub-clouds and distributed mmWave BSs respectively. With the proposed approach, mmWave miss-detection probability as low as 10% can be obtained compared to 90% with the conventional approach when mmWave BS are deployed in a sparse manner. The result

can be further improved by incorporating the user movement historical data, and to observe the result for different UE speed.

HO in mmWave Band: Traditional HO is based on the Received Signal Strength (RSS) wherein pilot signal strength measured by the UE determines the cell-edge and thus lends assistance in performing HO to the target cell. This approach is ineffective for addressing the unique challenges associated with the mmWaves. In mmWave cells, the RF reception changes drastically with UE speed and direction. Hence relying on the RSS to anticipate a cell edge may not suffice. Authors in [54] suggest a novel Inter-Beam HO Class (IBHC) concept combined with the HO control and radio resource management functionalities. Initially, the user is assigned to a mobility classes depending on its estimate speed. The corresponding HO frequency is defined such that the high velocity UEs are expected to observe more HOs than the pedestrians. The mobile user is assigned a group of beams as per mobility class, load conditions and the expected path of UE. Each beam in the group contains similar resource allocation to improve the reception quality. HO is thus performed only at the edge of the beam-group. The underlying assumption in the proposed scheme is that the individual signals of each beam are perfectly synchronized. This can be true for low speed users; however, it may not hold for the high-speed users. Another strong assumption is the perfect estimation of UE velocity. UE velocity estimation is a big challenge even in the existing mobile networks, where the number of HOs in a moving time window are used to estimate UE velocity. Emerging networks with dense deployment of multi-frequency networks will make the prediction of UE velocity even a bigger challenge. Concept presented in the [54] can be extended by considering the relationship between the maximum user velocity and the mmWave footprint where its beneficial for the mobile user to camp to the mmWave cell. The study should include the signaling cost and energy consumption in scanning for the mmWave cells. In [55], authors leverage the concept of moving cell for train communication using 60 GHz band. To avoid the large number of HOs in high speed train, authors propose to employ the Radio over Fiber (RoF) technique. The key

idea is to make the serving cells move together with the train and thus provide smooth uninterrupted transmission to the passengers. However, for this scheme to be practical, the train's velocity and the direction needs to be pre-known to achieve synchronization. Furthermore, due to the inability to cope up with randomness of user mobility, this concept is not appropriate for mobility management in indoor environments. The state-of-the-art literature work reviewed in this section is focused on managing mobility in a reactive way. Two of the key challenges in mobility management in emerging networks that are not addressed by the current reactive mobility management paradigm in the industry and the associated literature in academia are high latency of the HO process and the large signaling overhead. These challenges become more important with the increasing fraction of mobile UEs, more bandwidth hungry applications and the advent of delay sensitive use-cases like self-driven vehicles. Proactive mobility management is an emerging paradigm that has the potential to address these challenges. It's a vital component by which the network operators can guarantee the success of the futuristic mobile networks. Key concept of the proactive mobility management and the recent studies that have presented few novel ideas to achieve the proactive mobility management are discussed in the next section.

2.3 Proactive Mobility Management

It is a well-researched fact that people tend to visit the same places repeatedly in their daily life, e.g. workplace, school, gym, parks, shopping venues, etc. This makes their movement to feature a high degree of repetition and hence predictability. According to some large-scale studies, this perceptibility can be as high as 93% [56]. This intrinsic predictability in human mobility can be leveraged to build models to predict the UE mobility patterns. In cellular networks, these models can be built by harnessing the large volumes of UE mobility related data such as call detail records (CDRs), GPS traces, and data traffic from existing networks. Following is the list of some of the potential

use cases of mobility prediction in the current and emerging cellular networks:

- Enhancing the overall QoS and QoE by reserving and managing radio resources a priori for users expected to arrive in a cell [57].
- Prevent failures and minimize HO delay e.g. by proactively triggering HO [58, 59].
- Prevent ping-pong HOs.
- Efficient load balancing e.g. by predicting cell loads and emergence of hot spots.
- Assist in cell activation/deactivation, and hence, conserve energy consumption.

Mobility prediction models in literature can be classified into three broad groups:

1. History based prediction models: In this type of prediction models, UEs next target cell is predicted based on the statistical analysis of historical records such as HO records or CDR records.
2. Measurement based prediction models: Such prediction schemes derive probability of user transition to next cell based on the real time measurements e.g. RSSI, SINR, distance, etc.
3. Location based prediction models: Current user location and in some cases urban transportation infrastructure is used to predict the future user location in the location-based prediction models.

In the following, I discuss the recent studies in literature that have made use of the two types of prediction approaches for various use cases.

2.3.1 History Based Prediction

History based mobility prediction approaches can be further divided into the following categories:

CELL TRACE BASED PREDICTION

Location prediction based on cellular network traces has recently attracted a lot of attention. Zhang et al. propose NextCell scheme [60] that utilizes social interplay factor to enhance mobility prediction. Social interplay is characterized by the convolution between entropy of the average call duration between two users, and the probability distribution of these two users to be co-located in the same cell. NextCell predicts the user location at cell tower level in the forthcoming one to six hours. It shows that inclusion of the social interplay improves prediction accuracy by 20% when compared to behavior periodicity-based predictor. However, results were not compared with the existing prediction schemes. Authors in [61] presented a HO prediction scheme that combines signal strength/quality to physical proximity along with the UE context in terms of speed, direction, and HO history. The presented scheme achieves 33.6% reduction in HO latency when compared with conventional HO approach.

MACHINE LEARNING BASED PREDICTION

Complex interaction between different components of a network can be well captured by Machine Learning approaches. For the same reason, much of the history-based prediction works revolve around machine learning based approaches. Authors in [34] argue that most of the research involving behavior prediction of a single UE is an infeasible and impractical approach. The argument is backed by the fact that some HOs are coverage based, while some are network initiated (e.g. load balancing). They propose to address these challenges by employing the K-means algorithm to group the cells with the most similar HO behavior into a cluster. Next, the future HOs were forecasted, and abnormal HOs were identified. The main target of the proposal is to minimize the signaling load by avoiding the abnormal HOs. Now I present some of the research work done on specific machine learning algorithms:

Support Vector Machine: Authors in [62] capitalize on Support Vector Machine (SVM)

to predict the user location in the next 5 seconds. A framework to minimize HO delay using mobility prediction is proposed. However, they did not validate the framework, neither did they compare their work with the existing proposals. In [63], SVM predicts the next cell in a real-time manner, by combining GPS data, short-term Channel State Information (CSI), and long-term HO history. The presented model was applied on a synthetic Manhattan grid scenario. Results show that CSI results in almost 100% better prediction accuracy compared to using HO history alone. Using different shadowing values to represent different terrain and environment can further strengthen the idea practicality.

Neural Networks: Few works in [64, 65] have leveraged neural networks for mobility prediction. The basic idea is to utilize the neural network to learn mobility-based model for every user and then make prediction about the future serving cell. Authors in [64] performed clustering of the input RSS samples through k-means. The clusters and input RSS samples were then fed to a classifying model, where neural network was used to predict the user position. Results show that the prediction accuracy increase by just 5% when compared to the prediction using neural networks alone.

Markov Chain Based Prediction

A large number of research studies have used Markov chain-based approaches for mobility prediction for their ability to yield better accuracy than most other predictors with lower complexity [66]. In the following, I review recent studies for commonly used Markov Chain (MC) variants:

Standard Markov Chain: Standard Markov Chain is a memory-less algorithm as the next state depends only on the current state and not on the sequence of the events that preceded it. Authors in [67] extracted trajectories of 4,914 individuals using 27-day log of the mobile network traffic data. They compared the original Markov algorithm with the Lempel-Ziv (LZ) family algorithm [68]. The core operation of the LZ predictor is

by maintaining a prediction tree which adds more complexity compared to Markov. It was concluded that although slightly more accurate, LZ family algorithm consumes a lot more resources and time than Markov algorithm. Most of the mobility prediction algorithms only consider spatial factors to predict future movements. Authors in [68] improved Markov Chain based model by adding a temporal factor and achieved 6% higher accuracy. Humans usually follow regular paths as discussed earlier, however, they may deviate from their accustomed routine at some instances. Authors in [69] proposed a practical model based on State Based Prediction (SBP) method to predict the place to be visited when the user's trajectory exhibits unexpected irregularities. When user diverts from the routine, SBP is employed to conduct the prediction. Experiments reveal that the accuracy of proposed model can reach more than 83%, which is higher than the accuracy of 60% achieved by LZ predictor used in [68]. Authors in [70] proposed an implementation architecture for the MOBaaS (Mobility and Bandwidth prediction as a Service). The MOBaaS can be readily integrated with any other virtualized LTE component to provide the prediction information. Spatial information (location history) and temporal information (time and day data) are collected and analyzed. The results show a 33% reduction in access time for the requested content using the MOBaaS prediction information can be achieved. Due to its appeal, several extensions of MOBaaS were proposed later. For example, in [71], authors stressed that MOBaaS can be implemented in a cloud based mobile network architecture and can be used as a support service by any other virtualized mobile network service. Authors also evaluated the feasibility and effectiveness of the proposed architecture. Fazio et al. [72] propose Distributed Prediction with Bandwidth Management Algorithm (DPBMA). The algorithm uses Markov Chains to predict the user movement at each BS in a distributed way. This makes the proposed solution different from many other studies [67, 69, 70] where Markov chains are used to improve system utilization by reserving resources prior to the HO. This helps in preventing the call drop occurrences. However, distributed algorithm means BS needs to do a lot of processing making this solution not an attractive option

for low cost BS or small-cells.

Enhanced-Markov Chain: In [73], subscriber's mobility is predicted using the enhanced Markov chain algorithm. The core idea is to add the behavior pattern and temporal data of the users from CDR into the Local Prediction Algorithm (LPA) and the Global Prediction Algorithm (GPA). LPA and GPA are based on first and second order Markov processes where transition probability to next cell depends only on the present cell, and both present and previous cell respectively. Results show that the proposed prediction methodology achieves prediction accuracy of 96% compared to GPA with prediction accuracy of 81.5%. However, users without any historical record in the training process showed poor prediction accuracy. Techniques such as particle filter or Kalman filter can be employed to increase accuracy for new users.

Semi-Markov Model: Authors in [74] argue that both discrete and spatial Markov Chain assume human mobility as memory less. By using these approaches, I can achieve spatial prediction of future cell, but time factor cannot be incorporated. To address this concern, authors predicted HO to the neighboring BS using Semi-Markov Model. Semi Markov process allows for arbitrarily distributed sojourn times. Experimental evaluation leveraging on the real network traces generated by the smartphone application showed prediction accuracy of 50% to 90%. An extension of this approach can be to have ping-pong HO predictions.

Hidden-Markov Model (HMM): Ahlam et al. [75] proposed HO decision algorithm (OHMP) using HMM predictor to accurately estimate the next femto-cell using a) the current and historical movement information, and b) the strength of the received signals of the nearby BSs. The performance of OHMP is validated by comparison with the nearest-neighbor and random BS selection strategies. Results show that the number of ping-pong HOs reduce by 7 times when considering dense deployment of femto cells. Results in [75] are demonstrated for a single user scenario only and does not portray futuristic cellular networks with large number of users. To address this concern, same

set of authors extended their idea in [76] by incorporating multiple UEs. They take into consideration the available BS resources of serving femto-cell and interference level from the target femto-cell. The presented OHMP-CAC algorithm introduced a proactive HO scenario where HO is triggered when SINR of the serving cell reaches a predefined threshold. OHMP-CAC minimized the number of HOs by 64% and reduced the average HO decision delay by up to 75% when compared with the traditional RSSI based scheme. As discussed earlier, mobility prediction using Markov chain is a memory-less system as future state can only be determined by the current state. On the other hand, enhanced Markov Chains are based on historical data, but their application is very complex. Moreover, mobile operators may not be allowed by the customers to use their historical data due to privacy concerns. Even if historical records are accessible, HO delay might still be observed due to the extraction and processing complexity of historical records. Due to these factors, history-based prediction algorithms might render impractical.

2.3.2 Measurement Based Prediction

Measurement based mobility prediction approaches are more accurate than history-based mobility prediction schemes. However, the processing complexity due to the measurement procedure cannot be ignored.

RSSI Based Prediction

Soh and Kim [77] introduced RSSI based mobility prediction while keeping in view different UE velocities. They incorporated UE trajectory and road topology information to yield better prediction accuracy. The prediction goal is to achieve timely HO and limit the probability of forced termination during HOs. In addition, bandwidth reservation scheme was proposed that dynamically reserves radio resources at both participating BSs during the HO procedure. Results show that proposed mobility prediction scheme helps achieve almost similar forced termination probability as the benchmark scheme

with perfect knowledge of the mobile UE's next cell and HO time. Authors in [78] proposed an RSSI-based prediction scheme to reduce VoLTE end-to-end delay and HO delay under different UE velocities in mixed femto-cell and macro-cell environments. The core idea is to send the measurement reports based on user velocity and predict when and where to trigger HO procedure. As a result, HO delay is reduced by 28%. For ultra-dense BS deployment, mobile UE may not perform HO to each BS on its trajectory. Future work can include the consideration of load condition, so that both low latency and adequate resources can be guaranteed for improved QoE. The decision to skip the HO to a better radio condition cell can be based on dwell time or cell load condition. Next femto-cell prediction based on radio connection quality and cell load status is presented in [79]. Authors proposed two cell selection methods; a) BS prediction after analyzing the collected data of average RSSI from nearby femtocells, b) using cognitive radio to sense neighboring femtocells load before triggering HO. Results show that appreciable number of HOs can be avoided when compared with only RSS based HO approach. Thus, data interruption during HO and chances of Radio Link Failure e.g., due to ping-pong HOs can be avoided. Authors in [80] argue that RSS alone should not be considered when performing inter-RAT HO. Instead current RSS predicted RSS and available bandwidth should be considered. They proposed Fuzzy logic based Normalized Quantitative Decision (FNQD) scheme which aids in eliminating ping-pong effects in HetNets. This work can help realize improved mobility management for LTE-Unlicensed (LTE-U). However, the key performance metrics such as throughput and HO delay should be added for validation purposes.

Measurement Report Based Prediction

Song et al. used Grey system theory in [81] to predict the (N+1)th measurement report (MR) from Nth MR for high speed railways. The key idea is to utilize the predicted MR to make proactive HO trigger decisions. Their findings showed that the difference

between predicted MR and actual MR is within 1%. Thus, the proposed scheme is capable of proactively triggering HO in advance and HO success probability is enhanced from 5% to 10%.

User Direction Based Prediction

Authors in [82] present a user mobility prediction method for ultra-dense networks using Lagrange's interpolation. They predicted user's arrival into their neighboring femtocells based on users moving direction and the distance between users and neighboring cells. The presented approach increases the prediction accuracy when compared with only distance based and direction based mobility prediction. However, the performance of their proposed prediction scheme is not compared with other existing schemes to quantify the performance gains.

User Velocity Based Prediction

Higher UE velocity imposes additional threat to reliability making prediction of UE velocity extremely important to help tune the parameters more effectively. 3GPP based solution assigns mobility states (high, medium, low) depending on certain number of HOs in a moving time window. However, this technique will be inefficient in 5G networks with unplanned and highly dense deployment of heterogeneous BS having variable cell radius. UE velocity was estimated in [83] based on the sojourn time sample and accuracy was analyzed via Cramer Rao Lower bound. Numerical results show that the velocity prediction error decreases with the increase in BS density. The authors in [83] further extended their idea in [84]. The predicted UE velocity was used to assign the appropriate mobility state. Validation was done by gathering statistics of the number of HOs as a function of UE velocity, small cell density, and HO count measurement time window. The results show similar conclusion as in [83] that the accuracy of a suitable mobility state detection (known from UE velocity) increases with increasing small cell

density. Authors in [85] observed that mobility in urban areas depends on the traffic laws and is affected by the behavior of other people (red signal, other driver brakes etc.). They predicted user mobility based on the observation that a UE with constant velocity will probably go straight, while a UE decreasing in velocity might indicate stoppage on red light or a turn to a different direction. User location in their model is estimated from uplink time difference of arrival or provided by the UE via AGPS while velocity estimation is achieved by increasing sampling rate of location or by Doppler shift. Results showed that overall throughput can be enhanced by 39%, 31%, and 19% for UE velocities ranging from 25, 50, 75 km/h respectively.

2.3.3 Location Based Prediction

The knowledge of UE location can assist in an improved mobility prediction. Effective localization when combined with the mobility prediction algorithms can yield more efficient HO related QoE results. Soh and Kim in [86] presented a decentralized Road Topology Based mobility prediction technique where the GPS equipped UEs shall perform mobility prediction based on approximated cell boundary data that was shared by the serving BS. Cell boundary data is represented by a set of points at the cell edge and is populated based on historical measurement reports sent by UEs. UE at the cell edge will thus report the corresponding location ID back to the BS, and proactive resource reservation at potential BS can be achieved. Results show considerable reduction in forced termination compared to a reactive HO approach without mobility prediction. This approach can be applied to the macro-cells but is not reasonable to small cells as mobile UEs will have to send a lot of high-powered uplink messages at cell edge (high path loss condition). This can lead to an increase in HO failure due to high uplink RSSI. Moreover, UE battery consumption will be high. Authors in [86] proposed mobility prediction scheme based on road topology information. The main idea is based on the approximated cell boundary based on prior HO instances, being configured by

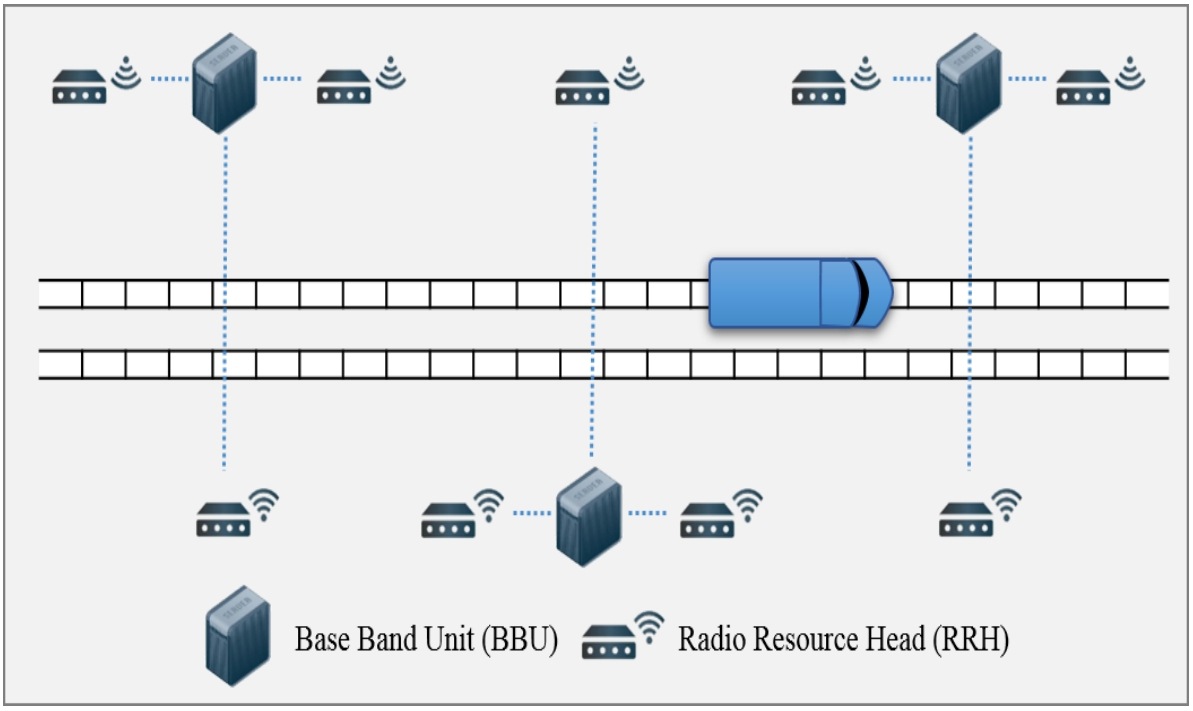

Figure 2.4: Directional network deployment using RRHs [3].

the serving cell. The authors in [86] extended their idea in [87] to add the temporal component to mobility prediction. The scheme uses linear extrapolation from a UE positioning data to predict its HO cell and time. 70% mobility prediction accuracy was achieved compared to 60% in their prior work [86]. Location based mobility prediction approaches assume all cell phones to have an accurate position information, which cannot always be guaranteed. Moreover, security concerns of the subscribers may hinder the collection of necessary data to realize accurate cell boundaries. While proactive mobility management seems to be a great fit to address the stringent QoE requirements in the emerging cellular networks, the trivial network dimensioning tasks should be planned while keeping in view the effect of mobility on the deployed network.

2.4 Mobility Oriented Network Planing and Optimization

Realizing massive potential of network densification to address the capacity crunch has introduced additional network planning challenges as discussed by Azar et al. in [88]. One such challenge will be faced due to larger fraction of the mobile users in the network;

hence, the network must be planned while considering mobility management in mind. Suitable network architecture can help achieve QoS goals while keeping the cost (e.g. signaling) to a minimum, and ultimately help attain higher network efficiency.

2.4.1 Signaling Minimization by Reduction in Handovers in High Speed Trains

Since considerable signaling overhead is being generated due to a single HO, network planning and architecture aimed to reduce the number of HOs can certainly be very effective. High speed train users are subjected to frequent HO as they move along the track. Apart from a huge amount of signaling data generation, they can also encounter severe issues like RACH failure, late HO, Radio Link Failure (RLF), and Release with Redirect (RwR). Futuristic mobile networks with smaller footprint small cells will cast an even bigger risk. To address this problem, authors in [89] presented a HO minimization technique where they propose to install an antenna on top of the train that will perform connectivity and trigger HO with covering BSs. Network deployment approach has been demonstrated in Fig. 2.4. This elevated antenna interfaces with an inner-train network to serve the passengers. Thus, instead of several users performing HOs simultaneously, only one HO will be performed by the elevated antenna. This not only reduces signaling load, but also minimizes the risk of HO failure as UEs will not experience penetration loss of 20-30 dB inside the train. Field trial conducted on a 2.4km run showed downlink throughput of 1.25Gbps. The concept of elevated antenna seems practical and is studied even by 3GPP [3]. However, single point of failure lies on its very foundation; if elevated antenna fails and observes HO failure then the multiple users being served under that antenna will have disrupted data transmission. Intelligent switching of the elevated antennas based on proximity to the BS can not only avoid HO failure but also deliver high throughput due to better SINR, but at the cost of complexity and cost. Another drawback will be the latency due to the addition hop between the top-mounted antenna

and the inside-train UEs. As a result, self-driven trains in the near future might not achieve the required latency QoE goal.

2.4.2 Changing Core Network (CN) to Achieve Latency Goals

Authors in [14] studied the latency, HO execution time, and coverage of four live LTE networks based on 19,000 km of drive tests. The test was conducted in a mixture of rural, suburban, and urban environments. Their measurements reveal that the lion's share of latency comes from the core network rather than the air interface. Based on the study in [14], Johanna et al. [90] proposed a new entity called the edge node that integrates MME and control plane part of SGW and PGW. Each edge node covers several BS, and when UE moves to coverage of another edge node, the application server and gateway is also shifted to minimize the latency. This approach helps to reduce latency for every HO done within BSs connected to the same edge node. However, HO associated with inter-edge node is followed by IP address reassignment and application-server transfer, which adds to delay and data interruption. Keeping in view that the number of 5G subscriptions will be 2.6 billion by the year 2025 [4], authors in [8] suggested a simplified 5G core network which will be connectionless, and will incorporate the best effort without the support for node mobility. The core idea is to have a legacy internet-like core network that will not be QoS centric, and the majority of the traffic will flow through default bearers only. Experiments were conducted on a smartphone to show that video streaming, web browsing, and messaging will work well, thus, the future core network can be radically simplified, resulting in a cost-effective solution. The authors in [8] mainly focused on a simplified core network with low complexity. Over-simplification of core network is not a practical approach as major functionalities of billing and access control cannot proceed. Similarly, IP re-allocation at every single HO is not feasible and may result in high latency or even packet loss.

2.4.3 C/U Plane Split

With improvement and advancement in the hardware technology, telecom operators can benefit from decoupling control and user plane (see Fig. 6.1). By doing so, future mobile networks with the composite of macro-cells and small cells can be used intelligently for efficient resource utilization. Moreover, signaling overhead from large number of HOs can be minimized by assigning control plane and user plane to macro-cells and small cells respectively. Authors in [91] address mobility support for high density, flexible deployment of small cell architecture with flexible backhaul using Localized Mobility Management (LMM) technique. The first step centralizes control-plane from small cells to a Local Access Server (LAS). The second step allows individual small cells to handle the mobility events, but still requires the LAS to act as a mobility anchor. Analytical model based on discrete time Markov chain is used to evaluate the average HO signaling cost, average packet delivery cost, average HO latency and average signaling load to the core network. Results show that average HO latency decrease by 10ms compared to the 3GPP scheme [14]. Authors in [92] minimized signaling overhead in a 5G network with a high density of mmWave BSs serving users under the umbrella of macro BSs. C/U split was employed where macro BS provides the control plane and several mmWave cells were group into clusters. Inter-cell HO signaling was curtailed by using a gateway cluster controller, resulting in signaling reduction in the core network as well. Results show that normalized X2 signaling overhead reduces from 100% to 10% as the density of the deployed mmWave cells increases. Authors in [93] targeted latency minimization in their proposed novel mobility management scheme for intro-domain handover (HO within the same SDN domain) and inter-domain handover (HO across different SDN domains). Layer 2 information and buffering approach was used to achieve HO latency of just 400ms compared to the legacy DMM with 100ms of HO latency. While proactive mobility management and mobility-oriented network planning seem to deliver promising results, the constant temporal variations in a live network and the importance of key

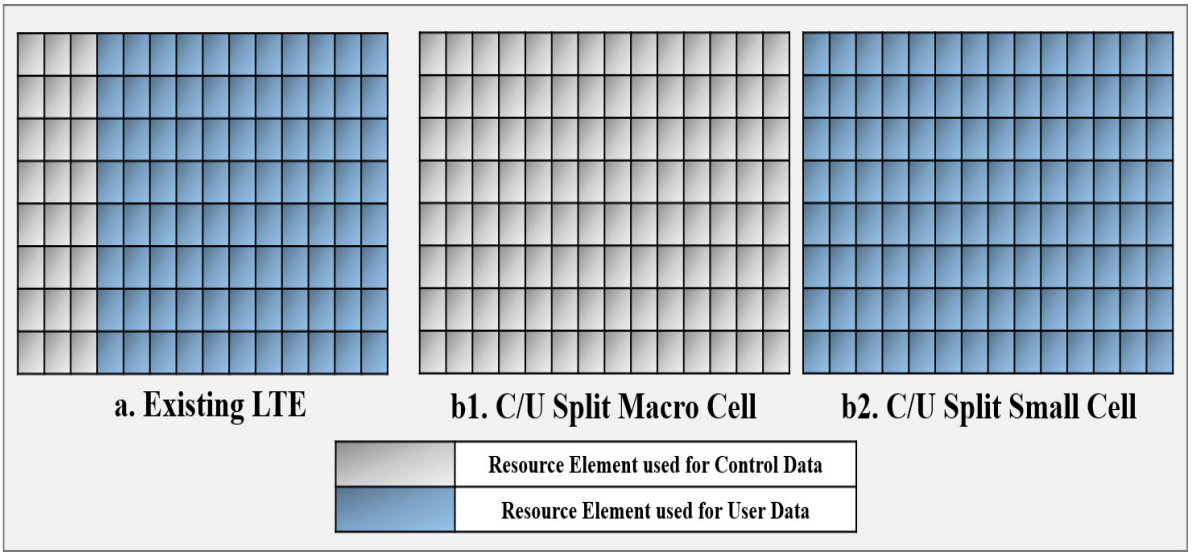

Figure 2.5: Frame structure for legacy LTE vs C/U plane split architecture.

landmarks can be addressed by introducing Artificial Intelligence (AI) to the cellular network domain.

2.5 AI-Assisted Mobility Management

In recent years, AI has gained much popularity for proactively managing mobility in future cellular networks. This is primarily because of an increasing number of configuration parameters and due to the complex interaction between network parameters and associated KPIs (as illustrated in Fig. 1.7). Once the research community is able to overcome those complex challenges, AI-assisted solutions will have a revolutionary effect on the telecom industry. The tutorial section of this chapter can help researchers understand the convoluted interplay between the network parameters and affected KPIs. Now I present some of the AI enabled mobility management solutions present in the literature.

2.5.1 Mobility Prediction using AI

The mobility prediction algorithm is presented in [94]. Authors use realistic mobility patterns to capture the human movement and a 3GPP compliant 5G simulator was used

to represent the HetNets scenario. Results show that mobility prediction accuracy of almost 87% can be achieved for 2dB shadowing with XGBoost compared to 78% with Deep Neural Network (DNN). The work can be extended by using time series predictors such as recurrent neural network or LSTM.

Based on HO attempts per hour, authors in [34] cluster cells into different groups with similar HO profiles using the K-means algorithm. For each cluster, hourly HO attempts were forecasted using linear regression, polynomial regression, neural networks and gaussian processes. the highest R2 value of 0.99 was obtained when using the gaussian process. The proposed model then checks for abnormal HO behavior e.g. ping-pong. Future work can be to proactively predict abnormal HO behavior ahead of time and to recommend suitable proposed parameters to prevent HO KPI degradation.

2.5.2 Leveraging AI to improve HO in HetNet

Authors in [95] employed XGBoost supervised machine learning algorithm to perform partially blind HOs from sub-6GHz to co-located mmWave cell. Authors show that this machine learning-based algorithm to achieve partially blind HOs can improve the HO success rate in a realistic network setup of co-located cells. The proposed algorithm should be compared with the existing HO approach in terms of energy efficiency and RLF to further validate the efficacy of the algorithm.

The idea of inter-frequency HO from a macro-cell to a non-co-located high frequency cell with a much lower footprint is presented in [96]. The authors use the Random Forest classification approach and also presented a use case of load balancing by which an efficient resource utilization for the static users can be achieved. The shortcoming in the presented approach is that for high-speed users, the load balancing based HO to smaller footprint cell may be inefficient due to large HO rate and the resultant signaling overhead and chances of HO failure.

Authors in [97] develop a Reinforcement Learning (RL) based HO decision algorithm

for the mmWave cells by taking into account the user experience as a weighted sum of throughput and HO cost. Based on the user's mobility information, the optimal beamwidth is selected by considering the trade-off between the a) directivity gain and b) beamforming misalignment. The algorithm approves the HO trigger for mobile users depending on UE velocity and BS density. The work can be extended by evaluating the signaling overhead reduction and throughput gain achieved when compared with other existing algorithms in the literature.

2.5.3 AI-Assisted RLF Avoidance

Authors in [98] predicted the RSRP of the serving and the HO target cell using Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN). The algorithm also predicts RLF instances with an accuracy of 84% using only RSRP as an input feature. An extension to [98] has been made in [99] where other features like SINR, out-of-sync identifier, RACH issues, and max RLC retransmission have been used for RLF prediction. A wrong HO avoidance algorithm has been proposed in [100]. It uses neural networks to prevent the HO to BSs which are affected by the undesirable radio propagation scenarios in the network, e.g., coverage hole caused by an obstacle. The proposed algorithm enables a UE to learn from past experiences (coverage unavailability) to select the best cell for HO in terms of QoE. The authors show that their algorithm helps achieve users to successfully complete the downlink transmissions more than 93% of the time. However, the simulation environment is quite simplistic where the UE traverses a straight line with only three BSs along the way. Hence, the movement of UE is almost deterministic, and the Neural Network can easily learn its pattern and can identify the optimal BS to perform HO. Furthermore, a single test UE gives a limited evaluation of the proposed algorithm. Elaborated results with a HetNet scenario and arbitrary movement of multiple users will have more realistic results.

CHAPTER 3

SyntheticNET: A 3GPP Compliant Simulator for AI Enabled 5G and Beyond

3.1 Introduction

Mobile cellular networks are one the most complex and expensive engineered systems in existence today. Given a typical modern Base Station (BS) has thousands of configuration parameters, optimal planning, configuration and continuous post-deployment optimization of the nation-wide mobile network often containing hundreds of thousands of diverse sites is already one the most challenging and resource hungry engineering problem. In wake of internet of everything, e-governance, e-commerce, e-health and ubiquitous consumption of infotainment, optimal design and operation of emerging mobile networks is one the key propellers of the emerging digital society.

While rapid evolution of the cellular technologies towards 5G and beyond is a vital step forward to meet the capacity crunch, it further aggravates the complexity challenge being faced by the operators today. This is because the number of parameters per site and number of sites per unit area continue to rise making mobile network too complex to optimally design, configure, operate, and manage. This calls for tools that can enable investigation and realistic evaluation of a myriad of new system level configurations and features in various deployments and use case scenarios.

Academic research community has heavily relied on mathematical models e.g., such as ones employing stochastic geometry, to get insights into the system level performance of various deployment scenarios [101, 102, 103, 104, 105, 106]. However, to achieve tractability, these models have to build on countless restrictive assumptions and simplifications with respect to user and BS location distributions, transceiver architecture

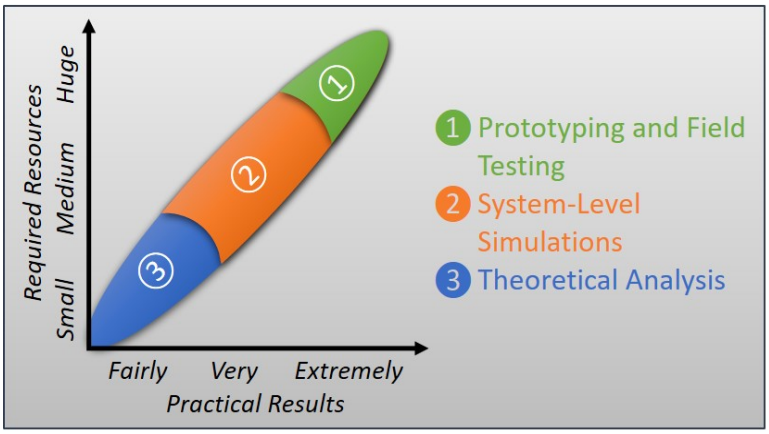

Figure 3.1: Role of Simulators in Network Performance Analysis.

and configurations and propagation characteristics to name a few. Furthermore, such models are static in nature and fail to capture the impact of dynamics that are peculiar to mobile networks such as user mobility, handover (HO) and so on.

Field trials offer the most realistic evaluation of a new network design, solution, or feature. However, relying on field trials alone to test every proposed design, solution or feature is not practical owing to the cost, time and effort required to conduct field trials. For this reason, only the most promising designs can be worthy of investment and resources needed for the field trials. In addition, given the large investments and stakes at risk, mobile operators want to minimize the chances of significant network performance impairment of a live mobile network even during the trial phase.

For 5G networks, given further increase in complexity, the process of designing an optimal network configuration that can maximize all the Key Performance Indicators (KPI) such as coverage, capacity, retainability and energy efficiency is even more challenging task. Identifying and maintaining the optimal network configuration is necessary for network operators to fulfill the promises of the much anticipated 5G networks. Deploying the new 5G network and innovative network functionalities and solutions being proposed for efficiency enhancement in 5G and beyond, particularly AI based network automation solutions as proposed in [107, 108, 109], in a real world without prior testing, will be a costly process and cannot be done practically. (See Fig. 3.1).

To address this problem, system level simulators are widely used in both industry and academia. Many 5G simulators emerged to date but, as per survey conducted by the authors and concluded in Table 3.1, none of them comprises of all the key components of 5G standard. Most importantly, as can be seen in Table 3.1, none of the existing known simulators have the specific features and flexibility needed to implement and evaluate an AI based design and zero touch automation framework envisioned for emerging networks as proposed for the first time in [107]. To tackle this problem, I have newly developed a system level simulator in Python platform, named SyntheticNET. The SyntheticNET simulator is modular, flexible, microscopic, versatile, and built in compliance with the 3GPP Release 15 [110]. The presented simulator supports a large number of unique features such as adaptive numerology, actual HO criteria and futuristic database-aided edge computing to name a few. Instead of an Objected-Oriented Programming (OOP) based structure like existing simulators [111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123], SyntheticNET simulator supports commonly used database files (like SQL, Microsoft Access, Microsoft Excel). Site and user information, configuration parameters, antenna pattern etc. can be directly imported to the simulator. As a result, the simulation environment is more realistic and closer to actual deployment scenarios.

Python based platform and the flexibility of different input and output data formats in SyntheticNET simulator allows validation of different Self Organizing Networks (SON) related features as well as new AI based network automation solutions [107]. Mobile operators can use it for planning, evaluating or even optimizing 5G networks. Research community can also benefit from it by implementing the new ideas on a true 3GPP-based realistic 5G system level simulator.

3.1.1 Related Work

Recently many simulators targeting 5G network characteristics have been developed [111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123]. However, the survey of

Table 3.1: Comparison of SyntheticNET simulator with existing 5G simulators.

Name	Platform	Adaptive Numerology	HO Support	Realistic Mobility	QCI Support	Scheduling Support	mmW Support	Cloud Based	Short Description
SyntheticNET	Python	✓	✓	✓	✓	✓	✓	✓[a]	Link-Level simulator System-Level simulator Conforms to 3GPP-based 5G standard Support of Database-aided edge computing
Matlab/Simulink [111]	Matlab	✓			✓	✓			Link-Level simulator
ns-3 [112]	C++					✓	✓		Link-Level simulator Event Driven: *a. Not suitable for large networks* *b. Takes long time*
OMNeT++ [113]	C++		✓			✓	✓		Event Driven: *a. Not suitable for large networks* *b. Takes long time*
OPNET [114]	C/C++					✓			Event Driven: *a. Not suitable for large networks* *b. Takes long time*
OpenAirInterface [115]	C++					✓			Network-Level simulator Support of core network: *a. Increased complexity* *b. Limits the #nodes*
5G-K [116]	C++					✓		✓	Link-Level simulator System-Level simulator Network-Level simulator
Vienna-5G [117]	Matlab	✓				✓	✓		Improved version of Vienna 4G
NYUSIM [118]	Matlab						✓		mmWave heterogeneous networks
C-RAN [119]	Matlab					✓	✓	✓	5G Cloud Based Networks
GTEC [120]	N/A								Link-Level simulator
X.Wang et al. [121]	Matlab					✓			Link-Level simulator System-Level simulator
V.V.Diaz et al.[122]	N/A						✓		Pathloss model for different scenarios: *a. Rural/Urban* *b. Macro/Small Cell* *c. LoS/NLoS*
Ke Guan et al. [123]	N/A								Ray Tracing with use-case of V2V communication

[a] SyntheticNET supports database-aided edge computing.

these concluded in Table 3.1, reveals that each of this simulator represents only a selected set of features present in 5G standard [110]. Among the available 5G simulators, Matlab [111] is the most advance 5G link-level simulator having the support of flexible frame structure, ability to select one of different resource scheduling techniques available and can incorporate mmWave channels as well. However, unlike SyntheticNET, it is not a system level simulator. Few important features that MATLAB based 5G simulator does not support include realistic mobility and HO mechanism, categorization of User Equipment (UE) as per QoS Class Identifier (QCI) and having cloud-based network deployment to name a few. Another popular simulator is Vienna 5G simulator [117] that is an open source system-level simulator for academic purposes and is based on Matlab platform. Unlike [111], Vienna does support cloud computing as well. However, this simulator lacks a key feature i.e., realistic mobility modelling and HO support.simulator also lacks some vital features to mimic a real cellular network such as realistic mobility modeling and HO support.

There are few other popular discrete-event 5G network simulators such as ns-3 [112], OMNeT++ [113] and OPNET [114]. Event driven simulators have a major portion of the protocol stack implemented in them, and the packet-oriented nature of these simulators exhibits quite accurate link-level results. However, their high computational and network deployment complexity hinders them from modeling large Radio Access Networks (RAN) needed for more realistic analysis. These simulators are more suitable for core side modeling and they cannot provide visualization of the crucial RAN KPIs such as coverage and capacity. The need of implementing and testing large RAN deployments can be highlighted from the fact that 5G networks will have an ultra-dense BS deployment with huge number of mobile subscribers which will include sensory devices, self-driven cars etc.

3.1.2 Contributions

The need for a Python based system level simulator for 5G and beyond stems from the new use cases and design features anticipated in 5G and beyond [124]. These include smart vehicles and transport, critical control of remote devices, human machine interaction, and broadband and media everywhere.

During the planning and development of SyntheticNET simulator, I make sure flexibility and modularity when implementing both existing network functions (e.g., propagation models, scheduling algorithms) and new network functions (user mobility, HO criteria, database-aided edge computing etc.). Thanks to the modular code structure, the SyntheticNET simulator is well-suited to the requirements of emerging network scenarios and its use cases, even beyond the scope of 5G.

For academic purposes, a free version of SyntheticNET will be available soon. In the following, I highlight the key features that make SyntheticNET simulator fit for the simulation of emerging cellular networks.

- SyntheticNET simulator is first python-based 5G network simulator and thus has the capability to handle large amount of data and access to Python based Machine Learning (ML) libraries. This unprecedented capability makes SyntheticNET simulator the first simulator that is purpose-built to test AI based network automation at all layers and validate the already standardised SON and next generation SON features.

- First microscopic simulator where each cell can have a unique parameter configuration that can be loaded in various industry compliant formats to model real datasheet-based features such as antenna patterns, clutter, BS amplifier, etc. Unlike OOP based BS deployment where BSs are deployed as per an underlying distribution, SyntheticNET simulator can import a database of site information corresponding to real deployment. Such is maintained by mobile network opera-

tors with a detailed description corresponding to individual BSs. The database constitutes of location, tilt, power, azimuth, height, signal propagation description, and even other low-level details.

- Similar to BS database, SyntheticNET simulator can import UE specific details such as location, type (static/mobile), height, number of antennas from a real database.

- Ease of calibration through ML based models [125] from traces of real RSRP data, and then predicting accurate RSRP information on the bins where real measurement data is unavailable or has similar characteristics (terrain, BS tilt, user mobility).

- Support of vendor specific or measurement realistic antenna pattern specifications for individual cells.

- Apart from having well known mobility patterns like random way point, SLAW model etc., SyntheticNET simulator can replicate realistic user mobility patterns through integration of Simulation of Urban MObility (SUMO) simulator [126]. This realistic mobility path not only includes the user commute from home to office and back, but random trips to marketplace and entertainment areas can be configured as well. This can help us achieve realistic spatial and temporal load distribution across the deployed network area.

- The SUMO based mobility module in SyntheticNET simulator can also incorporate actual street, highway, walkway topographic data in various formats.

- Historical UE mobility paths can be imported directly from a real network, and can be leveraged by SyntheticNET simulator to have better insights related to mobility, load distribution, user experience etc.

- SyntheticNET simulator supports flexible frame structure of 5G and corresponding Physical Resource Block (PRB) size and Transmission Time Interval (TTI).

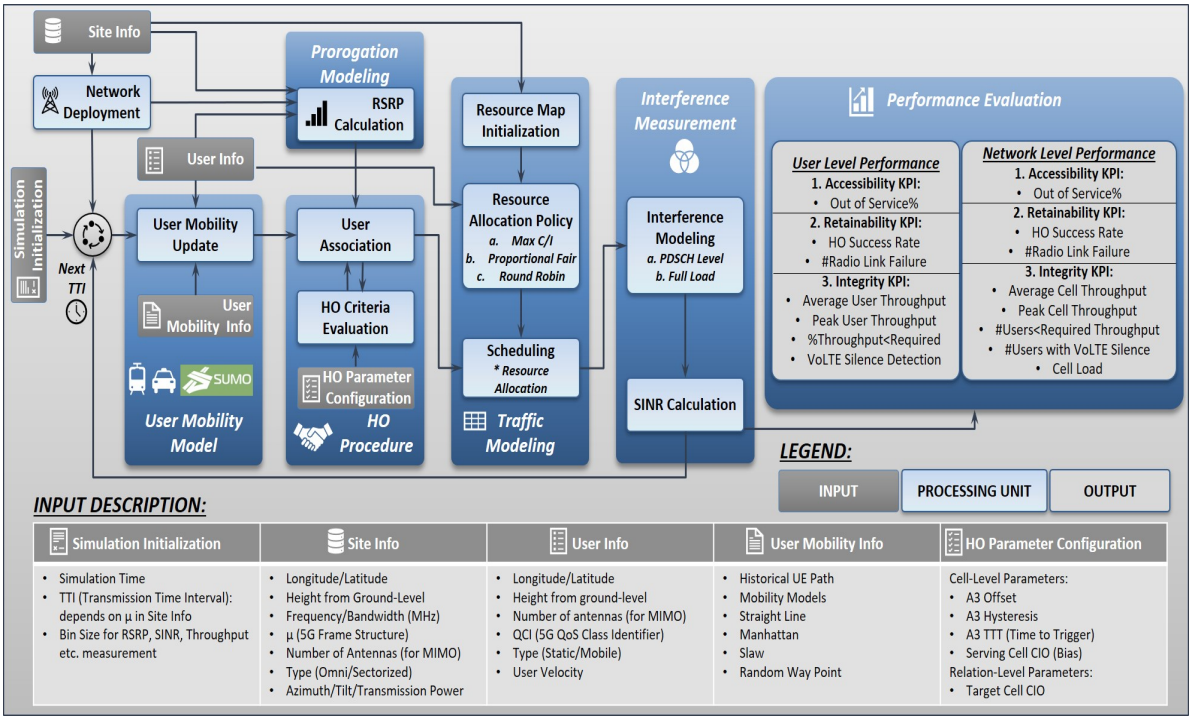

Figure 3.2: SyntheticNET simulator high-level block diagram.

- SyntheticNET simulator models realistic HO criteria where each cell can have individual HO related parameter configurations. This feature alone makes SyntheticNET simulator unique as existing simulators model mobility and HO using at most few parameters such as CIO and cell offset. On the other hand, SyntheticNET simulator models the mobility management to its utmost depth by incorporating dozens of parameters that affect the mobility related KPIs and other system level KPIs in an intricate and interdependent fashion.

- SyntheticNET simulator incorporates both cell-level parameters (e.g. A3 offset) and relation-level parameters (parameters affecting adjacent BSs on same frequency and inter-frequency e.g. CIO.)

- Database-aided edge computing support where KPIs known through simulation or by importing csv files can be used to test AI enabled proactive network features (e.g. proactive mobility management, proactive resource allocation, proactive load balancing).

This chapter describes the key features of SyntheticNET simulator. The rest of the chapter is organized as follows. Section 3.2 provides high-level overview and explains the execution flow of the simulator. Section 3.3 then presents the key salient features of SyntheticNET simulator that makes it distinct from existing simulators. These include adaptive numerology, realistic propagation modeling, detailed 3GPP compliant HO triggering and execution mechanism modeling, database-aided cloud computing and the support of realistic mobility pattern. In section 3.4, I present a use case where I show how SyntheticNET simulator features such realistic mobility pattern integration and HO procedure can aid in realistic evaluation of different mobility prediction techniques. Section 3.6 concludes the chapter.

3.2 Simulator Structure and Execution

One of the goals of SyntheticNET simulator is to act as a key enabler for AI based revolutionary planning and optimization solutions for 5G cellular networks and beyond. To have an overview of the structure of the presented simulator is thus necessary to understand the capabilities and usefulness of SyntheticNET simulator. In this section, I provide this high-level overview of the simulator without attempting to explain the functionality of each simulator module in detail.

3.2.1 Simulator Block Description

The overall structure of the SyntheticNET simulator is shown in Fig. 3.2. As shown in this figure, the SyntheticNET simulator can be divided into eight basic blocks. These blocks are briefly described below.

Network Deployment Unlike the existing OOP based simulators, SyntheticNET takes the input in the form of commonly used database format. such as the widely used csv format. This block imports the individual BS characteristics which include location, cell

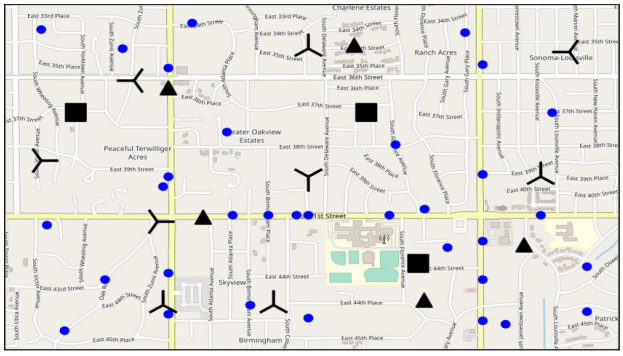

Figure 3.3: Sample heterogeneous network layout with sectorized BSs, omni-directional BSs (square), small cells (triangle) and UEs (dots).

id, type of BS, operating frequency, transmission power, electrical/mechanical tilt, azimuth angle, number of available antennas, Cell Individual Offset (CIO - herein referred to as cell bias) and even antenna pattern. The screen shot of a sample heterogeneous network deployment has been shown in Fig. 3.3.

Prorogation Modeling Module SyntheticNET allows incorporation of wide range of propagation models and associated data. Custom pathloss empirical models, measurement data based models or ray tracing based models can be used for realistic signal strength calculation. SyntheticNET also allows importing of the detailed topographic data in various industry compliant formats for more realistic pathloss modeling. One key feature of SyntheticNET is its ability to support newly emerging AI based propagation models such as [125].

User Association Module UE location can be imported to the simulator in a similar fashion as the network configuration file described above. This module's main responsibility is to compute the signal strength of each UE located in the defined network bound. UE is associated to the serving cell with the smallest distance to the UE, or to the BS having the highest Reference Signal Receive Power (RSRP). UE is associated with the cell only if the RSRP is higher than a certain threshold (defined as qRxLevMin in 3GPP [110]). The latter approach is recommended for heterogeneous BS scenario where transmission power of each BS can vary due to location, surrounding, type of BS

etc. While calculating the signal strength, height of BS and UE, and the angular separation between the UE and the respective BS azimuth angle is taken into consideration. A key unique feature of this module in SyntheticNET is that it allows testing of custom AI enabled more sophisticated user association criteria that can take into account advance KPIs such as cell current and future load, network energy consumption, mobility pattern, QoE requirements, caching requirements, caching on edge statistics, and UE battery.

User Mobility Location of mobile UEs are modelled by this block. UE location is updated based on the vselocity assigned for that individual mobile UE, and the direction is known from one of the selected model (Manhattan model, Random Waypoint, SLAW model etc.). In addition to predefined or historical UE path, SyntheticNET also supports realistic mobility pattern by integrating SUMO [126] in its mobility pattern modeling module (Section 3.3.4). This feature can help advance AI enable proactive and holistic mobility management solutions.

HO Procedure HO procedure module provides a realistic 3GPP-based HO criteria evaluation so that vital retainability KPIs like HO attempt and HO success rate can be evaluated. For HO procedure, configuration files which include the cell level and relation level parameters can be configured internally or can be imported to model a newly proposed or vendor specific HO implementation. More detail on this can be found in Section 3.3.2.

A unique feature of SyntheticNET in this context is that unlike most existing simulators that consider only one or two basic HO parameters thus offer inaccurate results on mobility related KPIs, the HO module in SyntheticNET incorporates all 20+3GPP defined configuration parameters that affect mobility in a real network. Modeling these parameters in a simulator is a key step to enable holistic AI enabled network automation. These parameters not only affect mobility related KPIs but also determine overall

signaling overhead, capacity, UE battery life and QoE.

Traffic Modelling This block first creates a resource map relative to each participating BS. Resource map constitutes of several Physical Resource Block (PRB) arranged in accordance with the μ parameter. μ defines the Sub Carrier Spacing (SCS) as per 3GPP release 15 [110]. Resource map size is dependent on the bandwidth of the central frequency BSs are operating on.

Next, the UEs served by the respective BSs are allocated resources according to selected scheduling scheme. The scheduling scheme can be custom or standard such as Round Robin, Proportional Fair, Max C/I etc. While allocating resources, priority criteria can also be defined. In default criteria, priority is given to UEs as per their QCI. Delay and jitter sensitive voice users are scheduled with the highest priority, followed by UEs corresponding to Vehicle to Everything (V2X) QCI. The remaining PRBs are allocated uniformly to FTP users. For each QCI class, UEs are allocated resources as per the scheduling approach described earlier.

Interference Measurement The SINR plays a vital role in determining the performance, quality and hence user experience in cellular networks. A large set of accessibility, performance, and retainability metrics, such as coverage, capacity, and mobility related KPIs are heavily dependent on SINR [107]. In most simulators reported in literature, for simplicity just RSRP based interference estimation is done. This abstraction introduces error that makes KPIs estimated by these simulators far different from the performance in a real network. To avoid this source of error, SyntheticNET uses the actual PRB level interference calculation. More detail on the PRB level SINR calculation can be found in Section 3.3.1.

Performance Evaluation User level and network level performance is evaluated in performance evaluation module. In this module, accessibility KPI can be estimated based

on the number of static and mobile users located in areas where RSRP or SINR is below the cell association threshold. The default cell association threshold is the RSRP or SINR level below which UE is unable to camp on the cellular network due to lower message decode % in either or both uplink\downlink direction. Uplink messages are not decoded properly due to higher pathloss (low RSRP) where UE transmission power is not adequate to maintain the desired signal quality at the BS. On the other hand, low downlink message decodes % is mainly due to high interference (low SINR). Retainability KPI can be computed in a similar manner as accessibility KPI, if during the connected mode or HO phase, RSRP or SINR remains below the Radio Link Failure (RLF) threshold for a certain duration. Configured HO parameters can thus be evaluated from retainability KPI. This feature can thus enable design of AI enabled algorithms to determine optimal HO parameters - an important use case for industry.

SyntheticNET supports a variety of data rate calculation methods to represent vendor specific implementations and deployment. In SyntheticNET simulator, default UE specific and cell level maximum throughput (Mbps) is computed by employing the 5G NR [110] max data rate (Γ) equation:

$$\Gamma = 10^{-6} \sum_{j=1}^{J} \left(v_{Layers}^{j} Q_m^j f^j R_{max} \frac{N_{PRB}^{BW^{j,\mu}} \cdot 12}{T_s^{\mu}} (1 - OH^j) \right), \tag{3.1}$$

where J is the number of component carriers, v_{Layers}^{j} is the maximum number of layers, Q_m is the maximum modulation order, f is the scaling factor, $R_{max} = 948/1024$, μ is the numerology which denote SCS as described earlier, T_s is the OFDM symbol duration, N_{PRB}^{BW} is the number of PRBs allocated to UE and OH is the overhead.

Another distinct feature of SyntheticNET simulator is to identify the silence period where voice users cannot communicate due to either uplink or downlink issues. Silent period is usually observed when UE experiences RLF or the RSRP and SINR drops below the silence threshold.

Accessibility, RLF and silence thresholds are typically dependent on associated network

parameters and to a certain extent the equipment vendor. SyntheticNET simulator supports the use of AI based techniques to identify the respective threshold for a current network deployment. Detail procedure to identify the respective threshold for a given network layout is beyond the scope of this chapter.

3.2.2 Simulator Execution Overview

Initial setup of SyntheticNET simulator requires setting up following items:

- Simulation duration.
- Transmission Time Interval (TTI) length - which is dependent on the 5G μ parameter.
- Network deployment - BS types, tilt, azimuth, power, scheduling scheme, operating frequency, bandwidth, number of tiers etc.
- Network-level parameter configuration - which may be unique for individual BSs e.g. HO parameters.
- Relation-level parameter configuration - which includes parameters affecting adjacent BSs on same frequency and different frequency as well, such as CIO.
- UE description - location, static/mobile, height, number of antennas etc.
- UE mobility features - Random waypoint, SLAW Model, Manhattan model, real traces etc.
- (optional) UE historical location - to help evaluate AI based advance mobility management for example proactive HO management, load balancing and energy efficiency.
- (optional) Database of historical KPIs and parameter value pairs for enabling AI based network automation.

Upon execution, SyntheticNET simulator starts processing the data through each module described earlier. In each TTI, network level KPIs and user experience KPIs are calculated. As the simulation proceeds, SyntheticNET simulator executes the HO to a better cell if the HO criteria are met. Next, the resource allocation takes place based on the selected resource scheduling scheme and then PRB-level interference measurement takes place for all scheduled UEs. When the number of TTI elapsed reach the simulation duration, an output file is generated which gives the UE level and network level KPIs. Moreover, a log file is generated which contains signal level from all the nearby BSs along with other network level statistics of interest. This log file provides additional insights that can be leveraged to propose better interference mitigation and mobility management solutions.

3.3 Detailed Feature Description

Elaborating each of the 5G specification [110] features incorporated in SyntheticNET simulator is not possible within the scope of this chapter. Therefore, I present four of the key features that make SyntheticNET simulator superior to existing 5G simulators[111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123]. These are the vital components of 5G standard, and are hence, essential to accurately and realistically simulate a 5G network. To the best of authors' knowledge, none of the existing 5G simulator incorporates all four network features described in following subsections.

3.3.1 NR Adaptive Numerology

In 5G, 3GPP provides adaptive numerology in order to accommodate diverse services (eMBB, mMTC, URLLC) and the associated user requirements. The key idea is to adapt the transmission configuration to address the stringent QoE constraints considering the effect of UE mobility and varying channel conditions.

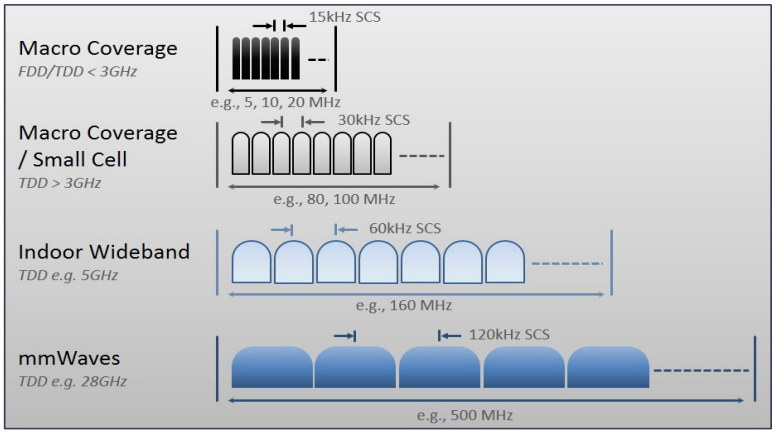

Figure 3.4: 5G NR adaptive numerology.

5G frame structure in SyntheticNET simulator supports adaptive numerology where the TTI duration and the number of PRBs per TTI vary in accordance with the flexible SCS. Structure of the 5G flexible frame and the SCS is governed by the μ parameter. When importing site info, the value of μ associated with each carrier frequency should be assigned so that PRB allocation and interference calculation takes place according to the respective frame structure.

3.3.2 NR Handover Criteria

User mobility has been the *raison d'etre* of wireless cellular systems. To maintain reliable connection, it is incumbent upon the mobile users to perform HO from serving cell to the next *suitable* cell along their trajectory. HO frequency is mainly dependent on the mobile user speed and network deployment characteristics (BS density, heterogeneity, HO parameter configuration etc.). 5G networks will have a large HO rate, primarily because of network densification and a large fraction of mobile UEs. 5G standard follows break-before-make HO approach similar to LTE where mobile user may observe HO failure due to poor signal strength of participating BSs, sub-optimal HO parameter configuration or high user velocity.

Therefore, apart from coverage and capacity, retainability is a vital KPI to measure user experience in 5G. For this reason, SyntheticNET simulator models the detailed

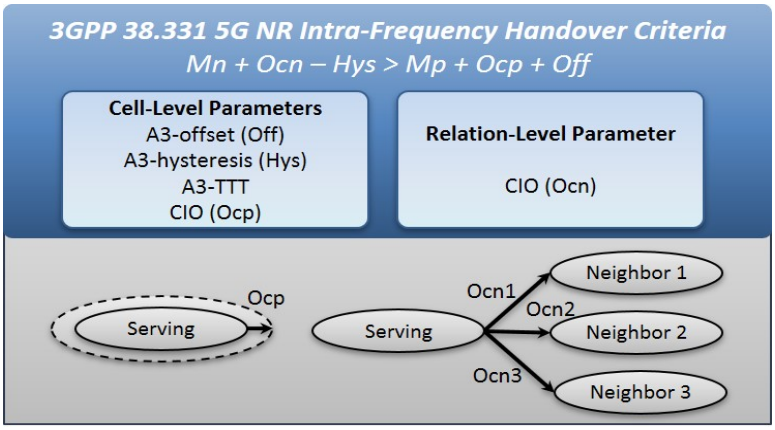

Figure 3.5: 5G intra-frequency HO parameters.

3GPP-based HO evaluation and execution process for mobile users. For each cell, intra-frequency Hand Over Margin (HOM) is calculated based on A3-offset, A3-hysteresis, serving cell CIO (Ocp or cell bias) and target cell CIO (Ocn). HO evaluation procedure initiates when RSRP of target cell exceeds the RSRP of serving cell by HOM. Next, SyntheticNET simulator's mobility block ensures HOM condition is fulfilled for each (Transmission Time Interval) TTI up till when the Time-To-Trigger (TTT) timer is expired. This is followed by a HO execution from serving to target cell and during this procedure, serving RSRP and SINR are recorded to help realistically quantify user throughput and retainability KPI for evaluating QoE metric. For more details of 3GPP defined HO execution mechanism, as implemented in SyntheticNET, see [109] and Fig. 1.7 therein.

HO parameter configuration files corresponding to each cell in the network are imported to SyntheticNET simulator as discussed in Section 3.2. For HOM calculation, two types of configuration files are needed: a) cell-level HO parameter list, and b) relation-level parameter list. Respective parameters needed for intra-frequency HO criteria evaluation are shown in Fig. 3.5. A more detailed diagram detailing all 28 mobility related parameters and their associations with all 8 mobility related KPIs dictated by these parameters is given in Fig. 1.7, in [109]. SyntheticNET simulator also supports 3GPP based inter-frequency HO. Description of inter-frequency HO has been omitted in this chapter and can be found in [109].

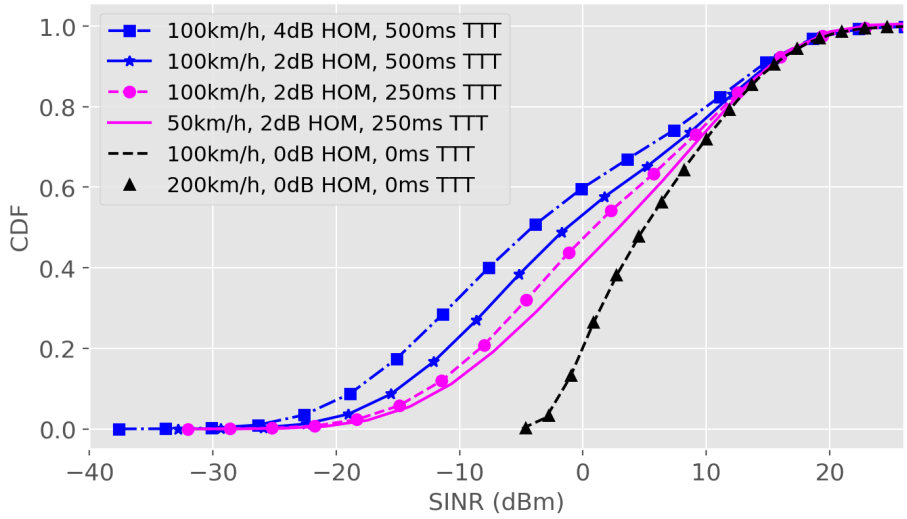

Figure 3.6: SINR CDF of a mobile user traveling across the network layout (Fig. 3.3).

Fig. 3.6 shows the SINR CDF of a mobile user traversing through the network layout shown in Fig. 3.3. During HO criteria evaluation (Fig. 3.5) i.e., during the time needed to execute HO, mobile user penetrates through the coverage of the neighboring cell without performing HO. As a result, UE observes temporal negative SINR (on dB scale) due to strong interference from the best server (HO target cell). The magnitude of negative SINR during HO phase increases with user velocity as user penetrate deeper into the coverage of neighboring cell. Similarly, larger HOM and/or TTT may contribute to more severe SINR dilapidation. This is illustrated in Fig. 3.6 for different user velocity and HO parameter configuration.

For example, Fig. 3.6 illustrate that for 0dB HOM and 0ms TTT, UE always stays on the best server while interference is observed from non top-1 cells. Consequently, UE observes positive SINR (dB) most of the time. However, there are instances where UE observes negative SINR due to strong interference from multiple non top-1 cells. Since UE always stays on the top-1 cell and HO delay due to 3GPP HO criteria (Fig. 3.5) is not observed, the effect of user velocity on SINR distribution is negligible. This can be verified in Fig. 3.6 where UE velocity is changed from 100km/h to 200km/h, but the SINR CDF remains unchanged.

Fig. 3.6 also shows SINR distribution plot for various HO configuration parameters. UE SINR decreases with more stringent HO criteria. This is in line with the temporal

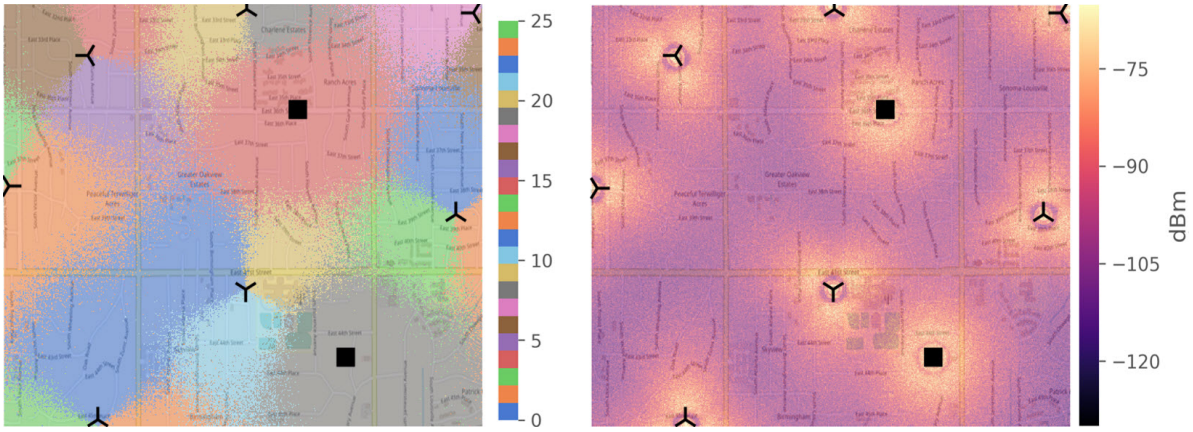

Figure 3.7: Network area binning (5x5m bins) based on Top-1 Physical Cell Identifier (PCI) and RSRP with 5dB shadowing standard deviation and realistic antenna patterns.

SINR degradation during HO criteria evaluation discussed earlier. There is a trade-off between ping-pong HOs and HO delay duration. Because of shadowing, ping-pong HOs increase dramatically when HO configuration demands UE to stay on best server or when UE HO criteria is easily fulfilled. Conversely, ping-pong HO reduces for tighter HO condition, but HO delay increases causing negative SINR or sometimes Radio Link Failure (RLF) especially for high speed users. More detail on this can be found in [109].

It is worth highlighting that most existing simulators do not model HO procedures and associated configuration parameters in such detail to capture aforementioned and other mobility related important phenomena and the associated impact on overall throughput and user QoE that is inevitably experienced in real network.

3.3.3 Futuristic Database Aided Edge Computing

SyntheticNET simulator also supports database aided edge computing approaches deemed essential for futuristic mobile networks [127]. SyntheticNET simulator divides the target area into a custom bin map whose size can be user defined (see Fig. 3.7). For a given network layout, SyntheticNET simulator quantifies several KPIs which include SINR distribution, Channel Quality Indicator (CQI) distribution, spectral efficiency, available resources, VoLTE Silence%, HO rate, HO failure% etc.

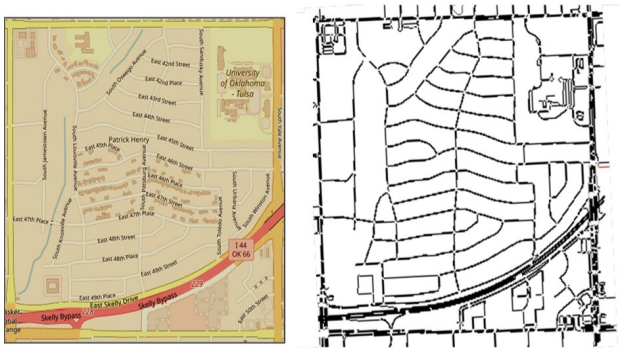

Figure 3.8: Realistic Road Map from SUMO.

SyntheticNET simulator then allows to build and store historical database of the above KPIs and selected measurements for the bands of interest such as RSRP, SINR, CQI, PRB usage, mobility traces, QoE indicators such as RLF reports etc. Using tools from machine learning and stochastic optimization, this database can be then leveraged to design algorithms for Data Base Station (DBS) aided cell discovery and selection, proactive radio resource allocation and switching ON/OFF DBS proactively instead of reactively, to jointly maximize both spectral efficiency and energy efficiency without compromising QoE.

In addition to highlighting the areas with poor coverage or high interference, the database of a list of key KPIs can be utilized to propose and evaluate novel SON and AI based network automation features. For example, by feeding the historical UE location, we can predict user location and can proactively perform inter-frequency HO to avoid low retainability KPI. Similar approaches can be designed and tested to achieve better load balancing in a multi-tier heterogeneous network.

3.3.4 Realistic Mobility Pattern

Existing 4G or even 5G simulators [111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123] are limited to much simpler and non-realistic mobility models like random waypoint, SLAW model, Manhattan model etc. SyntheticNET simulator on the other hand, incorporates realistic mobility pattern by integrating the Simulation of Urban

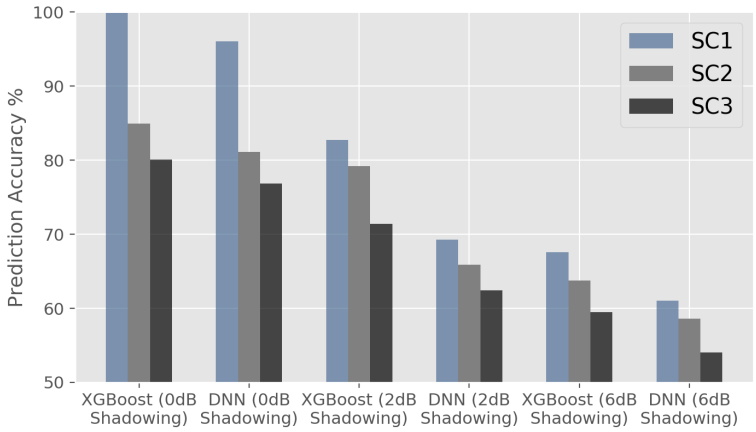

Figure 3.9: Performance of AI-assisted mobility prediction techniques in HetNets.

MObility (SUMO) [126]. SUMO is an open source, highly portable, microscopic and continuous road traffic simulation package designed to handle large road networks. It allows for more realistic simulation including pedestrians and comes with a large set of tools for scenario creation.

SUMO can help simulate a given traffic demand where the network scenario consists of individual vehicles moving through a given road network. Each vehicle can be modeled explicitly, has an own route, and moves individually through the network. Mobility patterns in SUMO are deterministic by default but there are various options for introducing randomness. Randomness can be added for certain aspects of test case scenario which include speed distribution, departure times, number of vehicles, vehicle type, route distribution etc. SUMO also supports traffic stops, departure speed, arrival speed, intersections, yield lane with low priority etc.

Thus, SUMO empowered extremely realistic mobility modelling capability of SyntheticNET that can also incorporate realistic road maps and mobility traces, makes SyntheticNET simulator first of its kind 5G simulator capable to investigate a large set of mobility management and optimization problems.

3.4 A Case Study Using SyntheticNET: AI-Assisted Mobility Prediction for HetNets

In this section, I give one example utility of SyntheticNET simulator through a case study that is not possible with simulators that do not realistically model mobility patterns and mobility management and HO procedures in the network.

This case study briefly shows how we can achieve AI-assisted mobility prediction of mobile subscribers through realistic traffic modeling obtained from SUMO. User Mobility Prediction can be one of the key enablers for AI based network automation and next generation proactive SON [128]. This can enable the reservation of network resources in future identified cells for seamless HO experience [104] as well as for traffic forecasting purposes for load balancing [129] and driving the energy saving SON functions [130, 105] as well as optimizing battery life [103].

In the first stage, I setup the network deployment by importing the site-info having the location of several macro and small cells, along with other associated parameters (power, height, tilt, azimuth etc.). Then I feed the realistic user mobility traces taken from SUMO into the SyntheticNET simulator. To get the required user mobility data from SUMO, SyntheticNET first passes the network file and population definition file to SUMO. The network file describes roads and intersections where the simulated vehicles move during the simulation. The population definition file has a general statistical information which includes the number of households, locations of houses, schools and workplaces, free time activity rate, etc. Mobile users by default travel from home to workplace and vice versa. However, additional trips where users visit the entertainment area or grocery shop as per a defined percentage are configured as well. The additional trips are considered as a proxy for increasing randomness in user trajectories. Moreover, randomness in the daily user routes between home and workplace is configured as well.

SUMO then performs the simulation on the input data from SyntheticNET and generates realistic mobility pattern for the mobile users over the configured time interval.

The realistic mobility traces generated by the SUMO are then used for mobile users in the user mobility module of the SyntheticNET simulator. During the trajectory, users perform HO as they move across cells. SyntheticNET simulator also keeps track of user location and serving cell id to be used as an input to AI enabled solutions for mobility prediction purposes.

In the exemplary scenario, I run the simulation for 10 days over the map of city of Tulsa obtained from open source map (see Fig. 3.8). Several macro BSs and small cells are deployed in the test area. After assigning the home, workplace and entertainment locations, I obtain the realistic mobility traces with different degrees of randomness in user paths. Scenario 1 (SC1) represents zero randomness between user trajectories, whereas in medium randomness scenario 2 (SC2), user makes equal number of random trips to any entertainment location as between home and workplace. For high randomness scenario (SC3), in addition to randomness in the user trips as defined in SC2, users follow different routes between home and office 50% of the time.

Finally, I run eXtreme Gradient Boosting trees (XGBoost) and Deep Neural Network (DNN) multi-class classification algorithms on the data obtained from the simulation scenarios described above. Data is split into 70% training data and the remaining 30% into test data. Results of the AI-assisted mobility prediction techniques on test data for different shadowing standard deviation of RSRP can be found in Fig. 3.9. Fig. 3.9 shows that prediction accuracy of more than 90% can be achieved for certain scenarios. However, the prediction accuracy decreases as the randomness in the training data increase.

Mobility prediction will be one of the key enablers AI enabled network automation including next generation proactive SON solutions that aims at efficient resource management of emerging cellular networks. SyntheticNET simulator can help the research community implement their novel research ideas on a realistic 3GPP complaint network simulator and thus can have a better proposal evaluation. Detail description of the

algorithm is out of scope for this chapter and has therefore been exempted here. For detailed description, see [131].

3.5 Future Work

The viability, usefulness and uniqueness of SyntheticNET can only be ensured by putting continuous efforts in the development phase of SyntheticNET. In the following, I have identified few of the key features I will incorporate in SyntheticNET:

1. Matrix Pre-Calculation: Existing mobile network simulators take significant amount of time when executed for a realistic network deployment with considerably large number of both BSs and UEs. SyntheticNet will use a novel approach to lower this simulation processing time by pre-generating matrices of signal strength and signal quality across the deployed network while still maintaining the effect of shadowing. In a similar manner, UE mobility patterns will also be pre-generated. This approach can reduce simulation interval by avoiding the calculation of the UE location and UE association related signal indicator values every TTI.

2. Support of Multiple Frequency: Futuristic mobile networks heavily rely on the deployment of several layers of frequency to meet the capacity crunch. By employing, SyntheticNET should be able to support multiple frequency deployment and have the capability of simulating events involving different layers such as inter-frequency handover.

3. Measurement Gap and Inter frequency Handover: Most of the currently available simulators only supports intra-frequency handovers or hand-over between base stations with similar frequency due to its convenience in terms of implementation. Due to its additional complexity such as modeling measurement gap, incorporation of inter-frequency handover or handover between different frequencies is mostly taken for granted. With SyntheticNet, I understand the importance of

inter-frequency handover in terms of its effect on the network performance. I therefore find it as a must to include inter-frequency modelling and implementation to SyntheticNet.

4. Radio Link Failures: One of the most common problem experienced by mobile users is Radio Link Failure (RLF). RLF affects some of the core network KPI's such as retainability and throughput. Realistic implementation of RLF will be beneficial in learning why this event happens and in finding ways how it can be avoided. Thus, is it essential to capture and incorporate RLF event to SyntheticNet.

5. Support of Complete List of Mobility Events (A1, A2, A3, A4, A5): 3GPP-standardized mobility events used in LTE will still be utilized in 5G. As a simulator that supports legacy and futuristic network, this necessitates intricate modeling and implementation of the most commonly used HO parameters to SyntheticNet. This will ease experimentation to learn how KPI behaves with changes on these parameters.

6. Load Balancing Algorithms: Though heavily researched, load balancing is still a challenge in today's cellular network. Since 3GPP left the load balancing algorithm for innovative purposes, SyntheticNet will model the approach being used by major telecom vendors. Moreover, I will develop new innovative load balancing features and will evaluate the efficacy of the developed algorithms by comparison with the load balancing algorithms currently employed by major telecom vendors.

7. Idle Mode Mobility: Modeling of idle mode users is frequently left untouched in most simulators available. However, it is essential to model these users in order to realistically capture the network dynamics. Even though idle mode users don't transmit any data, they do use signaling which affects the network specially in the uplink direction. With that in mind, SyntheticNet will include modeling of

idle mode UEs to capture the key KPIs like signaling, battery consumption and accessibility.

3.6 Conclusion

The importance of a realistic yet practical simulator adhering to 3GPP standard for cellular networks can be mirrored by the expected complexity of 5G and beyond networks. However, simulators which are currently available are bounded by too much simplifications, unrealistic assumptions and are lacking in implementation of vital network features making them insufficient in capturing the complexity and dynamics of a real cellular network. To address these challenges, I have developed the first 3GPP 5G standard (Release 15) compliant network simulator called SyntheticNET simulator. SyntheticNET provides a more realistic and practical evaluation of different network scenarios as well as implementation of several key network features.

Unlike existing OOP based simulators where BS locations depend on an underlying distribution and cannot be preassigned, SyntheticNET simulator is microscopic where individual elements (BS and UE) of the network can have unique and hard coded parameters (azimuth, tilt, antenna pattern, height, transmission power etc.) which is the case in an actual network deployment. With the modular approach of SyntheticNET simulator, it is effortlessly possible to further extend the already implemented network functionalities with 3GPP release 16 and upcoming updates making this simulator future proof. With the flexible implementation of SyntheticNET simulator, it is possible to simulate large-scale networks with several thousand active heterogeneous BSs and several user types, without the need for specialized simulation hardware.

SyntheticNET simulator is the first and only simulator built to date which model more than 20 parameters essential to implement a detailed 3GPP-based HO process. With the added support of realistic user mobility traces, vital mobility KPIs like retainability and HO success rate can be precisely evaluated. In addition to mobility, other key com-

ponents of SyntheticNET simulator includes ray tracing-based models to give accurate signal strength calculation, and adaptive frame structure to help meet several 5G use cases (eMBB, URLLC, mMTC) requirements.

SyntheticNET simulator is the first Python based simulator with inherent ease to process, manipulate and analyze large data sets. Similarly, it has easy access to wide range of machine learning algorithms. This makes SyntheticNET simulator relatively easier to implement and evaluate AI based solutions for autonomous configuration and optimization of network parameters in a given multi-tier heterogeneous network deployment making it beneficial for research community and industry alike.

The presented use case on mobility prediction showcased the power of SyntheticNET in providing practical network deployment, hand over procedure and ease of incorporating realistic mobility patterns from other sources to provide a realistic evaluation of several machine learning techniques in predicting user mobility which would have been impossible or inaccurate using the currently available network simulators.

CHAPTER 4

QoE-Aware Smart EN-DC Activation Using Artificial Intelligence

4.1 Introduction

5G New Radio (NR), with innovative use cases of enhanced Mobile Broadband (eMBB) for large volume transmissions, massive Machine Type Communications (mMTC) for sensors and IoT devices, and Ultra Reliable Low Latency Communications (URLLC) for self-driven vehicles comes with unprecedented Quality of Experience (QoE) goals. Studies project that 5G subscriptions will top 2.6 billion by the end of 2025 [132]. While in 5G, the capacity crunch will be addressed primarily by ultra-dense Base Station (BS) deployment and mmWave band utilization [107], ensuring QoE with a conglomeration of new and legacy technologies remains an open challenge of utmost importance.

As per 3GPP Release 15 specification 37.863 [133], E-UTRAN New Radio Dual Connectivity (EN-DC) allows 5G capable User Equipments (UEs) to simultaneously connect to an LTE eNodeB (eNB) that acts as a master node and a 5G gNodeB (gNB) acting as a secondary node. This non-standalone 5G network deployment will meet the UE capacity demands, help mobile operators to reduce the CAPital EXpenditure (CAPEX), and will accelerate the penetration of 5G networks. However, the added complexity involves primarily the signaling overhead, and the decision when to activate/deactivate EN-DC mode.

EN-DC activation comes with an intrinsic trade-off between 5G network utilization and potential QoE degradation due to RLF and voice call muting. To the best of authors' knowledge, there does not exist a framework to quantify and optimize this trade-off. While the goal is to effectively activate EN-DC to benefit from the 5G features, sub-optimal configuration can lead to excessive amount of ping-pong EN-DC activa-

tion/deactivation, recurrent Radio Link Failures (RLFs), and exasperating voice call muting. RLF is the radio interface disruption between Base Station (BS) and UE, and is typically caused by coverage hole or poor signal quality as a result of high interference. UE observes high interference either during handover (HO) process due to sub-optimal HO parameter configuration, or due to the interference from neighboring cells usually at the cell-edge. On the other hand, voice call muting refers to the instances where the either of the UEs from the call originating or call terminating side are unable to receive audio packets. This situation is observed mostly due to poor radio conditions at either of the voice call participating UEs.

By accelerating the EN-DC activation in an attempt to increase network efficiency, EN-DC may be triggered at poor Radio Frequency (RF) conditions at either 4G or 5G network. This can result in call disconnect, and service interruption. Following the service disruption, repeated re-accessibility attempts not only increase signaling but degrade UE energy efficiency as well. Thus, optimal configuration to activate/deactivate EN-DC is essential to maintain the expected QoE and network efficiency of 5G network.

4.1.1 Related Work

The concept of dual-connectivity has been studied extensively in literature [134, 135, 136]. A detailed review of these studies can be found in a recent survey on the topic of mobility management in emerging networks [109]. More specifically, the analysis of dual-connectivity gain in terms of delay and throughput [137], mobility [138], energy efficiency [139], reliability [140], and low latency [141] exists in literature as well. However, to the best of authors' knowledge, no study in existing literature addresses the QoE-aware criteria to activate dual-connectivity between two different mobile technologies viz a viz 4G and 5G. Particularly, RLF and muting instances in the context of dual-connectivity, have not been studied at all.

Most of the RLF related literature [142, 143, 144, 145] addresses intra-frequency HO

issues by controlling the system common parameters. For example, in [142], time-to-trigger (TTT) and HO margin are adjusted based on the type of RLF observed during HO. Similarly, [143] considers tuning another know parameter called A3-offset to prevent RLF between intra-frequency neighbors. Authors in [144] categorize HO failure into too early, too late and wrong cell HO to adjust TTT and A3-offset accordingly. Apart from optimizing intra-frequency HO parameters, authors in [145] propose transmission power changes to adjust coverage holes in an attempt to avoid RLF. RLF detection approach in [146] uses an RF threshold to detect possible RLF situation and accelerates HO to a better cell if available. However, the mechanism of setting appropriate RF threshold, is not defined.

On the other hand, voice call muting (specifically IP based Voice over LTE 'VoLTE' muting) is rarely studied by the research community, and the primary reason for that is bi-fold; a) the low penetration rate of VoLTE calls - most subscribers are redirected to circuit switch based 3G networks when making voice calls (this is due to incapability of mobile handset, inability of BS, or reluctance of the network operators to enforce VoLTE calls), b) the voice muting prevention techniques are normally based on traditional optimization methods (coverage hole avoidance, SINR improvement, seamless and timely handover, resource availability). Research community up till now assumes those traditional optimization techniques can suffice to avoid call muting. However, although the RLF avoidance approaches discussed above [142, 143, 144, 145, 146] may help minimize voice call muting as well, the optimization techniques aimed specifically at voice muting prevention, need to meet more stringent criteria than traditional approaches. This is because unlike traditional HTTP/FTP traffic, voice call requires real-time low-latency packet transfer for high definition, jitter-free communication. For the same reason, in an attempt to camp the voice call UE on the best available frequency, network operators use a different set of mobility parameters compared to when an ordinary data call is active.

In the context of voice call muting, only the study of mobility (HO between WiFi access points and not cellular tower) [147] and resource scheduling [148] exists in literature. However, none of the existing studies aims to investigate a scheme for a QoE-aware dual-connectivity (EN-DC) establishment. Furthermore, as concluded earlier, most of the RLF prevention approaches proposed in literature target intra-frequency HO optimization and do not identify actual measurement thresholds to detect possible RLF. Therefore, there is dire need for a framework to detect potential RLF threshold and potential muting threshold (signal strength and quality), and utilize that information to configure optimal inter-RAT (Random Access Technology) parameters for resource efficient and QOE aware EN-DC activation.

4.1.2 Contribution

This chapter presents the first framework whereby leveraging real network data measurements, I have quantified and optimized the tradeoff between 5G network utilization and QoE degradation due to potential RLF or potential muting. This work is an extension to our work in [149] where I have only presented a preliminary study on potential RLF identification. Potential RLF and potential muting refers to the UE RF condition where actual RLF or muting may not be observed, however, the UE is under the RF environment that can ultimately lead to actual RLF or muting (e.g. through the expiry of relevant timers and counters). I first obtained the potential RLF thresholds by taking into account the 3GPP [150] based low level measurements ($N310$, $T310$, $maxRACHattempts$, $maxRLCretransmissions$) from the real network.

I first gather RLF and call muting related data from the real network and investigate several approaches to address the data imbalance issue, which include Random over sampling, SMOTE, NearMiss, CNN, Tomek Links, ENN, NCL and GAN. More detail on these algorithms can be found in section 4.3. I then design, develop and evaluate various machine learning algorithms including regression, KNN, SVM, Naive Bayes,

XGBoost and deep learning algorithms, each with a range of hyper parameters. Results show that the best performance for the potential RLF identification model is obtained when data augmentation is performed using Tomek Links, and the enriched data is trained using deep learning technique. The trained deep learning based potential RLF model gives 25% more accuracy than the raw data where data augmentation is not performed.

UEs with voice service requirements are typically configured with a separate set of parameters to ensure successful transmission and reception of the real-time voice packets. Keeping in view the stringent requirements of voice calls, I train, develop and evaluate a two stage AI model, where potential RLF model with stricter potential RLF identification criteria is used as a first stage model. I use data from the actual voice call in the training phase of this two-stage model. Similar as for potential RLF model, I evaluate the results of the second stage AI model using several machine learning techniques. The best accuracy is obtained when Generative Adversarial Networks (GAN) enriched data is used to train the deep learning model. With GAN, I am able to increase the F1 score from 0.45 to 0.87.

I then formulate two different QoE aware activation of EN-DC problems solving which can allow operators to fine tune the trade-off between maximizing 5G utilization and the risk of RLF and call muting respectively. I establish the non-convexity of the formulated problems and leverage Genetic Algorithm (GA) to obtain the optimal EN-DC activation criteria. Results show that GA can yield near optimal solutions in just 335 and 969 iterations for non-voice and voice service UEs, compared to 741,321 iterations needed with brute force. The overarching contribution of the chapter is a data driven optimization framework (Fig. 4.2) to optimally activate EN-DC while taking into account QoE by adjusting the 4G and 5G RSRP and SINR thresholds. Our results show that for equal weight (w=0.5) between EN-DC activations and RLF, the proposed framework can reduce RLF 1328 to just 229 while reducing number of EN-DC activations from

6125 to 4051. Similarly, number of mute instances can be decreased from 3208 to just 360 while reducing number of EN-DC activations from 6125 to 3571.

The contributions of the chapter are summarized below:

- First of its kind QoE-aware dual connectivity criteria using real network data.
- Leverage deep domain knowledge to identify
  - potential RLF occasions using low-level counters specified by 3GPP.
  - potential muting instances based on stringent potential RLF criteria.
- Two stage AI model trained on domain knowledge-based problem identification, that identifies potential RLF/mute for a given RF condition in a 4G or 5G network.
- Novel work that utilize real world VoLTE call data to build a classification-based AI-model.
- Using unconventional techniques such as Tomek links and GAN to address the class imbalance.
- Simple yet effective objective function where operators can assign weightage to control the rate of EN-DC activations, and RLF/muting avoidance.

The rest of the chapter is organized as follows. In Section 4.2, I briefly describe the 3GPP based a) EN-DC activation procedure, b) RLF trigger conditions, and c) the support of voice calls over cellular networks. Real LTE network measurement data collection, exploration and development of AI model to predict potential RLF and potential muting is presented in Section 4.3 and Section 4.4 respectively. Optimization problem formulation for an efficient RLF/mute-aware EN-DC activation criteria can be found in Section 4.5. Simulation results in Section 4.6.1 shows how Minimization of Drive Test (MDT) data can be used to determine suitable EN-DC activation configuration parameters while minimizing chances of RLF by making use of the AI based RLF prediction model developed in Section 4.3. Finally, I conclude the chapter in Section 4.7.

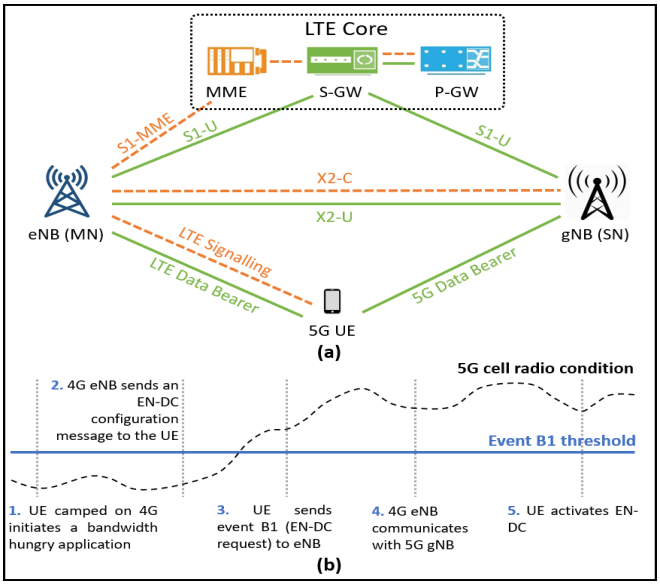

Figure 4.1: (a) EN-DC signaling and data connections, (b) EN-DC activation process.

4.2 Background

In this section, I briefly describe the 3GPP standard based procedures of EN-DC activation, RLF trigger criteria, and the support of voice calls over cellular networks.

4.2.1 EN-DC in 3GPP Release 15

A major focus of 3GPP Release 15 [133] is to get a first incarnation of 5G NR into the field that complements 4G LTE. Primarily, due to the higher frequency bands standardized in 5G networks, it is deemed better to enable UEs to connect simultaneously to 4G and 5G NR. This is referred to as Dual Connectivity option 3X or EN-DC. UE traditionally camps on 4G eNB, referred to as Master Node (MN) in EN-DC terminology. Later on, the network may configure EN-DC if the UE initiates the services that can benefit from EN-DC. Fig. 4.1(a) illustrates EN-DC signaling and data connections.

EN-DC activation process starts by the MN sending the EN-DC configuration (having the 5G frequency information and event B1 measurement criteria) to the UE. Event B1 configuration defines the 5G RF threshold which UEs are required to meet before initiating EN-DC request. EN-DC capable UE sends event B1 to the MN if as per the

configuration, the Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ) or Signal to Interference and Noise Ratio (SINR) of the 5G cell becomes better than the B1-threshold (see Fig. 4.1(b)). Mathematical form of event B1 fulfilment is shown in (4.1) where Mn is the measurement result of the inter-RAT neighbor cell, not taking into account any offsets, Hys is the hysteresis parameter, Ofn and Ocn are optional frequency and cell offset parameters respectively.

$$Mn + Ofn + Ocn^{\smile} Hys > B1 - threshold \tag{4.1}$$

As per the 3GPP standard, the UE has to transmit event B1 measurement report to the MN if it measures 5G radio condition to be higher than the configured event B1 threshold. The (4.1) shows that the 3GPP standard requires event B1 configuration to be either RSRP or SINR based, and not both. This can lead to a situation where 5G RSRP is above the configured B1 threshold, however, 5G SINR is poor due to excessive interference from the neighboring cells. Since the B1 measurement report encapsulates RSRP and SINR of both the serving 4G cell (MN), and the candidate 5G cell, in this work, I extend this process by applying another filtration inside the BS. By having another condition set inside the MN, we can ensure that the RSRP and SINR of both 4G and 5G networks is above an optimal threshold. This ensures that the QoE is not degraded due to RLF or voice muting after EN-DC activation. This is vital because EN-DC requires signaling data to be transmitted through the LTE BS. Hence, while good 5G radio condition is essential to deliver high quality service, having an additional check on both 4G and 5G radio condition before the EN-DC activation ensures that LTE connection to the UE can reliably serve the UE well. More detail on this can be found in Section 4.5.

Upon B1 reception, MN communicates with the 5G gNB and EN-DC is activated after the admission control check, and capability enquiry. 5G gNB upon EN-DC activation is referred to as Secondary Node (SN).

4.2.2 Radio Link Failure in 3GPP

The event where the UE abnormally detaches its connection with the serving cell is known as Radio Link Failure, commonly abbreviated as RLF. RLF procedure in 5G networks is same as in 4G, and is observed when either of the following three conditions are met continuously for a certain period. Each of these RLF condition is controlled by one or more parameters.

- Upon timer $T310$ expiry after configured consecutive out-of-sync indication (n_1) represented by $N310$ parameter.

  where t_1 represents the timer activated when n_1 equals or exceeds the value of configured parameter $N310$. Algorithm 1 gives a detailed explanation of this RLF criteria.

- After the configured number of consecutive unsuccessful RACH attempts (n_2) have been reached, as explained in the algorithm 2 below.

- When the number of consecutive RLC retransmissions represented by n_3 equals the value of the parameter $maxRLCretransmissions$. See following algorithm for details.

A network operator may prevent RLF by configuring higher thresholds mentioned above. However, in that case, UE would remain stuck in the poor RF condition. Though RLF causes service disruption, it gives the UE under poor RF condition a chance to reset its struggling connection, and UE can camp on the cell offering a better coverage. Optimization of these RLF related parameters to minimize the RLF are beyond the scope of this chapter and can be the subject of a future study. Here I am interested in developing a model that can predict the RLF. Such model will be then used for an intelligent EN-DC activation decision.

Algorithm 1: (RLF Criteria 1)

```
Initialize local counters n_1, n'_1 and timer t_1 to zero;
while UE in connected mode do
    for Every wireless frame do
        if UE is out of sync then
            Increment n_1 by 1;
            if n_1 ≥ N310 then
                ;                                          // Start timer if n_1 = N310
                Increment t_1 by 1;
                if timer t_1 equals T310 then
                    ;                                      // timer expires
                    Execute RLF;
                else
                    Do nothing;
                end
            else
                Do nothing;
            end
        else
            ;                                              // UE is in sync
            if timer t_1 has not started then
                Reset n_1 to zero;
            else
                Increment n'_1 by one;
                if n'_1 equals N311 then
                    Reset n_1, n'_1, t_1 to zero;
                else
                    Do nothing;
                end
            end
        end
    end
end
```

Algorithm 2: (RLF Criteria 2)

```
Initialize local counter n_2 to zero;
while UE in idle or connected mode do
    for Every RACH attempt do
        if RACH fails and timer t_1 not running then
            Increment n_2 by 1;
            if n_2 equals maxRACHattempts then
                ;        // Maximum Number of consecutive RACH fail attempts reached
                Execute RLF;
            else
                Do nothing;
            end
        else
            ;                                                    // Upon RACH success
            Reset n_2 to zero
        end
    end
end
```

Algorithm 3: (RLF Criteria 3)

```
Initialize local counter n_3 to zero;
while UE in connected mode do
    for Every instance of downlink data schedule do
        if RLC retransmission is required then
            Increment n_3 by 1;
            if n_3 equals maxRLCRetransmission then
                ;          // Maximum allowable consecutive retransmissions reached
                Execute RLF;
            else
                Do nothing;
            end
        else
            ;                                                // Upon successful decoding
            Reset n_3 to zero
        end
    end
end
```

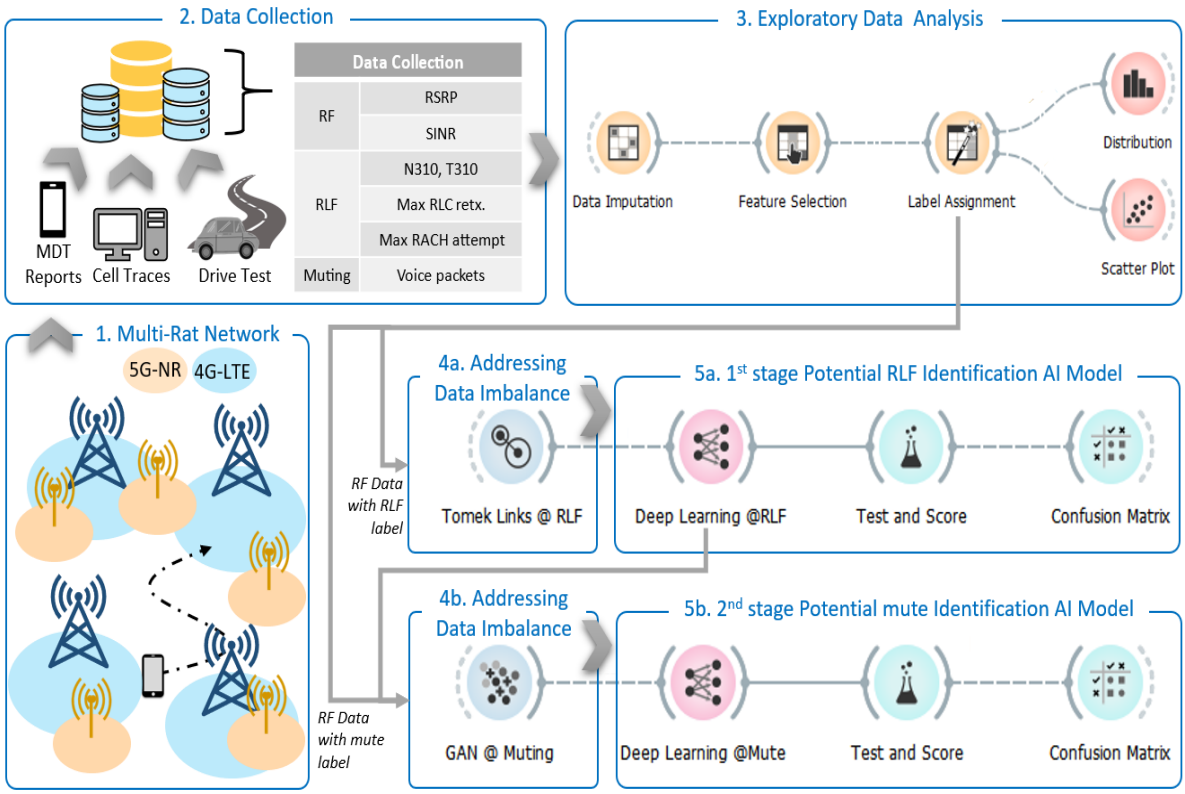

Figure 4.2: High level overview of the proposed AI-enabled EN-DC activation FRAMEWORK.

4.2.3 Voice Over Cellular Networks

Voice telephony is the primary reason why the cellular network came into existence. The legacy 4G networks and the latest 5G NR networks support voice services through VoLTE and Voice over NR (VoNR) respectively. The prerequisite of both VoLTE and VoNR call establishment are the UE capabilities and network configuration to support voice services. 3GPP [151] standardises QoS Class Identifier (QCI) value of 5 for voice call signaling. As a result, voice capable UE is configured with a QCI 5 bearer as soon as it comes under the coverage of the BS providing voice services. However, the voice packets themselves are sent through QCI 1 which is established only during the duration of an active call session. Session Initiation Protocol (SIP) is used for control signaling, while Real-time Transport Protocol (RTP) is used for the delivery of voice packets. Resource scheduling for a voice activated user is achieved through Semi-Persistent Scheduling (commonly known as SPS). SPS allocates with high priority

a fixed number of resources in a periodical manner and at predefined location within the bandwidth. This is done to minimize the scenario where a UE with an active voice call is being starved of the shared resources due to resource congestion.

The packet switch-based VoLTE and VoNR delivers high definition voice with much lesser jitter and delay than traditional circuit switch networks. However, the voice call is susceptible to muting under poor RF conditions. Due to the drop or loss of voice packets, a UE under poor radio condition cannot hear or transmit the audio to the call participant. Under worst circumstances, the voice muting can extend even for seconds, and this can be detrimental to user experience.

Unlike RLF which has underlying counters that dictate the network operations when to trigger RLF, voice muting is not dependent on any underlying parameters. This is due to the real-time flow of packets between the two participants, and call muting can be observed almost instantly whenever the signal strength or quality deteriorates. Nevertheless, voice muting is observed whenever the UE observes RLF or when the UE RSRP or SINR degrades. Thus, we can expect call muting to happen whenever actual RLF, or the condition that can lead to an actual RLF (referred herein as potential RLF) is observed. Note that the real-time nature of voice packets calls for a stringent condition to classify the radio condition under which a UE can observe call muting. This requirement has been taken into consideration in the two-stage deep learning model presented in Section 4.4, where I develop an AI model to predict voice call muting using actual muting data from a live network.

4.3 AI Model for RLF Prediction to Enable Smart EN-DC Activation

This section describes how actual measurement data from a real 4G network are collected to develop a deep learning based AI-model to help identify the set of RSRP and SINR conditions that correspond to potential RLF. This model is then used to design criteria to activate EN-DC mode after the MN receives the RSRP and SINR of the 4G and

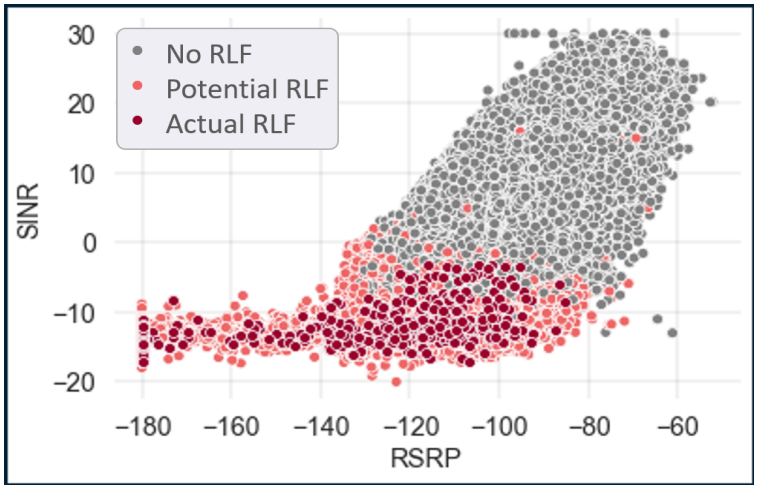

Figure 4.3: Potential RLF occurrences versus the UE RSRP and SINR measurements.

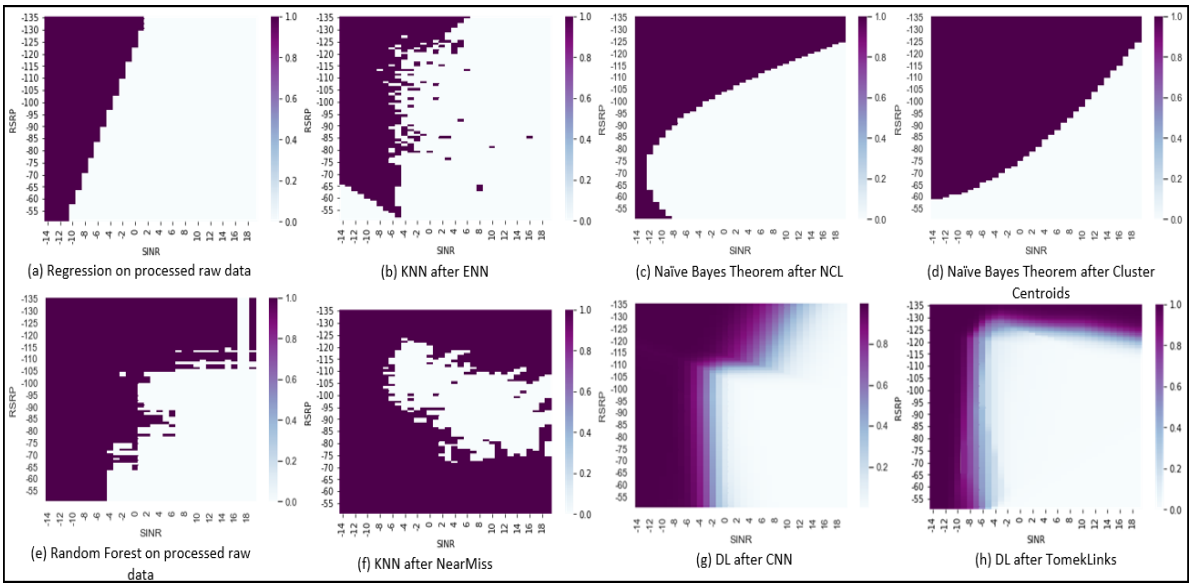

Figure 4.4: Decision boundary of THE potential RLF models shown in Table 4.1

5G cell inside the B1 measurement report from the EN-DC capable UE. Since the RLF criteria in 4G and 5G NR is same, the proposed RLF prediction AI model is applicable for both 4G and 5G NR [133]. Fig. 4.2 illustrates the high-level overview of the proposed AI-powered EN-DC activation method.

4.3.1 Data Collection, Cleansing and Pre-Processing

A drive test in a commercially deployed LTE network is conducted for a total of 13 hours, and RSRP and SINR measurements are recorded at a time interval of 100ms.

Table 4.1: Applying data-imbalance resolution techniques on the potential RLF class.

Classification Algorithm	Metric	Raw Data	Random over sampling	Smote	Random under sampling	Near Miss	CNN	Tomek Links	ENN	NCL	Cluster Centroids	GAN
Regression	Accuracy	97%	88%	89%	88%	95%	95%	97%	98%	97%	90%	97%
KNN	Accuracy	98%	98%	95%	98%	88%	97%	97%	96%	97%	98%	97%
SVM	Accuracy	97%	89%	89%	89%	89%	97%	89%	89%	97%	94%	97%
Naive Bayes	Accuracy	97%	88%	90%	97%	95%	95%	96%	95%	96%	90%	96%
Decision Trees	Accuracy	97%	93%	90%	90%	48%	93%	97%	96%	96%	88%	96%
Random Forest	Accuracy	79%	94%	93%	93%	57%	97%	98%	97%	97%	94%	97%
XGBoost	Accuracy	78%	93%	91%	91%	78%	97%	98%	97%	97%	94%	97%
Deep Learning	Accuracy	74%	89%	89%	89%	72%	97%	99%	72%	98%	94%	97%
Regression	F1	0.75	0.88	0.49	0.88	0.68	0.67	0.74	0.74	0.74	0.53	0.75
KNN	F1	0.78	0.78	0.68	0.78	0.44	0.75	0.78	0.72	0.77	0.69	0.79
SVM	F1	0.73	0.88	0.88	0.88	0.88	0.75	0.88	0.88	0.74	0.63	0.76
Naive Bayes	F1	0.7	0.88	0.5	0.7	0.62	0.69	0.7	0.66	0.69	0.51	0.72
Decision Trees	F1	0.75	0.89	0.9	0.9	0.16	0.57	0.75	0.73	0.74	0.47	0.75
Random Forest	F1	0.86	0.9	0.92	0.92	0.2	0.77	0.79	0.78	0.79	0.66	0.8
XGBoost	F1	0.88	0.91	0.91	0.91	0.31	0.76	0.88	0.78	0.74	0.64	0.79
Deep Learning	F1	0.87	0.88	0.88	0.88	0.1	0.76	0.93	0.1	0.8	0.62	0.76

Moreover, the low level RLF related parameters mentioned in Section 4.2.2 are also registered. Out of the 0.45 million data samples recorded, only 543 actual RLF are observed ($\sim$7 RLF every 10 minutes). This data, if used as it is to train an AI model, can lead to a poorly performing model due to the class imbalance in the training data. For that reason, and to incorporate all the chances of *possible* RLF, using domain knowledge, I mark those rows of the data as potential RLF where even though actual RLF is not observed but the underlying RLF related parameters ($T310, N310, N311$, $maxRACHattempts, maxRLCretransmissions$) showed abnormality.

Next, some of the RLF related higher layer parameters were not received in sync with the physical layer RSRP and SINR data during the logging of drive test data. The incomplete data as a result of the sync issues were filled in with the appropriate RF information. For example, as UE attempts RACH with the target cell only during the HO procedure, the RF data of the target cell was filled in against the respective RACH results. The processed RF data with potential RLF instances label has been plotted in the Fig. 4.3. The tail of the scatter plot in the bottom left area are poor RSRP samples due to late HO instances, where UE is unable to perform HO to the best cell due to poor SINR.

I incorporate the low-level RLF related parameters to have a better insight to the RF condition leading to actual RLF. This also alleviate the data imbalance issue as I obtain 27,794 potential RLF samples as compared to just 543 actual RLF instances. However, even 27,794 potential RLF samples are a small fraction of of the total 0.46 million total samples. This data when fed into the training phase of machine learning gives very poor performance as the non-potential RLF data is dominant. This has been shown under the Raw Data column of table 4.1.

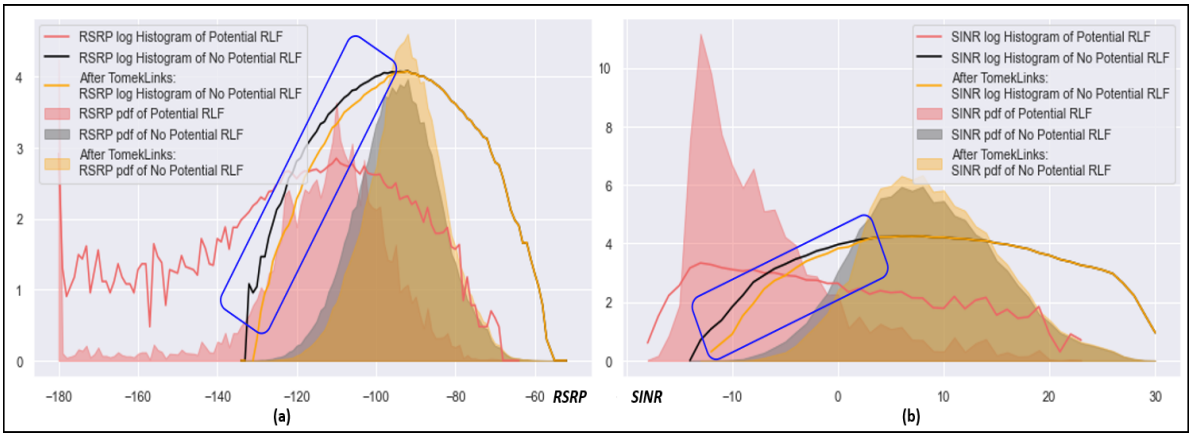

Figure 4.5: Effect of Tomek Links in addressing data imbalance and improving class isolation.

4.3.2 Addressing Data Imbalance

A key challenge in creating an RLF model is the training data class imbalance. If used without a class balancing technique, most machine learning models trained on our data will be biased towards the majority class i.e. no RLF. The resultant accuracy paradox, where the high accuracy of machine learning model is driven by the majority class, and the minority class showing poor performance will be detrimental to the fidelity of the RLF model. In our context, minority class (potential RLF class) is actually the class of interest, and for that reason data imbalance problem must be addressed to have meaningful results. In the following, I briefly discuss the approaches I have leveraged to address data imbalance problem. Here I represent minority class and majority class as C_{min} and C_{maj} respectively.

- Random over sampling randomly duplicates observations from the C_{min} to reinforce its signal.
- Synthetic Minority Oversampling Technique (Smote) synthesises new minority instances.
- Random under-sampling randomly removes observations from the C_{maj}.
- After identifying the two nearest samples in the distribution belonging to different classes, the near miss algorithm eliminates the majority class data point, thereby

trying to balance the distribution.

- Condensed Nearest Neighbor Rule (CNN) works by classifying each majority sample using kNN, and mis-classified sample is assigned to C_{min}.

- A pair of data instances (x_i, x_j) where $x_i \in C_{min}$, $x_j \in C_{maj}$ and $d(x_i, x_j)$ is the distance between x_i and x_j, is called a Tomek link if there is no data instance x_k ($x_k \in C_{min}$ or $x_k \in C_{maj}$) such that $d(x_i, x_k) < d(x_i, x_j)$ or $d(x_j, x_k) < d(x_i, x_j)$. The Tomek link algorithm removes the unwanted overlap between C_{min} and C_{maj} by removing majority class sample from Tomek link data pair. This is done based on the assumption that for the data points that form a Tomek link, either one of them is a noise or both are in the borderline.

- Edited Nearest Neighbor Rule (ENN) removes any instance whose class label is different from the class of at least two of its three nearest neighbors.

- Neighborhood Cleaning Rule (NCL) modifies the ENN where three neighbors of each data data point are found. If the classification of the data point $x_j \in C_{maj}$ given by its three neighbors contradicts the original class of x_i, then x_i is removed. Conversely, if the data point $x_i \in C_{min}$ and the three neighbors miss-classify x_i as a majority class sample, then the nearest neighbors that belong to the majority class are removed.

- Cluster Centroids find the clusters of the majority class with K-mean algorithms. Then it replaces the cluster points with cluster centroids as the new majority samples.

- In Generative Adversarial Network (GAN), two neural networks contest with each other in the training phase. The goal of the first neural network is to befool the second neural neural network by generating synthetic data that resembles the input training data. The role of the second neural network is to correctly identify

the synthetically produced data. In the context, GAN can be used to oversample the C_{min}.

Of the aforementioned techniques to address class imbalance, our results indicate that Tomek links outperforms other techniques (see table 4.1). The highlighted blue rectangle in Fig. 4.5 illustrates that Tomek Links focus on the class boundary to help improve the isolation between the overlapped classes by removing majority samples at the border area.

In the following I describe the training, testing, and validation of the machine learning algorithms after applying the data imbalance resolution approaches discussed above.

4.3.3 Model Building and Validation

The prepared data is scaled and used to train and test several AI techniques for creating a best performing model for RLF prediction as function of observed RSRP and SINR. After splitting the processed data into a training and a test dataset, I develop and validate several classification algorithms that include KNN, decision trees, regression and deep learning-based models. Table 4.1 shows the accuracy and F1 score of the minority class (potential RLF class) for various machine learning models trained on the same data to predict RLF. F1 score observed for majority class for all the machine learning algorithms is higher than 0.9 and has not been included in Table 4.1. Deep learning with data imbalance problem addressed by Tomek Links shows the best results in terms of accuracy, F1 score and domain knowledge (decreasing RSRP and SINR induces more chance of RLF). The decision boundary for the aforementioned model is shown in the Fig. 4.4(h).

Deep Neural Network algorithm belongs to a special class of machine learning, called deep learning and creates a multi-layer perceptron to find the input-output associations. Its basic structure consists of an input layer, output layer and one or more hidden layers

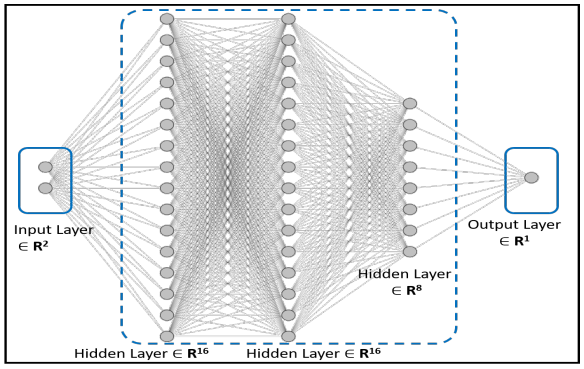

Figure 4.6: Structure of the deep learning based model for predicting potential RLF. The model is trained, tested and validated after addressing data imbalance using Tomek link.

between them, each containing several neurons (or nodes). The number of neurons in the input layer are typically equal to the number of input features, whereas output layer in case of binary classification model consists of a single neuron that holds the prediction output. Number of hidden layers and its neurons are variable and depends on the complexity of the model it is trying to learn. To avoid under or over fitting, I investigate a variety of Deep learning neural network architectures with a range of hyper-parameters as shown in Table 4.2. Our experiments show that a deep learning model with fully connected three hidden layers having 16, 16 and 8 neurons respectively as shown in the Fig. 4.6, yields the best results. The model was trained using epoch size of 50 and batch size of 1.

Table 4.2: Deep Learning Hyperparameters for potential RLF model.

Hyperparameter Name	**Search Range/Value**
DNN depth d	{1,2,3,5}
DNN width w	{5,8,10,16}
Activation Function (Hidden Layers)	Relu
Activation Function (Output Layers)	Sigmoid
Optimizer	Adam (Gradient Descent)
Loss Metric	Binary Cross Entropy

4.4 AI Model for Voice Muting Prediction to Enable Smart EN-DC Activation

Similar to Section 4.3 but for voice call muting, this section describes how actual measurement data from a real VoLTE network is collected and used to develop a deep

learning based AI-model to help identify the set of RSRP and SINR conditions that correspond to potential voice call muting. This model is then used to help activate EN-DC mode if the UE requiring voice bearers send EN-DC activation request (event B1) to the MN having the RSRP and SINR of the MN and SN. Since the handover criteria in 5G is same as in 4G, with the assumption that the same RSRP and SINR measurements of 4G and 5G networks correspond to similar voice muting performance, we can apply the learned voice muting prediction AI model on both LTE and 5G NR. Both 4G and 5G networks broadcast specific signals at different but known frame locations for the UE to measure RSRP, while the SINR is calculated as the ratio of signal to interference and noise. Hence, it can be safely assumed that same pair of RSRP and SINR for both 4G LTE and 5G NR exhibits exactly same muting behavior. However, note that this assumption may be untrue if the difference between the 4G and 5G carrier frequencies is large. This is because of dissimilar channel characteristics and different uplink performance of the two carrier frequencies. Similarly, huge deviance in transmission power can have dissimilar results as well. In that case the presented framework will work but the exact model will have to be retrained using data from the 5G network.

Fig. 4.2 illustrates the high-level overview of the proposed AI powered EN-DC activation framework. For muting prediction, I develop a cascaded 2-stage AI model where the labels obtained from stage-1 potential RLF model (Section 4.3) are used alongside actual voice muting samples obtained from real world measurements. Finally, the second stage AI model is trained to predict potential voice muting.

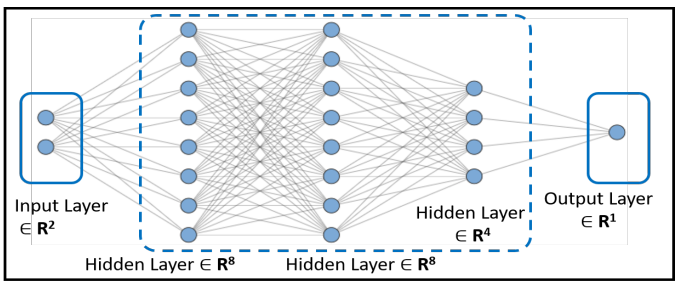

Figure 4.7: Structure of the deep learning based model for predicting potential voice muting. The model is trained, tested and validated using GAN enriched real data.

Table 4.3: Applying data-imbalance resolution techniques on the potential mute class.

Classification Algorithm	Metric	Raw Data	Random over sampling	Smote	Random under sampling	Near Miss	CNN	Tomek Links	ENN	NCL	Cluster Centroids	GAN
Regression	Accuracy	99%	91%	89%	91%	98%	99%	99%	99%	99%	91%	99%
KNN	Accuracy	99%	95%	97%	93%	95%	99%	99%	99%	99%	98%	99%
SVM	Accuracy	99%	94%	96%	93%	94%	99%	99%	99%	99%	87%	99%
Naive Bayes	Accuracy	99%	89%	86%	90%	91%	99%	99%	99%	99%	87%	99%
Decision Trees	Accuracy	99%	97%	97%	89%	88%	97%	99%	99%	99%	90%	99%
Random Forest	Accuracy	99%	97%	98%	92%	92%	99%	99%	99%	99%	97%	99%
XGBoost	Accuracy	99%	96%	97%	93%	91%	99%	99%	99%	99%	29%	99%
Deep Learning	Accuracy	99%	93%	96%	93%	94%	99%	99%	99%	99%	96%	99%
Regression	F1	0.41	0.61	0.11	0.71	0.3	0.45	0.41	0.44	0.43	0.12	0.51
KNN	F1	0.43	0.59	0.21	0.73	0.28	0.45	0.48	0.45	0.48	0.39	0.57
SVM	F1	0.41	0.63	0.27	0.73	0.28	0.43	0.44	0.47	0.49	0.24	0.76
Naive Bayes	F1	0.4	0.6	0.09	0.7	0.27	0.37	0.39	0.42	0.4	0.09	0.79
Decision Trees	F1	0.41	0.6	0.2	0.69	0.29	0.17	0.45	0.46	0.44	0.11	0.46
Random Forest	F1	0.43	0.59	0.33	0.62	0.31	0.42	0.44	0.49	0.49	0.41	0.84
XGBoost	F1	0.41	0.6	0.33	0.63	0.35	0.45	0.45	0.5	0.48	0.47	0.86
Deep Learning	F1	0.45	0.63	0.22	0.62	0.32	0.47	0.51	0.47	0.48	0.26	0.89

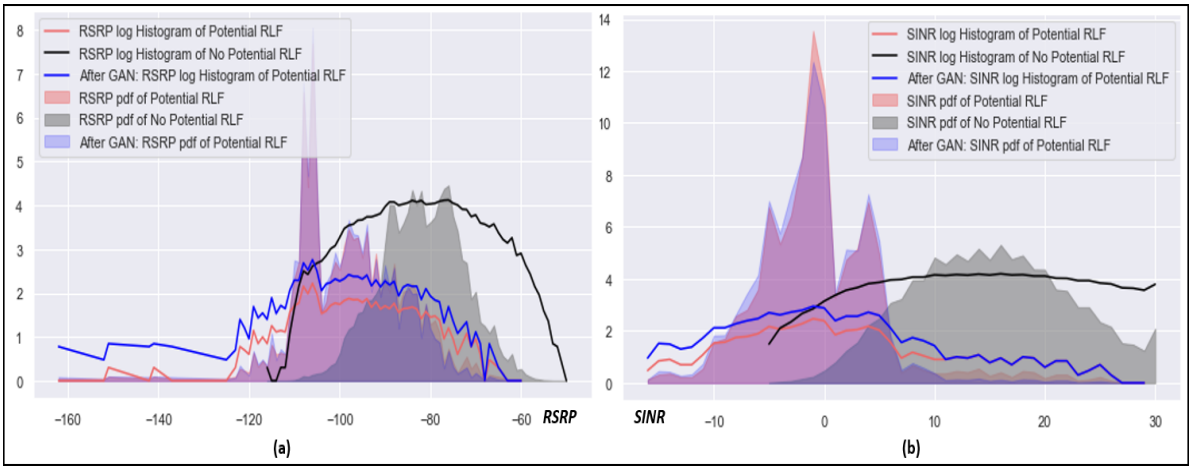

Figure 4.8: Effect of GAN in mitigating class imbalance issue.

4.4.1 Data Collection, Cleansing and Pre-Processing

VoLTE call based drive test is conducted for eight hours and RSRP, SINR measurement are recorded every 100ms. Muting can be observed due to poor RF condition at either the caller or callee location. To accurately identify the RF condition that can lead to call muting, I place one of the call participant static UE under good RF conditions. The other participant UE is placed under a moving vehicle with continuously changing RF condition, and the call muting related data is recorded. Out of the 0.3 million data samples recorded, I observe 2092 actual voice mute occasions (∼4.36 mute occasions per minute). Real-time Transfer Protocol (RTP) packets are continuously exchanged between the UE and BS during the call period, and the absence of RTP packets is observed not only during muting occasions, but also when no call is ongoing. For that reason, I label the rows of data as voice mute only if RTP packets are absent while the voice call is in established phase.

4.4.2 Model Building and Validation

I took the similar model building procedure as for potential RLF model in Section 4.3, hence, for the sake of conciseness I skip the initial details. Table 4.3 shows the performance of various machine learning algorithms that I investigate to determine the best performing potential mute prediction model.

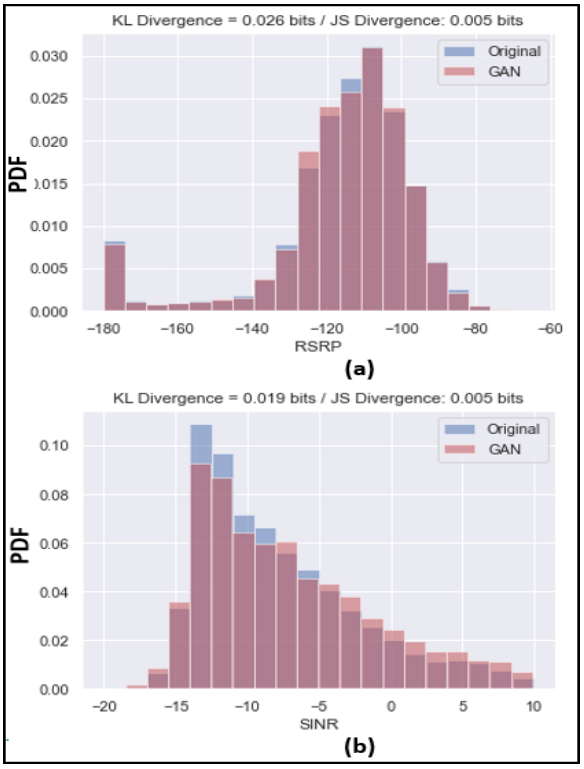

Figure 4.9: Comparison of the original minority class (mute data) and the synthetic data generated from GAN.

Unlike potential RLF model, here I use GAN to address class imbalance, as it shows the best results in terms of F1 score. KL and JS divergence of the GAN generated samples from the real data, along with the probability density function (PDF) of the original minority class and the synthetically generated data is shown in the Fig. 4.9. The results show that GAN manage to produce synthetic data that closely resembles real data.

Fig. 4.8 shows that the class separation in the collected voice call data is much pronounced than the data collected for RLF prediction (see Fig. 4.5 for comparison). This difference in the class distribution stems from the fact that: a) both classes belong to different metrics i.e., potential RLF and potential muting, b) network configures voice bearer activated UE with a different and more aggressive set of mobility parameters to keep the UE in a better RF condition at all times. In the potential RLF case, the class distribution in Fig. 4.5 is more overlapped and Tomek Links successfully improves the class border isolation, resulting in an improved AI model. On the contrary, here I have applied GAN on the minority class alone to have more synthetic samples representing

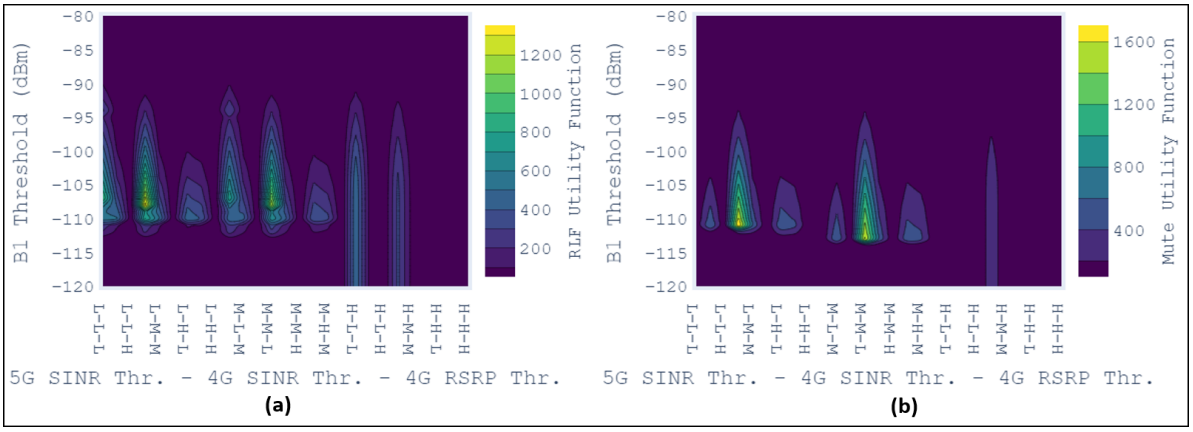

Figure 4.10: Objective function of (a) RLF (4.5) and (b) Mute (4.6) optimization problem.

the minority class. The resultant enhancement of the minority class helps to improve model accuracy.

Building on insights from training RLF prediction model, that show that a deep learning-based model out performs other machine learning models, here I focus on deep learning models only. I investigate a range of deep learning architectures with a variety of hyper-parameters to prevent under- or over-fitting as shown in table 4.2. Our experiments show that for voice mute prediction problem, with used training data, a deep learning model with fully connected three hidden layers having 8, 8 and 4 neurons respectively as shown in the Fig. 4.7 out performs all other experimented architectures. The model was trained using epoch size of 50 and batch size of 1.

4.5 QoE Aware EN-DC Activation

Following are the objectives mobile network operators should take into consideration when enabling EN-DC in their network:

- Maximize EN-DC request by the EN-DC capable UE to have more chances to leverage 5G NR features.
- Facilitate EN-DC activation for every EN-DC request, i.e., minimize ($\sum$ EN-DC Requests - $\sum$ EN-DC Activations).

- Avoid degradation in retainability due to RLF at either 4G or 5G network after EN-DC activation.

- Prevent voice muting after activating EN-DC for UE with voice service demands.

Table 4.4: List of acronyms used in optimization problem formulation.

Symbol	Description	Symbol	Description
U	Set of all UEs	u	Any user u $\in U$
U_c	Set of UEs with EN-DC configuration	U_a	Set of EN-DC activated UEs
δ^u_{5R}	5G RSRP of u	θ_{B1}	5G RSRP threshold
δ^u_{4R}	4G RSRP of u	θ_{4R}	4G RSRP threshold
δ^u_{5S}	5G SINR of u	θ_{5S}	5G SINR threshold
δ^u_{4S}	4G SINR of u	θ_{4S}	4G SINR threshold
$\mathbf{\Delta}^u$	$[\delta^u_{5R}, \delta^u_{4R}, \delta^u_{5S}, \delta^u_{4S}]$	$\mathbf{\Theta}$	$[\theta_{B1}, \theta_{4R}, \theta_{5S}, \theta_{4S}]$
$\mathbf{\Delta}^{u,4}$	$[\delta^u_{4R}, \delta^u_{4S}]$	$\mathbf{\Delta}^{u,5}$	$[\delta^u_{5R}, \delta^u_{5S}]$
α	EN-DC Activation function	ζ	Potential RLF AI-Model
β	RLF function	η	Potential Muting AI-Model
γ	Muting function	-	-

Using the notations defined in Table 4.4, I can write EN-DC activation function as:

$$\alpha(\mathbf{\Delta}^u, \mathbf{\Theta}, U_c) = \sum_{u \in U_c} 1\left[\mathbf{\Delta}^u_i > \mathbf{\Theta}_i \; \forall \; i\right] \tag{4.2}$$

where 1 is the indicator function, and the subset $U_c \subseteq U$ is the set of EN-DC capable UEs configured with B1 measurement report. $\mathbf{\Delta}^u_i$ is the i-th element of the set of RF condition of user $u \in U$, i.e., for any user u, $\mathbf{\Delta}^u_1 = \delta^u_{5R}$, $\mathbf{\Delta}^u_2 = \delta^u_{4R}$, $\mathbf{\Delta}^u_3 = \delta^u_{5S}$, $\mathbf{\Delta}^u_4 = \delta^u_{4S}$. Similarly, the i-th element of the set of thresholds is $\mathbf{\Theta}_i$, where $\mathbf{\Theta}_1 = \theta_{B1}$, $\mathbf{\Theta}_2 = \theta_{4R}$, $\mathbf{\Theta}_3 = \theta_{5S}$, $\mathbf{\Theta}_4 = \theta_{4S}$.

UEs upon EN-DC activation may experience RLF due to poor RF condition. RLF function denoted here by β can be defined as:

$$\beta(\mathbf{\Delta}^u, \zeta, U_a) = \sum_{u \in U_a} max\left(\zeta(\mathbf{\Delta}^{u,4}), \zeta(\mathbf{\Delta}^{u,5})\right) \tag{4.3}$$

where $U_a \subseteq U_c$ is the set of UEs with EN-DC activated, and ζ is the potential RLF AI-Model, which takes in $\mathbf{\Delta}^u$ as input and outputs a prediction of 1 or 0 representing

occurrence of potential RLF and no RLF respectively. The output of the potential RLF AI-model is represented as $\zeta(\mathbf{\Delta}^u)$.

Similarly, for the set of UEs requiring voice services, function of mute η is denoted as:

$$\eta(\mathbf{\Delta}^u, \eta, U_a) = \sum_{u \in U_a} max\left(\eta(\mathbf{\Delta}^{u,4}), \eta(\mathbf{\Delta}^{u,5})\right) \tag{4.4}$$

The potential muting AI-Model η takes in $\mathbf{\Delta}^u$ as input and outputs a prediction of 1 or 0 representing occurrence of potential muting and no muting respectively. The output of the potential muting AI-model is represented as $\eta(\mathbf{\Delta}^u)$.

Operators can increase EN-DC activations by configuring lower values of EN-DC thresholds Θ. This, however, can lead to RLF or voice muting soon after EN-DC activation, rendering the dual connectivity procedure useless. Keeping in view this tradeoff, the optimization problem in subsection 4.5.1 and 4.5.2 is formulated to achieve maximum utility and resource efficiency.

4.5.1 RLF Aware EN-DC Optimization

I formulate a multi-objective optimization problem to smartly maximize EN-DC activations while preventing the chances of RLF occurrences:

$$\begin{aligned}
\underset{\Theta^r=[\theta^r_{B1},\theta^r_{4R},\theta^r_{5S},\theta^r_{4S}]}{\text{maximize}} \quad & \frac{\alpha^w}{\beta^{(1-w)}} \\
\text{subject to} \quad & \theta^r_{B1,low} \geq \theta^r_{B1} \leq \theta^r_{B1,high}, \\
& \theta^r_{4R,low} \geq \theta^r_{4R} \leq \theta^r_{4R,high}, \\
& \theta^r_{5S,low} \geq \theta^r_{5S} \leq \theta^r_{5S,high}, \\
& \theta^r_{4S,low} \geq \theta^r_{4S} \leq \theta^r_{4S,high}, \\
& w \leq 1.
\end{aligned} \tag{4.5}$$

where w is the operator defined weight that can be used to adjust the relative importance of EN-DC activations (α), and RLF (β). The range of optimization variables

and constraints indicate (4.5) is a large-scale non-convex NP-hard problem due to the inherent coupling of optimization parameters and the EN-DC requests. Non convexity stem from the fact that I am dealing with four integer metrics (RSRP and SINR of 4G and 5G) in a heterogeneous multi-RAT network deployment where randomness in UE location, and resource requirement results in variable cell loads that affect 4G and 5G SINR metric. In addition, the 4G and 5G RSRP metric does change with the distance from the BS, however, non-uniform BS deployment along with user mobility makes the RSRP metric non deterministic. For the RSRP range [-120dBm, -90dBm], and the SINR range [-10dB, 10dB], I have total 741,321 distinct combinations of the four optimization parameters. The plot of the RLF and mute objective function using brute-force is shown in the Fig. 4.10 wherein RSRP categories are defined as: L: <-110dBm, M: -110dBm to -100dBm, and H: >-100dBm. Similarly, SINR rages are defined as <-3dB for L, >3dB for H, and -3dB to 3dB for M category. The non-convex nature of the problem can be seen from the visualizations in Fig. 4.10(a) and 4.10(b).

4.5.2 Voice Muting Aware EN-DC Optimization

Similar to (4.5), optimization function for potential mute model can be formulated as below:

$$
\begin{aligned}
\underset{\Theta^m=[\theta^m_{B1},\theta^m_{4R},\theta^m_{5S},\theta^m_{4S}]}{\text{maximize}} \quad & \frac{\alpha^w}{\gamma^{(1-w)}} \\
\text{subject to} \quad & \theta^m_{B1,low} \geq \theta^m_{B1} \leq \theta^m_{B1,high}, \\
& \theta^m_{4R,low} \geq \theta^m_{4R} \leq \theta^m_{4R,high}, \\
& \theta^m_{5S,low} \geq \theta^m_{5S} \leq \theta^m_{5S,high}, \\
& \theta^m_{4S,low} \geq \theta^m_{4S} \leq \theta^m_{4S,high}, \\
& w \leq 1.
\end{aligned}
\tag{4.6}
$$

where w is the operator defined weight that can be used to adjust the relative importance of EN-DC activations (α), and mute (γ). Compared to UEs with an active HTTP/FTP

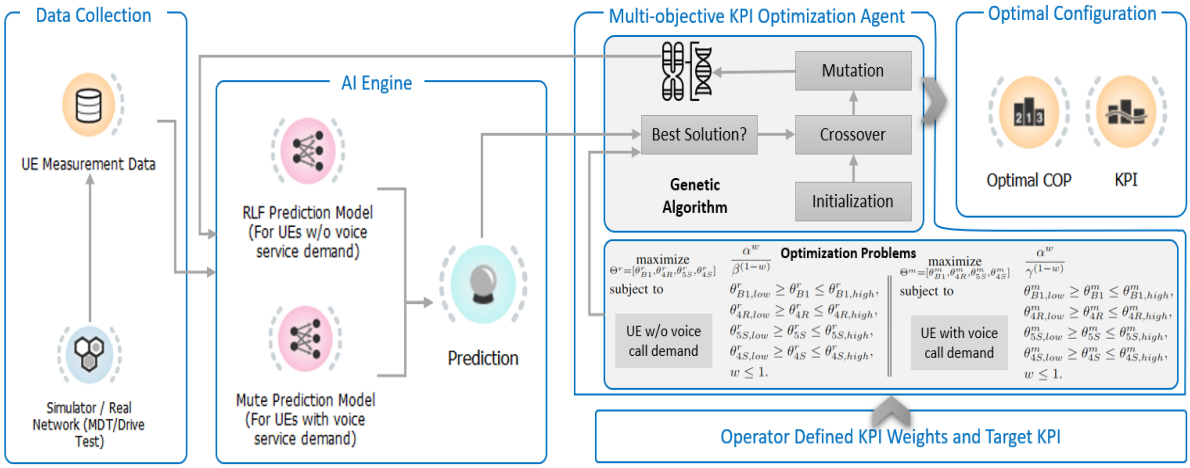

Figure 4.11: Proposed smart EN-DC activation framework.

session, the UEs with an active voice bearer are configured with a different set of mobility and retainability parameters. The goal of this approach is to keep the UE undergoing a voice call to good radio conditions, and other factors like load balancing and HO rate are given less priority. The optimization problem in (4.6) is formulated while keeping in view the same strategy where UEs with voice services will have a different set of EN-DC activation thresholds.

4.6 Proposed Smart EN-DC Activation Framework and Simulation Results

The proposed smart EN-DC activation framework effectively use the learning from the AI model presented in section 4.3 and 4.4. The framework requires the data collected from a live network to be fed into the AI engine where RLF (or mute) prediction takes place. Next, the optimization agent evaluates the multi-objective KPI optimization problem formulated in the previous section. This is done keeping in view the operator defined weightage to the number of EN-DC activations and the number of RLF/mute. I solve this non-convex problem using GA heuristic. As illustrated in the Fig.4.11, the optimization agent frequently pools the RLF/mute prediction from the AI engine, and the optimal Configuration and Optimization Parameters (COPs) that yield the maximum utility function are obtained.

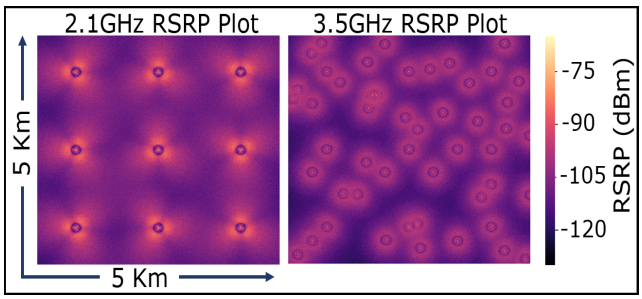

Figure 4.12: RSRP plot of deployed 4G and 5G network.

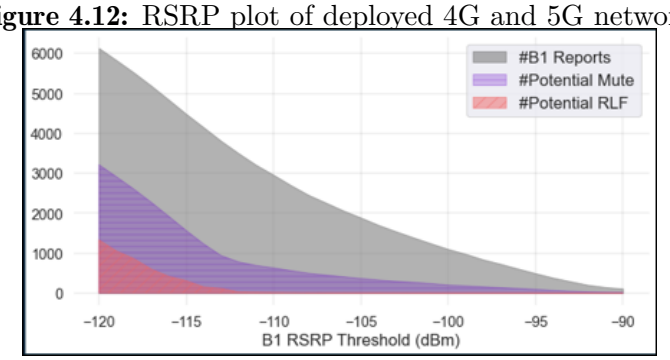

Figure 4.13: Number of UE generated B1 reports (EN-DC activation requests) against RSRP threshold.

4.6.1 Simulation Setup

I implement the proposed framework in a state of the art 3GPP compliant network simulator called SyntheticNET [152]. SyntheticNET has many key features missing in other simulators including 3GPP based mobility management, adaptive numerology, and allocated resource element(s) based SINR calculation.

Table 4.5: Simulation details for Smart EN-DC activation.

Technology	**4G LTE**	**5G NR**
Frequency	2.1GHz	3.5GHz
Cell Type	Macro Cell	Small Cell
Antenna Type	Directional	Omni
Number of Cells	27	16
Transmit Power	40dBm	30dBm
Base Station Height	30m	20m

A multi-RAT (Random Access Technology) network with nine macro 4G eNBs each having three sectors, and sixteen higher frequency omni directional 5G gNBs are deployed in a square of 25km^2 area. LTE eNBs are laid out uniformly in a grid form, while 5G small cells are deployed randomly representing hotspot locations. A total of 300 mobile

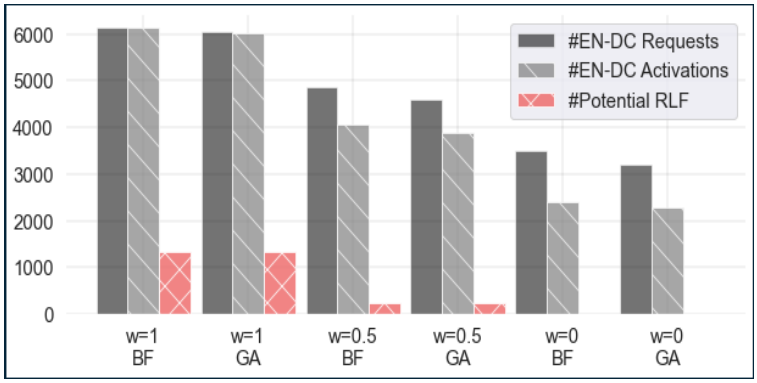

Figure 4.14: Number of EN-DC activations and RLF observed when using optimal parameters in Table 4.6.

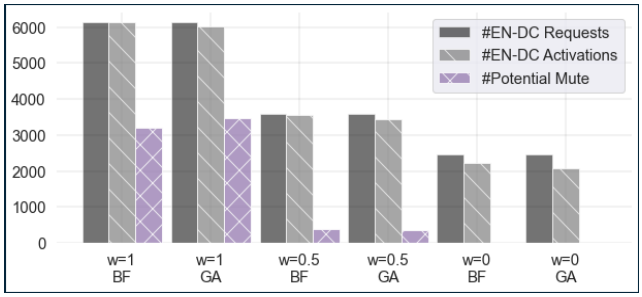

Figure 4.15: Number of EN-DC activations and RLF observed when using optimal parameters in Table 4.7.

UEs traverse the area following random way point mobility model. RSRP plot of the deployed network is shown in the Fig. 4.12. Speed of the users is set to 120km/h and the simulation run for 12,000ms. More detail about the network configuration can be found in Table 4.5.

UEs are configured to measure RF condition of 5G gNB every 0.5s, and event B1 measurement report is sent to the MN if the B1 criteria is met. Fig. 4.13 shows the

Table 4.6: Optimal parameters obtained from genetic algorithm (GA) for a UE with a data call requirement.

w	Algo	Iterations	Utility	Optimal Parameters $\Theta^r = [\theta^r_{B1}, \theta^r_{4R}, \theta^r_{5S}, \theta^r_{4S}]$
1	BF	741,321	1246.5	-120dBm, -120dBm, -6dB, -7dB
	GA	335	1225.5	-120dBm, -119dBm, -8dB, -6dB
0.5	BF	741,321	47.2	-112dBm, -118dBm, -3dB, -2dB
	GA	2890	46.1	-112dBm, -118dBm, -7dB, -2dB
0	BF	741,321	2.2	-108dBm, -118dBm, -1dB, -2dB
	GA	5543	2.1	-108dBm, -118dBm, -2dB, -2dB

Table 4.7: Optimal parameters obtained from genetic algorithm (GA) for UEs requiring voice call services.

w	Algo	Iterations	Utility	Optimal Parameters $\Theta^r = [\theta^r_{B1}, \theta^r_{4R}, \theta^r_{5S}, \theta^r_{4S}]$
1	BF	741,321	1142	-120dBm, -120dBm, -6dB, -7dB
	GA	969	1130.4	-120dBm, -120dBm, -7dB, -10dB
0.5	BF	741,321	49	-115dBm, -110dBm, -5dB, -1dB
	GA	5607	46.6	-115dBm, -110dBm, -10dB, 0dB
0	BF	741,321	2.2	-112dBm, -110dBm, 0dB, -1dB
	GA	10987	2.1	-111dBm, -110dBm, -6dB, -1dB

effect of changing B1 threshold on the number of B1 reports (EN-DC requests), potential RLF and potential mute occurrences. Fig. 4.13 signifies the need for a smart EN-DC activation scheme i.e., the importance of optimally assigning B1 threshold. An incorrect B1 threshold may deteriorate retainability Key Performance Indicator (KPI) or integrity KPI through large number of RLF instances, and voice muting.

4.6.2 Performance Evaluation

The AI models (Fig. 4.2) are developed from the real world data, however, I evaluate the performance of our proposed framework using the simulated data obtained from SyntheticNET - a realistic 3GPP compliant network simulator.

Table 4.6 shows the optimal parameter set for w = [1, 0.5, 0], where the optimization techniques of Brute-Force (BF), and Genetic Algorithm (GA) are compared. A larger value of w tends to maximize EN-DC requests while giving low priority to RLF/mute reduction. Fig. 4.14 shows the number of EN-DC requests, EN-DC activations and the predicted RLF occurrences when w is varied. Results show that for the three values of w used during the evaluation, GA shows slightly less ENDC activations compared to BF. However, note that for the given network deployment, GA converges to the optimal parameters in much smaller number of iterations than BF (see Table 4.6). This is critical if the operators want to have separate set of optimal EN-DC parameters per cell, or if with the temporal change in load condition, the optimal EN-DC activation parameters

are changing, and new set of parameters need to be updated dynamically.

With no-condition scenario of w=1 in (4.5), the optimization function maximizes EN-DC activations and disregard RLF/mute occurrences. This is shown in the Fig. 4.14 where 1328 of the 6025 EN-DC activations results in RLF. Fig. 4.14 also shows that we can help reduce RLFs by decreasing w, and can totally eliminate the chances of RLF with w=0. This however comes at ∼50% loss of EN-DC activations. One more observation I can induce from the Fig. 4.14 is that for the given network model, although GA results in slightly lower EN-DC requests and activations than BF, the difference between EN-DC requests and activations is less in GA. This will lower the signaling overhead and UE energy consumption will be more efficient, as most of EN-DC requests will be successfully acknowledged with EN-DC activations. For w=0, BF solution results in 3501 EN-DC requests and 2413 EN-DC activations (1088 EN-DC requests were not entertained due to chances of RLF from poor RF condition). On the contrary, only 914 EN-DC requests were discarded when the optimal parameters obtained from GA were used that resulted in 3209 and 2295 EN-DC requests and activation respectively.

Similar behavior can be observed for the UEs with the requirement of voice services. Table 4.7 shows the optimal set of parameters needed to activate EN-DC while considering different weightage assigned to minimize the chances of voice muting. Similar as for Table 4.6, GA performs almost similar EN-DC activation count and chances of mute with much lesser iterations than BF.

Fig. 4.15 shows that zero chances of mute instances can be achieved by assigning more weightage to mute using w=0. However, since UE is more susceptible to mute rather than RLF, zero mute occasions can be obtained at the cost of lower EN-DC activations (2085) compared to the similar case with w=0 in Fig. 4.14 (where zero RLF chances are observed with 2295 number of EN-DC activations).

4.7 Conclusion

EN-DC mode addresses strict QoE requirements of the UE by enabling multi-connectivity to 4G and 5G cells. However, multi-connectivity can be beneficial only if the RF condition of participating 4G and 5G cells are above a certain threshold. Currently, there does not exist EN-DC mode selection scheme in literature that takes into account the risk of RLFs and voice mute. This chapter proposes a smart EN-DC triggering scheme by which RLF and mute due to poor RF conditions can be minimized. The scheme works by selecting the best B1 thresholds based on insights from a deep learning based AI model to predict RLF and mute. The core RLF prediction model is developed, trained, and validated using real network measurements of RSRP, SINR and underlying 3GPP based RLF related parameters. The value of these low-level parameters are used to identify potential RLF against RSRP, SINR values. I use Tomek Links approach to address the class imbalance and enhance the classification accuracy. The mute prediction model on the other hand, is a two-stage AI model that employs potential RLF AI-model along with the voice call muting samples extracted from real network measurements during a voice call. The class imbalance issue in the potential mute model is addressed using GAN.

Simulation results based on a state of the art 3GPP compliant network simulator show that for the analyzed network deployment, compared to the state of the art i.e., no smart conditioning on EN-DC, our proposed scheme can totally eliminate the RLF and mute occurrences. The optimal RSRP and SINR thresholds obtained from the presented optimization function help reduce RLF and mute occurrences from 1328 and 3208 cases to zero potential RLF and potential mute cases respectively.

CHAPTER 5

AI-Assisted Joint Search Method for mmWave Cell Discovery

5.1 Introduction

5G New Radio (NR), with innovative use cases of enhanced Mobile Broadband (eMBB) for large volume transmissions, massive Machine Type Communications (mMTC) for sensors and Internet of Things (IoT) devices, and Ultra Reliable Low Latency Communications (URLLC) for self-driven vehicles will provide much-anticipated use cases and innovative ideas that will benefit both commercial and consumer side in urban as well as rural areas of the world. Connected devices will be three times the global population in 2023 [153]. Studies project that global mobile data traffic will increase from 50 Exa Bytes per month to almost 230 Exa Bytes per month in 2026 [154].

The resultant spectrum gridlock will be broken primarily by a) utilizing wide-band mmWave frequency cells [107, 109], and b) by the re-use of Shannon's capacity through ultra-dense Base Station (BS) deployment in both High Frequency (HF) and mmWave band. However, ensuring the availability of the spectrum to the User Equipment (UE) remains an open challenge of utmost importance. 3GPP standards on Carrier Aggregation [155], and Dual Connectivity [149] make use of excessive signaling exchanges between the BS and UE to configure, activate and deactivate respective multi-connectivity approaches. Moreover, access to the mmWave band resources is yet another unprecedented challenge.

The wideband mmWave BSs though dramatically scales up the system capacity, comes with an unprecedented challenge to the cell availability for the UEs. The peculiar nature of the 30-300 GHz frequency range of mmWave spectrum due to the hallmarks of high path loss, directional transmissions, and sensitivity to minute environmental variation

Cell Discovery Method	Reference	Short Description	Pros	Cons
Exhaustive Search Method	[7]–[11]	A simple strategy for cell discovery could be to sweep through all possible antenna configurations looking for a rendezvous between BS and UE via Brute Force.	Equal chances among all UEs to get network access	Low efficiency due to high i. BS energy consumption ii. UE latency in network access
Hierarchy/Binary Search Method	[12], [13]	Extension of Exhaustive Search Method where every iteration successively divides the target UE search area by configuring smaller beams	Larger probability of UE network access	i. Too much BS resources spend for every UE ii. Not suitable for area with large number of UEs
Hybrid Cell Discovery	[14], [15]	UE estimated location is found through exhaustive search, and then hierarchy method will refine the beam alignment	Lower latency than the above two methods	Not suitable for area with large number of UEs
Context Based Cell Search	[16]–[19]	Search is focused towards the crowded areas known through for e.g., Call Detail Record (CDR) Data	achieves higher network efficiency, and adaptability to temporal shifting of crowded areas.	Low Fairness towards the UEs located in sparsely populated areas
Joint Search Method	[20]–[24]	Macro BS sharing the UE location to mmWave BS	Less mmWave BS resources spent on UE discovery, and lowest latency due to high hit-rate%[a]	i. Signaling load and configurational complexity due to UE dual connectivity towards macro and mmWave BSs ii. Susceptible to incorrect UE location due to GPS error/indoor location

Figure 5.1: Comparison of mmWave Cell Discovery Approaches.

and obstacles hinders the initial access procedure. Initial access procedure consists of cell discovery, extraction of system information, and random access procedure. Successful mmWave cell discovery in addition to beam alignment between UE and mmWave cell also requires Line-of-Sight (LoS) transmission path. Unlike HF cells, mmWave beams with Non-Line-of-Sight (NLoS) scenario won't be decoded at the UE terminal due to dramatic link quality deterioration.

5.1.1 Related Work and Motivation

Researchers are working on devising suitable strategies to achieve an efficient cell discovery process in mmWave systems, that reduces latency and overhead (number of pilot symbols). Table 5.1 summarizes some of the literature work directed towards mmWave cell discovery [156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173].

While most of the research papers [156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173] address the mmWave alignment issue between UE and BS, none of the proposed idea incorporate the realistic LoS scenario deemed essential for successful mmWave cell discovery. Authors in [157] make use of a heavily used exhaustive search method to propose optimal beamwidth design taking into account

the tradeoff between mmWave cell search delay and beamforming gain. Authors in [158] introduced the concept of beam discovery signal to help identify the beam used during the exhaustive search method. The beam discovery signals are also used to identify interfering beams, and later interference mitigation takes place using orthogonal codes.

The comparison of the exhaustive search method and hierarchical search method can be found in [161]. Authors in [161] concluded that hierarchical search can achieve similar beam alignment performance to exhaustive search with low overhead (fewer pilot symbols used). A hybrid method with the strengths of the exhaustive and hierarchical method has been proposed in [163] that outperforms the hierarchical search method in terms of miss-rate% (probability of misdetection), and exhaustive in terms of discovery delay.

Authors in [167] discusses the context-based cell search approach where intelligent mmWave BSs steers their beams through a known populated area for UE discovery. Unlike other approaches, this scheme increases hit-rate% by avoiding beam transmission towards sparsely populated areas, or towards blockages like trees, buildings, rivers, etc. Another promising cell discovery approach has been proposed in [169, 170, 171, 172, 173] where joint collaboration between macro BS and mmWave BS efficiently discovers the UE with the macro BS feeding the UE location to the mmWave BS. Authors in [173] exploit a probabilistic neural network to suggest the optimal beamwidth to achieve successful cell discovery. Simulation results in [173] show that latency can be decreased from 1.6ms using an exhaustive search method to just 0.18ms using the proposed joint search-based method. Moreover, a misdetection probability of as low as 0.08 can be achieved.

While research ideas like in [157, 171] presents analytical model for coverage probability, none of the proposed schemes in literature [156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173] incorporates Non-Line-of-Sight (NLoS) induced coverage hole in the cell search procedure. This is critical due to the high

sensitivity nature of mmWave to blockages as UEs under NLoS conditions might render the BS efforts taken towards mmWave cell search procedure totally useless. Moreover, the user experience of the mobile UEs served by mmWave cells may deteriorate rapidly once the mobile UEs enter NLoS areas due to blockage. This is detrimental both to the bandwidth-hungry eMBB applications and time-critical URLLC use-cases.

5.1.2 Contribution

The main contributions of this work can be summarized as follows:

- To the best of authors' knowledge, this is the first joint search-based mmWave cell discovery framework build using realistic mmWave network data of radio link failures (RLFs), coverage holes, and serving mmWave cell identifiers.
- I employ domain knowledge-assisted data sparsity techniques to predict the optimal mmWave cell in areas with sparse UE distribution. The optimal mmWave cell prediction on the unlabeled bins obtained from a real mobile network is made using
  - traditional interpolation techniques of Inverse Distance Weighted (IDW), Moving Average, and Natural Neighbor and Nearest Neighbor.
  - domain knowledge-based custom Algorithms of Nearest Neighbor Count (NNC), and Weighted nearest Neighbor Count (WNNC).
  - machine learning and deep learning models each with optimized hyper-parameters (Regression, Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Trees, XGBoost, and Deep Learning (DL)).
- A key output of the proposed framework is a map of optimal mmWave cell, which can then be used to attempt mmWave cell discovery. Similarly, UE in coverage hole due to no proximity to a mmWave cell or due to NLoS condition can be exempted

from the mmWave cell discovery procedure, hence conserving the resources of both UE and BS. This will ultimately lead to an efficient mmWave cell discovery procedure with a higher hit rate%.

- I leverage the architecture of E-UTRAN New-Radio Dual Connectivity (EN-DC) [149] to facilitate the proposed joint search-based mmWave cell discovery.
  - A case study is presented to show how 4G macro cells with known UE GPS location can be fed into the proposed framework to identify optimal 5G mmWave cells that can reliably service the EN-DC capable UE.
  - I compare the EN-DC activation rate after mmWave discovery procedure proposed in this chapter to the a) mmWave cell discovery method to the nearest mmWave cell and b) mmWave cell discovery based on the sparse data only (without addressing data sparsity).
  - Results show that compared to the other two approaches, our proposed framework effectively activates EN-DC for most of the UEs while keeping the unsuccessful mmWave search attempts to a minimum.

The rest of the chapter is organized as follows. Section 5.2 discusses the data collection from a realistic mmWave environment. The approaches to optimal mmWave cell identification for a given UE location have been demonstrated in Section 5.3. Section 5.4 discusses a case study to efficiently activate EN-DC for 5G leg of mmWave. Finally, I conclude the chapter in Section 5.5.

5.2 Synthetic Data Collection From a Realistic mmWave Environment

5.2.1 Challenges in Real Network mmWave Data Collection

Collecting the mmWave related network data from a live network though plausible in theory is impractical because of several reasons that include:

- mmWave networks are not fully deployed in our location.
- Even for other areas like Los Angeles where some network operators have already deployed mmWave network, data acquisition is difficult and might not be even useful because
  - Currently the number of mmWave enabled UEs is in scarce and we will not be able to collect adequate data samples.
  - Existing techniques of mmWave cell discovery are based on exhaustive search and the associated inefficiency in cell discovery will contribute to low UEs camping on mmWave cell. As a result, the number of available data samples will be very few. This further adds to the significance of our work, through which mmWave cell camping can be enforced for the candidate UEs.
  - Subscriber data confidentiality further hinders data collection from the existing mmWave UEs.
  - Drive test-based data collection in a congested place like Los Angeles would be expensive both in terms of time and resources. Furthermore, we will get data only from a subset of the target area, and that can lower the effectiveness of the presented framework.

5.2.2 SyntheticNET Upgrade

In the backdrop of the aforementioned challenges, I exploit a 3GPP-compliant state-of-the-art system-level simulator named SyntheticNET [152]. SyntheticNET simulator has been calibrated against real network measurements to ensure the validity of the data generated through it. However, although SyntheticNET, in its current form, supports features related to cell discovery such as 3GPP-based initial cell selection [174], it is tailored more to mimic a network operating on lower frequency bands (i.e. maximum 3.5GHz). To address this issue, I incorporate several upgrades to make Synthetic more

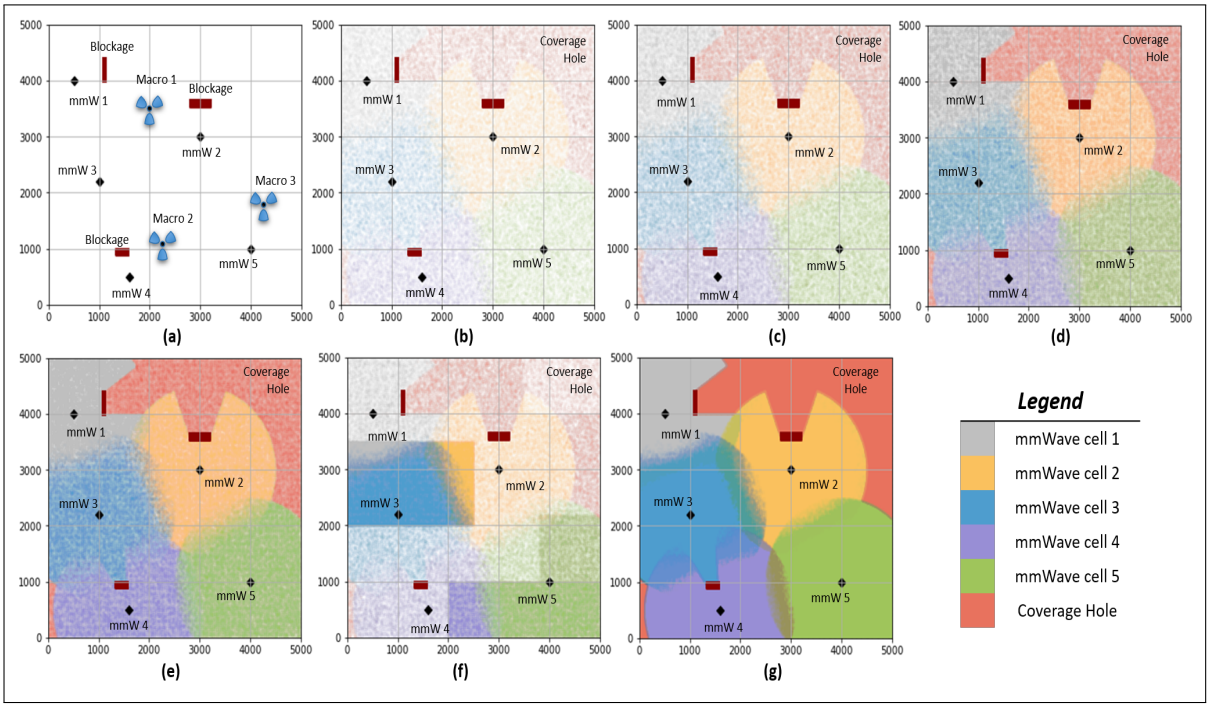

Figure 5.2: (a) System model with macro cell, mmWave cell and blockage locations. mmWave optimal cell coverage map for (b) Use Case 1 - 5% uniform sparse data, (c) Use Case 2 - 10% uniform sparse data, (d) Use Case 3 - uniform sparse 20% data, (e) Use Case 4 - uniform sparse 30% data, (f) Use Case 5 - non uniform sparse data, and (g) Ground Truth.

suitable for mmWave simulation.

To cater to the macroscopic propagation effects in a mmWave simulation environment, first, I utilize a real antenna patterns from a mmWave antenna available commercially [175]. The use of realistic antenna pattern helps in a more accurate mmWave propagation modeling. Instead of using two pathloss models for LoS and NLoS, I utilize a single pathloss model for LoS scenario. For NLoS situations, I model the attenuation caused by blockage by incorporating actual obstructions in the simulator. This approach is more realistic and practical as the location, dimensions and even signal degradation respective to each unique obstruction can be accurately configured instead of analytical approximations.

5.2.3 System Model Used for Data Collection

As previously mentioned, for both LoS and NLoS conditions, I utilize a single pathloss model. Using the calculated path loss, I determine the received power of the user. The downlink RSRP R_u^s from the serving mmWave cell s to user u is given by:

$$R_u^s = P_t^s G_u G_u^s \delta_u^s \alpha (r_u^s)^{-\beta} \tag{5.1}$$

where P_t^s is the transmit power of serving mmWave cell s, G_u is the gain of user equipment, G_u^s is the transmitter antenna gain of the mmWave cell s towards user u, δ_u^s is the shadowing observed from the mmWave cell s at the location of user u, α is the pathloss constant, β is the pathloss exponent and r_u^s represent the distance of user u from cell c. The values of α, β, and δ are based from the study conducted in [176].

5.2.4 Simulation Setup and Data Generation

I use an area of size 5km x 5km for the simulation as shown in Fig. 5.2(a). I deploy a heterogeneous network with two macro BSs radiating at 2.1GHz frequency, and five omni-directional mmWave BSs operating in the 28GHz band. Fig. 5.2(a) shows the system model diagram of the deployed 5km $\times$ 5km network area with the location of macro and mmWave cells. Moreover, several obstructions are put in place to realistically model the NLoS scenario.

5.2.5 Sparsity in Realistic Traffic Modeling

Even with the number of connected devices to be three times the global population by 2023 [153], certain areas of the globe will have incomplete traffic map due to sparse human populations. Similarly, industrial areas and high-tech factories with few IoT devices and robotics will result in a sparse UE distribution chart over the geographical area. Emerging mobile networks with mmWave bands will initially observe low traffic

Table 5.1: Description of Simulation Parameters

Parameter Description	Value
Simulation area	25 km^2
Number of macro BSs	2
Macro cell frequency	2.1 GHz
Number of mmWave BSs	5
mmWave cell frequency	28 GHz
mmWave cell height	10 m
mmWave Transmission Power	20 dBm
Pathloss Exponent	5
Shadowing Standard Deviation	8
Number of UEs per Use Case (UC)	UC1: 30, UC2: 60, UC3: 120, UC4: 240, UC5: 190
% of Mobile UEs	70%
Mobile UE velocity	60 km/h
Total Simulation Time	15000 ms

due to a low number of mmWave supported devices readily available. As a result, the mmWave network data from a real network is anticipated to be sparse.

To simulate sparsity, I utilize SyntheticNET and present five different use cases (UC) of user distribution, where UC1 to UC4 represents uniform user distribution of 5%, 10%, 20%, and 30% respectively. Additionally, UC5 represents a more realistic non-uniform distribution of UEs. 70% of all UEs in all five use cases are considered mobile. The network-level simulation parameters are summarized in Table 5.1.

During the simulation, UEs were configured to camp initially on 2.1GHz macro cells, that were providing coverage to UEs within the target area. I assume macro cell to have error-free UE location to share with the mmWave cells, thus conforming to the joint search method. Moreover, perfect alignment is considered between UE and mmWave cells. UEs served by macro cell then attempt mmWave cell camping to the nearest BS. Upon failure, UE attempts to camp on the second nearest mmWave BS, and the process continues unless no suitable mmWave cell is found or the distance between UE and BS exceeds cell-range denoted by κ. The parameter κ is a common parameter used in existing mobile networks that prevent far away UEs to camp on the overshooting cell. As a result, UEs can be ensured to have good signal strength and high uplink interference

from distance UEs can be reduced. In our simulation, I use κ value of 1500m to limit mmWave band small cells coverage to far away UEs. This conforms with the mmWave environment where mmWave cells with beamforming will have pencil-like beams and mmWave UE will have better signal reception than in the case of macro cells.

I record the radio link failure of the mobile UEs camped on mmWave cell as they travel through the designated network area. The radio link failure is observed due to dramatic signal deterioration from the NLoS reception induced by the blockage between UE and mmWave BS. Similarly, radio link failure can be observed due to UE getting further from the serving mmWave BS by a distance equal to the configured κ parameter. Similarly, the failed mmWave cell camping attempts due to no optimal mmWave cell for the static and mobile UEs are recorded as well. Both the radio link failure and failed cell camping is marked as coverage hole in this work. Fig. 5.2(b-f) illustrates the different sparsity levels in UC1 to UC5, and the resultant coverage hole and optimal mmWave cell coverage obtained from the simulation results. Finally, I increase the number of UEs to 2000 and uniformly dispersed the UE before running the simulation in order to obtain the ground truth (Fig. 5.2(g)) to verify our results presented in the next section.

5.3 Identifying Optimal mmWave Cell

In this section, I analyze various techniques to address the sparse data typically obtained from the mobile network. Addressing data sparsity is essential to predict optimal mmWave cell even for those areas where I do not have prior information of mmWave cell camping due to no UE activity. Using joint search method, a macro cell serving a mmWave capable UE can share the UE location to the known optimal mmWave cell for efficient and effective mmWave camping. The data sparsity techniques I study for optimal mmWave cell identification include traditional interpolation techniques, domain knowledge-based custom algorithms, and Artificial Intelligence algorithms.

5.3.1 Applying Traditional Interpolation Techniques to Determine Optimal mmWave Cell

Different spatial interpolation techniques could be leveraged to address the data sparsity challenge in cellular networks. In this work, I leverage some of the most common interpolation techniques including moving average, inverse distance weighted, natural neighbor, and nearest neighbor. These techniques work best if sparse available data is somewhat representative of the whole data or exhibits some degree of spatial correlation [177]. However, in situations where the available data is sparse and non-representative, these methods are likely to perform poorly. A brief description of each technique is shown below:

- Inverse Distance Weighted (IDW) - The simplest form of IDW method is also known as Shepard's method. It is based on the assumption that the distribution of signal samples is strongly correlated with distance. Some of the advantages of simple IDW method include its efficiency and ease of comprehension since it is intuitive. This interpolation works best with evenly distributed points. However, the simple IDW method's disadvantages are that it leads to the production of the "bull's eyes" effect, it is sensitive to measurement outliers, it introduces significant errors in case of non-uniform distribution measurements or unevenly distributed data clusters, the computational error becomes highly significant in the neighborhood of a data point, the calculation of missing value increases proportionally with the number of data points, leading to inefficiency of the method when the number of data points is large.

- Natural Neighbor - The natural neighbor (NaN) interpolation is based on Voronoi decomposition (tessellation) of a set of given points in the plane. The received signal strength value at a particular location is found from a weighted average of N from all available measurements which fall within its 'natural neighborhood'. The

Algorithm 4: Weighted Nearest Neighbors Count (WNNC)

```
Initialize K, GPSaccuracy, BINsize, and Label ;    // Label: vector of mmW PCIs
 with 0 representing Coverage Hole
for Every bin b do
    if b is unlabeled then
        Initialize CUMweight ;    // CUMweight: vector representing cumulative
         weight against each entry of Label vector
        for tier k = 1 to K do
            fetch bins_k from tier k compute weight w_k for index i = 1 to size(Label)
             do
                CUMweight[i] += w_k × count(Label[i] in bins_k)
            end
        end
        if max(CUMweight) > 0 then
            label b with PCI having maximum CUMweight
        else
            Do Nothing;              // Not enough data. Surrounding bins are empty
        end
    else
        Do nothing;
    end
end
```

natural neighbor interpolation method performs well with the non-homogeneous distribution of measurements as well. However, its major drawback is that it cannot find missing signal values that lie outside the convex hull of Voronoi polygons since it requires that the points to be interpolated be in the convex hull of the measurement locations as the Voronoi cells of outer data points are open-ended polygons with an infinite area.

- Nearest Neighbor - The nearest neighbor (NeN) method is also known as proximal interpolation or point sampling. Although the nearest neighbor approach is of low complexity, it results in sharp transitions between the individual signal level zones and increases noise, especially at the boundary of a given area, since it does not consider the influence of the sample data points apart from the nearest neighboring data point.

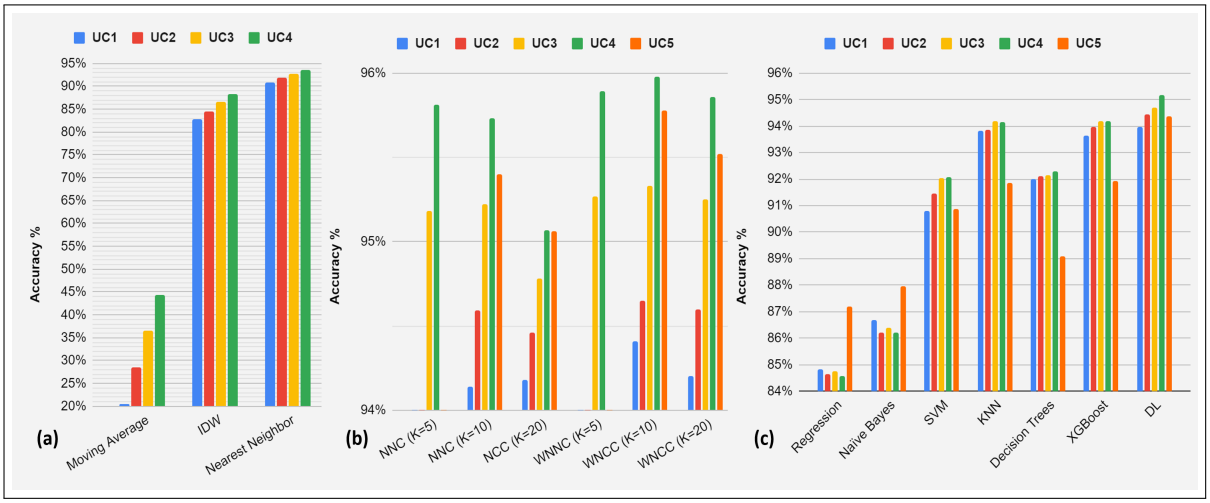

Figure 5.3: Optimal mmWave cell identification for Use Case 1-5 using (a) Traditional Interpolation techniques, (b) Custom Algorithms, (c) Machine Learning.

5.3.2 Custom Algorithms for Optimal mmWave Cell Identification

This subsection discusses two domain knowledge based algorithms to address the sparsity in the UE data fetched from the real networks.

Nearest Neighbors Count (NNC)

NNC fills up the unlabeled bin using the label with the maximum number of occurrences in surrounding tiers (represented as K).

NNC, when used with large number of tiers (K) tends to complete more empty bins, and this is useful for areas with ultra-sparse populations, or in industrial areas with only a few amounts of IoT sensors. However, using a large K might not be favorable under the mmWave environment. High LoS dependency of mmWave frequencies may tend to mislabel the empty bins where the UEs cannot be serviced by the mislabeled cell due to some narrow blockage(s). The problem can also aggravate when using large bin sizes.

Fig. 5.3(b) illustrates the accuracy for all five use cases obtained with K of 5, 10, and 20. Lower K may fail to predict 100% of the target area, and the bar chart is exempted therefore. This effect can be observed with more sparse data (UC1 and UC2). On the contrary, higher K will incorporate more neighboring bins to predict the missing bin at

Table 5.2: Deep learning hyper-parameters for optimal mmWave cell identifier model.

Hyper-parameter Name	Search Range/Value
DNN depth d	{1,2,3,4,5,6}
DNN width w	{5,8,10,12,16}
Activation Function (Hidden Layers)	Relu
Activation Function (Output Layers)	Sigmoid
Optimizer	Adam (Gradient Descent)
Loss Metric	Binary Cross Entropy

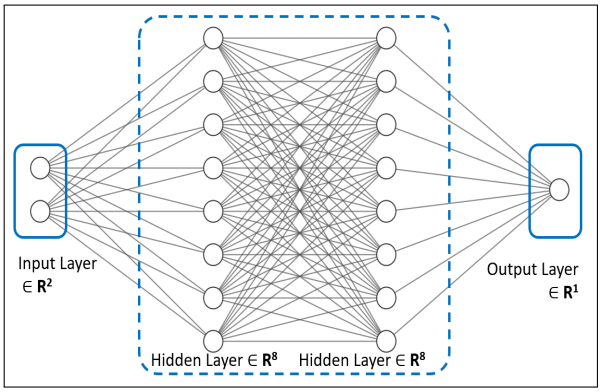

Figure 5.4: Structure of the deep learning based model for predicting optimal mmWave cell for a given UE location.

the cost of lower accuracy. K of 10 yields best accuracy in predicting optimal mmWave cell.

Weighted Nearest Neighbors Count (WNNC)

WNNC addressed the aforementioned problem by applying a unique weight w to each tier around the unlabeled bin. The weight w decreases gradually as I move away from the unlabeled data to the outer tier. Since the GPS accuracy varies globally with certain areas of the planet having lower accuracy than the others [178], the w in addition to bin size encapsulates GPS accuracy as well. The weight w_k assigned to a tier $k \in [1, K]$ can be represented mathematically as:

$$w_k = \frac{\psi}{k \times \omega} \tag{5.2}$$

Table 5.3: Time to build optimal mmWave cell map.

Use Case	Labeled Bins	Unlabeled Bins	Time to Build Optimal mmWave Map		
			Nearest Neighbor	WNCC (K=10)	DL
UC1	50,000	950,000	6.4sec	21Hrs 10mins	12mins
UC2	100,000	900,000	6.9sec	20Hrs 51mins	45mins
UC3	200,000	800,000	7.7sec	19Hrs 23mins	43mins
UC4	300,000	700,000	8.7sec	18Hrs 41mins	48mins
UC5	187,176	812,824	7.4sec	19Hrs 31mins	57mins

where ψ represents GPS accuracy and bin size is denoted by ω. Detail of WNCC has been shown in algorithm 4.

With the similar reason as for NNC, WNNC shows best result for $K = 10$ compared to K of 5 and 20, as shown in Fig. 5.3(b). WNNC yields better results than NNC with the inclusion of domain knowledge assisted weight metric that gives more weightage to lower-tier neighboring bins. Accuracy of as high as 96% can be achieved when using WNNC.

5.3.3 Artificial Intelligence (AI) Assisted Optimal mmWave Cell Identification

This subsection describes how machine learning algorithms can alleviate sparsity issues and can build up a map representing optimal mmWave cell. The available sparse data is scaled and used to train and test several AI techniques for creating a best-performing model for optimal mmWave cell as a function of UE location. After splitting the data into a training and a test dataset, using a range of hyper-parameters, I develop and validate several classification algorithms which include KNN, decision trees, regression, and deep learning-based models. Fig. 5.3 shows the accuracy of the predicted mmWave cell for various machine learning models trained on the same data. Deep learning yields the best results in terms of accuracy. The decision boundary for the aforementioned model is shown in Fig. 5.3.

Deep Neural Network algorithm belongs to a special class of machine learning, called

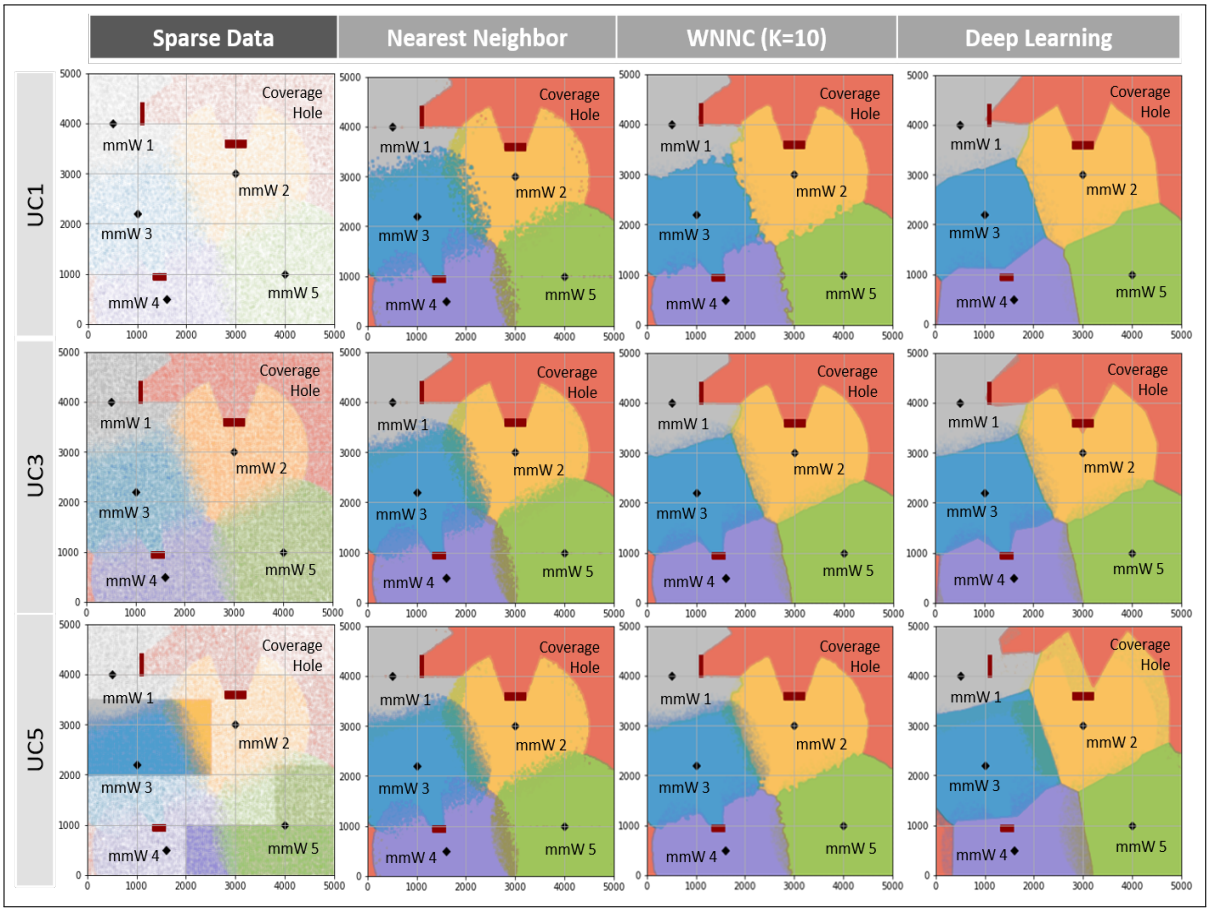

Figure 5.5: Optimal mmWave cell map predicted using Nearest Neighbor, WNNC (K=10), and Deep Learning.

deep learning, and creates a multi-layer perceptron to find the input-output associations. Its basic structure consists of an input layer, output layer, and one or more hidden layers between them, each containing several neurons (or nodes). The number of neurons in the input layer is typically equal to the number of input features, whereas the output layer in the case of a binary classification model consists of a single neuron that holds the prediction output. The number of hidden layers and its neurons are variable and depend on the complexity of the model it is trying to learn. To avoid under or over-fitting, I investigate a variety of deep learning neural network architectures with a range of hyper-parameters as shown in Table 5.2. Our experiments show that a deep learning model with fully connected three hidden layers having 16, 16, and 8 neurons respectively as shown in Fig. 5.4, yields the best results. The model was trained using an epoch size of 200 and a batch size of 10.

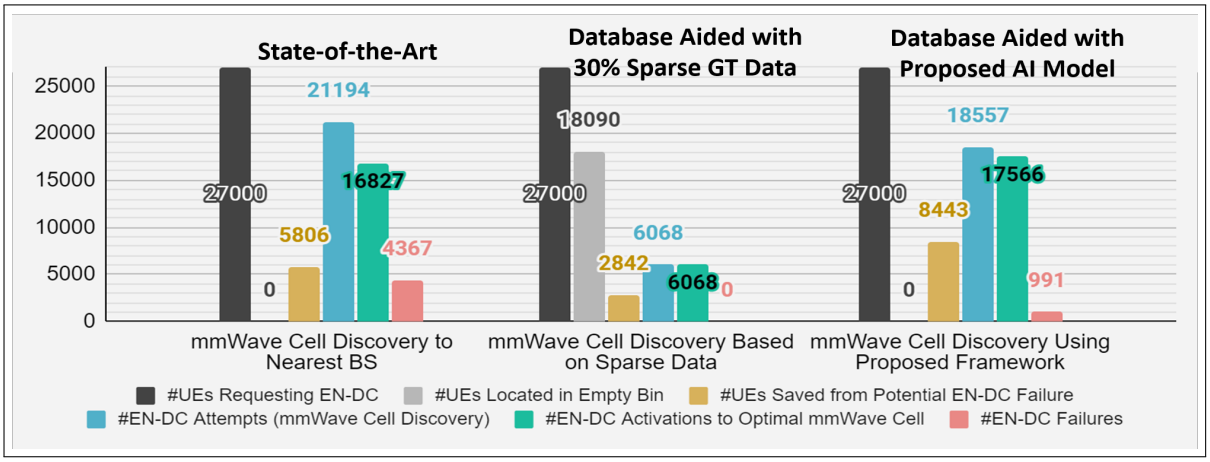

Figure 5.6: Comparison of EN-DC activation KPIs for 5G mmWave cells.

Fig. 5.3(c) illustrates the accuracy of optimal cell prediction increase with the increase in available samples (UC1 to UC4). It also shows that the DL model gives the best result with an accuracy of almost 95%. Deep learning-based models are effective especially for the scenarios where signal reception at different times of the day varies due to different UE mobility and traffic dynamics. As a result, the optimal mmWave cell needs to be continuously tuned with the dynamically changing conditions mentioned above.

Comparison of the accuracy metric obtained from all the approaches in subsection 5.3.1, 5.3.2, and 5.3.3 in Fig. 5.3 highlights WNNC as the best performing approach with accuracy of 96%. This however comes at the cost of high processing and delay where WNNC takes 18 to 21 hours in building the optimal mmWave coverage map. On the contrary, the deep learning model takes 45 minutes to build an optimal mmWave coverage map with high accuracy of 95%.

Fig. 5.5 shows the decision boundary plot for approach that yields the best results in each of the subsection 5.3.1, 5.3.2, and 5.3.3.

5.4 Case Study - EN-DC Activation for mmWave Band

The proposed joint search-based mmWave cell discovery framework is well suited for E-UTRAN New-Radio Dual-Connectivity (EN-DC) activation. As per 3GPP Release

15 specification 37.863 [133], EN-DC allows 5G capable UEs to simultaneously connect to a 4G and 5G BS. EN-DC activation requires UE to first establish a user-plane and control-plane to a 4G mobile network. Later on, UE searches for an optimal 5G BS and establishes a user plane upon successful discovery of a nearby 5G BS. This non-standalone 5G network deployment will help mobile operators to reduce the capital expenditure (CAPEX) and will accelerate the penetration of 5G networks in developing countries. More detail on EN-DC can be found in [149].

The huge resource requirements of bandwidth-hungry applications keeping in view the over-congested high-frequency bands can be addressed by activating EN-DC using mmWave band of 5G cells. EN-DC requires UE to first camp on an LTE cell, herein assumed to be a macro cell having accurate UE information through UE GPS location sharing or using Minimization of Drive Test (MDT). 5G mmWave cell discovery can be enabled through the proposed framework. The historical data from UE traces collected from both the 5G standalone, and EN-DC activated UEs contains the serving mmWave cell information against the UE location. The UE trace data also contains the poor radio link failure data observed due to either NLoS induced signal deterioration or due to a high pathloss situation (where UE and BS distance exceeds cell-range κ). The approaches mentioned in Section 5.3 can be applied to the data collected from the network to identify the optimal 5G mmWave cell. This can help not only in 5G mmWave cell discovery but also for mmWave cell alignment required to maintain reliable communication for mobile UEs.

I run a simulation for 300 EN-DC capable UEs, 70% of which move with a constant velocity of 60km/h using random waypoint model. The system model used is the same as shown earlier in Fig. 5.2(a), where 4G macro BSs act as coverage layer and 5G mmWave cells take the role of the capacity layer to address the needs to bandwidth-hungry applications by activating EN-DC where applicable. UEs already camped on 4G macro cell periodically request EN-DC activation, followed by macro cell initiating

mmWave cell discovery using joint search method where the UE accurate location is shared by the 4G macro cell to the respective 5G mmWave cell. I use the following three approaches to identify optimal 5G mmWave cell before initiating mmWave cell discovery:

- mmWave cell discovery to nearest BS - 4G macro cell directs the 5G mmWave cell located nearest to the candidate UE to establish the EN-DC connection, without taking into consideration the location of the blockage.
- mmWave cell discovery based on sparse data - historical data obtained at the UE location is leveraged to identify the optimal mmWave cell. mmWave cell discovery for EN-DC activation terminates if the UE is located in the bin where prior UE trace data is absent.
- mmWave cell discovery using proposed framework - Optimal 5G mmWave cell coverage map obtained after addressing sparsity on historical data, keeping in view the mmWave coverage hole induced by NLoS condition and large UE-BS distance is used for mmWave cell discovery.

Results in Fig. 5.6 show that the first approach with mmWave cell discovery to the nearest mmWave BS results in a large number of EN-DC attempts (21194), however, successful EN-DC activations are much less due to 4367 EN-DC failures. 4G macro cell in this case is unaware of the blocking locations and the absence of NLoS aware coverage map results in a large number of EN-DC failures. Note that EN-DC failure here refers to the UE with failed mmWave cell discovery or due to UE camping to the sub-optimal mmWave cell. On the other hand, the second approach outcomes zero EN-DC failures, but with very few numbers of EN-DC attempts. This is due to only 30% available bins being labeled i.e., sparsity of 30% considered in this case (similar as UC4). Fig. 5.6 shows that 18090 bins are unlabeled and for the UE in any of the unlabeled bin, the macro cell does not proceed with mmWave cell discovery due to the absence of optimal

mmWave cell information.

Finally, the best EN-DC KPIs are obtained using mmWave cell discovery as per the proposed scheme. Deep learning-assisted optimal mmWave coverage map has been used in this example. Fig. 5.6 shows that when compared to the other two approaches discussed above, the maximum number of EN-DC activations are observed when attempting mmWave cell discovery using our proposed approach. This is due to the interpolation of the sparse data which allows macro BS to effectively predict the optimal mmWave cell against the UE location. Moreover, macro BS avoid unnecessary mmWave cell search attempts due to knowledge of the UE location under coverage hole due to either a) the distance of the UE from mmWave BS being larger than cell-range κ, or b) UE under NLoS due to any blockage in the surrounding area. Approach one which attempts cell discovery to the nearest mmWave BS has only the knowledge of the number of UEs farther from the BS than κ. On the contrary, the proposed scheme knowing the number of UEs out of the configured cell radius κ, along with the NLoS aware coverage map results in an efficient EN-DC activation with just $\sim$5% EN-DC failures (991 failures out of 18557 EN-DC attempts).

5.5 Conclusion

One of the most effective ways to avoid the looming capacity crunch in emerging mobile networks is by efficiently make use of the wide channel mmWave band cells. The joint search method is the most promising cell discovery approach where the high-frequency macro cell aids mmWave cell discovery by sharing the UE location to the nearby mmWave cell. However, the knowledge of optimal mmWave cell is crucial to the success of-f mmWave cell discovery. This is due to the peculiar nature of mmWave cells where signal level deteriorates dramatically when UE goes under NLoS scenario. To address this issue, I propose a joint search-based mmWave cell discovery approach, where UE past traces can be leveraged to build an NLoS aware coverage map. This

map can then aid macro cell to identify the optimal mmWave cell against the given UE location. The optimal mmWave cell map is built while taking into account the data sparsity, a phenomenon common in mobile networks. Results from a mmWave-enable 3GPP-compliant simulator SyntheticNET show that I can predict optimal mmWave cell for cell discovery with an accuracy of 96% using a domain knowledge-based custom WNNC algorithm. Since UE mobility and traffic dynamics may affect signal reception in different times of the day, I demonstrate how deep learning can be used to build the optimal mmWave cell map in much lesser time than the WNNC algorithm, and with the accuracy of 95%.

I also present a case study where the proposed mmWave cell discovery can be utilized to efficiently enforce EN-DC transmissions between the EN-DC capable UEs and the participating 4G macro cells and 5G mmWave cells. Simulation results show that we can enable 17566 EN-DC activations to optimal mmWave cells, while keeping the number of unnecessary mmWave cell discovery attempts due to UE location in the coverage hole, to a minimum.

CHAPTER 6

Conclusion and Future Research Directions

6.1 Conclusion

The dissertation presents the mobility management frameworks in multi-RAT multi-band ultra-dense cellular networks. State-of-the-art 3GPP-based mobility criteria are studied in the light of futuristic mobile networks and user requirements. The panorama of mobility challenges arising in emerging mobile networks implies that if no drastic and timely measures are taken to rethink mobility management for future ultra-dense networks, user mobility management can become the bottleneck in practical deployments of ultra-dense networks despite advances in the hardware design of mmWave and conventional spectrum based small-cells. The dissertation not only presents the first-ever detailed taxonomy on mobility-related 3GPP network parameters and KPIs, but also presents the intricate interplay between the mobility-related network configuration parameters and the affected network KPIs. In addition, the dissertation also explicates a tutorial on 3GPP-based 5G mobility management procedures.

Since the mathematical models and existing network simulators fail to incorporate the realistic mobility management dynamics, this dissertation discusses the development and key attributes of SyntheticNET - the very first Python-based simulator that fully conforms to 3GPP Release 15 5G standard. The development of SyntheticNET is vital to incorporate the futuristic network and traffic modeling scenarios, and the python-based platform allows the effective application of Artificial Intelligence (AI) to various network functionalities.

This dissertation discusses the first-ever intelligent QoE-aware EN-DC triggering scheme by which RLF and mute due to poor RF conditions are minimized. The scheme works

by selecting the best B1 threshold based on insights from deep learning-based 2-stage AI model to predict radio link failure and voice mute. Using SyntheticNET, we show how our proposed scheme can eliminate the RLF and mute occurrences vis-a-vis state-of-the-art approaches i.e., no smart conditioning on EN-DC. The optimal RSRP and SINR thresholds obtained from the presented optimization function help reduce RLF and mute occurrences from 1328 and 3208 cases to zero potential RLF and potential mute cases respectively.

Finally, the dissertation presents a novel framework where real network data can be leveraged to provide database-aided mmWave cell discovery. A case study has also been presented that shows how efficient EN-DC can be activated to 5G mmWave leg keeping in view the out of coverage areas due to blockage. Simulation results on the 3GPP compliant SyntheticNET simulator show the proposed framework outperforms both the state-of-the-art mmWave cell discovery techniques and database-aided real data-oriented cell discovery method in terms of the number of EN-DC activations to optimal mmWave cell.

The results of the frameworks presented in this dissertation illustrate that AI together with domain knowledge has the potential to enable an efficient mobility management system required to achieve the ambitious QoE goals of the futuristic mobile networks.

6.2 Future Research Directions

Now I will discuss a few of the key points related to future research directions:

6.2.1 HO Delay Based SINR Distribution

Current SINR modeling is based on best-server association, however, the UE always camp on the second-best cell prior to HO. This is the result of the HO evaluation process [2] which ensures that the target cell is the best candidate cell for HO. A mobility-

oriented SINR distribution that capture the temporal negative SINR [179] before HO needs to be studied for more realistic throughput estimation.

6.2.2 HO Delay Based Uplink Interference

Current researchers do not consider the practical situation where due to intra-frequency HO delay, high mobility users are closer to the target cell while still being served by the comparatively farther located serving cell. Under those circumstances, high uplink power to achieve target SINR in the serving cell can cause strong temporal interference in the target cell. The issue can be aggravated under highly dense BSs deployed in an impromptu fashion. However, this problem can be tackled by utilizing an eICIC ABS (Almost Blank Subframe) scheme for highly mobile users. A proactive HO trigger can also eliminate the possibility of high uplink RSSI by performing timely HO.

6.2.3 Latency Goals

Another challenging aspect of the small cell deployment is that the small cells are typically not directly connected to the core network and lack Xn or N2 interfaces (for inter-cell communication) which are the real means of coordinating mobility procedures in the macro-cells. The lack of a low latency connection to the core network can contribute to significant HO signaling delays.

6.2.4 Energy-Efficiency

Achieving both UE and network-level energy efficiency is a big challenge for futuristic cellular networks, especially when considering ultra-dense BS deployment and the addition of a wide variety of user devices. Most of the existing energy-saving schemes have a common tenancy; cells are switched ON/OFF reactively in response to changing cell loads. A meritorious effort has been made by Hasan et al. in [57], where authors

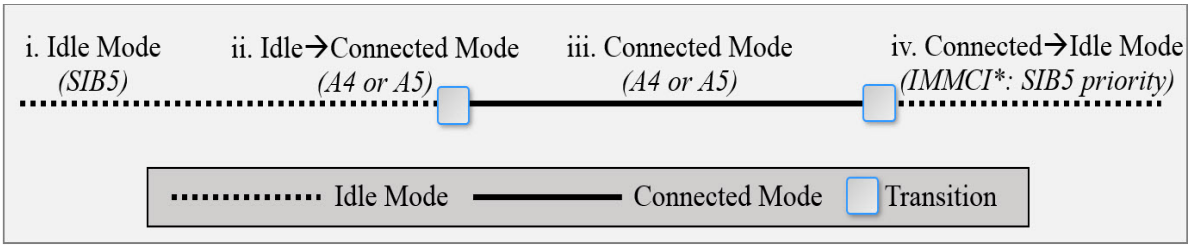

Figure 6.1: Load Balance (LB) opportunities (i, ii, iii, iv) in different stages of 5G UE connection.

proposed the AURORA framework in which the past HO traces are utilized to determine future cell loads. The prediction is then used to proactively schedule small cell sleep cycles. Load balancing is also achieved through the use of an appropriate Cell Individual Offset (CIO).

6.2.5 Smart Intra-Frequency Search

Dense deployment poses challenges for small cell discovery as conventional cellular networks broadcast a neighbor list for the user to learn where to search for potential HO cells. However, such a HO protocol does not scale to the large numbers of neighboring small cells and the underlying network equipment is not designed to rapidly change the neighbor cell lists as small cells come and go.

6.2.6 Smart Inter-Frequency Search

Inter-Frequency (IF) mobility is a vital component of cellular networks but has not got the attention it deserved in the research community. IF-mobility requires event A2 to be triggered, which is followed by the BS to configure measurement gap periodicity to the UE. However, this process interrupts data transmission and reception. This is because UE shifts the radio to measure appropriate IF-cell(s). Futuristic mobile networks with a variety of frequencies ranging from HF to mmWave band may require the UE to undergo an extensive search of available frequencies before initiating a mobility decision. This issue can be aggravated when considering the latency goal of $<$1ms.

6.2.7 Improving Mobility Load Balancing

Mobility Load Balance (MLB) is a vital component of heterogeneous multi-layer cellular networks and are open to the following challenges:

- LB can be achieved at four different instances as shown in Fig. 6.1. It can be triggered through i) idle mode SIB4 configuration, ii) after network access using A4 or A5 measurement report, iii) in connected mode using A4 or A5 measurement report (as configured), iv) when UE is released from connected to idle mode using 3GPP proposed IMMCI (Idle Mode Mobility Control Info). In IMMCI, traffic steering is achieved by varying the idle mode SIB5 priority of the serving or target layer. LB in idle mode is the most optimal as signaling and data interruption associated with connected mode LB can be avoided. Moreover, complexity in parameter configuration and management by IMMCI can be minimized. Research contributions are currently lacking for idle mode load balancing. Similarly, a new variant of IMMCI (SON-based) is needed which can adaptively steer traffic to achieve load balancing under varying load conditions.

- LB detail procedure has not been provided by 3GPP and is left intentionally to vendors for innovation purposes. LB requires the exchange of load information between participating BSs via the Xn interface. However, different vendors have their own proprietary version of LB implementation, thus, inter-vendor BS cannot perform LB due to a mismatch in LB metrics. The existing LTE networks deploy offloading features, where high load cell offload users to another vendor cell without considering its load condition. This can cause service rejection and ping-pong HO conditions. The frequent IF-search will disrupt continuous reception and will result in higher latency. 5G heterogeneous network can assume numerous vendors, and to benefit from the load balancing feature, a standard inter-vendor LB mechanism needs to be devised.

- Cells with smaller footprints will have few serving UEs, and mobility-based ingress and egress of even a single user can have drastic load imbalance among available frequency bands. Hence, ways to achieve proactive LB are mandatory to have fairness and efficient resource utilization.

6.2.8 Mobility in mmWave Networks

mmWave with bandwidth as large as 500MHz is the remedy to the spectrum saturation in the HF band, however, an intrinsic feature of narrow beams can pose serious challenges in supporting mobility in the emerging cellular networks. A few of the main challenges are presented here:

- Simic et al. [180] practically demonstrates mmWave to prove multi-Gbps connectivity but conclude that supporting mobility is a very challenging task due to the outage area of as high as 40% with 90BS/km2 deployment. The reason for the coverage hole is the high diffraction phenomena in mmWaves, and the absence of Non- Line of Sight (NLoS) paths.

- Corner Effect: Indoor areas have cell edge near doors, where the user is more likely to make a sharp turn and hence, time available for HO would be very less especially in the 60GHz mmWave scenario. This issue suggests that some sophisticated techniques, other than conventional methods are required for the HO trigger.

- Current mmWave standards such as IEEE 802.11ad follows the max-RSSI-based approach for UE-BS association, however, this solution appears rudimentary and ineffective for an emerging network with an ultra-dense BS density. There will be chances of an unbalanced number of users per BS, and ping-pong HOs will be highly likely.

- In addition, cell discovery for mobile users is a major challenge due to the absence of Reference Signal (RS) broadcast as in HF bands.

Presently, an overwhelming understanding of the research circle is to use mmWave-cells for static users only. Intricacies of mobility between the beams (of both intra-frequency cells and inter-frequency cells) need to be addressed to support mobility. One possible solution is to come up with a hybrid solution where HF macro-cells with much accurate UE location guide the UEs how, when and to which small cell they need to connect. This is similar to control-data split architecture with mmWave providing data support while UE is under the coverage of macro-cell providing control signals.

6.2.9 Low-Cost Multi-Connectivity

Dual connectivity architecture has been proposed to mitigate mobility management problems in HetNets by allowing UE to connect with the macro-cell for control connectivity as well as simultaneous data connectivity with small-cells. The effect of the user association on dual connectivity performance is an interesting research problem that needs to be investigated in detail. Researchers need to study the gain dual connectivity can yield in terms of HO overhead reduction, synchronization complexity, and radio resource efficiency. Most of the research work addresses reliability and latency goals through multi-connectivity, however, signaling load increment is not addressed. More efficient proposals with special consideration of signaling load need to be devised.

6.2.10 Accurate and Efficient Mobility Prediction

The mobility prediction schemes are seen as a driving force for context-aware cellular networks as they are used to proactively reserve resources, trigger LB, and activate/deactivate small-cells. Few challenges associated with mobility prediction are:

- Users not willing to share location information due to privacy reasons.

- GPS data acquisition consumes user battery and intermittent accessibility requests resulting in signaling or RACH issues (some RACH failure issues cannot be seen in the KPI data).

- Accuracy and reliability of 3GPP proposed Minimization of Drive Test (MDT) feature is needed to be evaluated since a multitude of factors like the GPS error [181], quantization resolution etc. affect the accuracy of the measurements reported by the UE.

- Although human trajectory exhibits high predictable component [56], however, mobility prediction is always bound to have some inaccuracy as can be understood through an example: an office employee may have lunch in a canteen, in a conference room, with colleagues in an outside restaurant, etc. These random variations are almost impossible to predict.

A possible solution can be resource reservation to be done in the multiple neighbors, however, the cost of signaling and available resource for other UEs especially during busy hours needs to be considered.

Bibliography

[1] 3GPP, "Nr; user equipment (ue) procedures in idle mode and in rrc inactive state," Tech. Rep., 2020. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3192

[2] ——, "Nr; radio resource control (rrc); protocol specification," Tech. Rep., 2020. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197

[3] ——, "36.936 - evolved universal terrestrial radio access (e-utra); study on mobile relay," Tech. Rep., 2017.

[4] Ericsson, "Ericsson mobility report," Tech. Rep., 2020. [Online]. Available: https://www.ericsson.com/49da93/assets/local/mobility-report/documents/2020/june2020-ericsson-mobility-report.pdf

[5] C. Systems, "Vni global mobile data traffic forecast update 2017-2021," Tech. Rep., 2017.

[6] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 1, pp. 336–361, 2013.

[7] F. Al-Ogaili and R. M. Shubair, "Millimeter-wave mobile communications for 5g: Challenges and opportunities," in *2016 IEEE International Symposium on Antennas and Propagation (APSURSI)*, 2016, pp. 1003–1004.

[8] W. Kiess, Yuan Xun Gu, S. Thakolsri, M. R. Sama, and S. Beker, "Simplecore: A connectionless, best effort, no-mobility-supporting 5g core architecture," in *2016 IEEE International Conference on Communications Workshops (ICC)*, 2016, pp. 367–372.

[9] N. G. M. Networks, "White paper on 5g vision and requirements v1.0," Tech. Rep., 2015. [Online]. Available: https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf

[10] Pingzhi Fan, "Advances in broadband wireless communications under high-mobility scenarios," *Chinese Science Bulletin*, vol. 59, no. 35, pp. 4974–4975, 2014.

[11] M. Khanfouci, "Distributed mobility management based on centrality for dense 5g networks," in *2017 European Conference on Networks and Communications (EuCNC)*, 2017, pp. 1–6.

[12] X. Gelabert, G. Zhou, and P. Legg, "Mobility performance and suitability of macro cell power-off in lte dense small cell hetnets," in *2013 IEEE 18th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2013, pp. 99–103.

[13] Cisco, "Mme administration guide, staros release 20 (chapter: Mobility management entity overview)," Tech. Rep., 2016. [Online]. Available: https://www.cisco.com

[14] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From lte to 5g for connected mobility," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 156–162, 2017.

[15] S. Batabyal and P. Bhaumik, "Mobility models, traces and impact of mobility on opportunistic routing algorithms: A survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1679–1707, 2015.

[16] S. Thangam and E. Kirubakaran, "A survey on cross-layer based approach for improving tcp performance in multi hop mobile adhoc networks," in *2009 International Conference on Education Technology and Computer*, 2009, pp. 294–298.

[17] E. Spaho, L. Barolli, G. Mino, F. Xhafa, and V. Kolici, "Vanet simulators: A survey on mobility and routing protocols," in *2011 International Conference on Broadband and Wireless Computing, Communication and Applications*, 2011, pp. 1–10.

[18] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in lte-advanced: Key aspects and survey of handover decision algorithms," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 64–91, 2014.

[19] J. Wu and P. Fan, "A survey on high mobility wireless communications: Challenges, opportunities and solutions," *IEEE Access*, vol. 4, pp. 450–476, 2016.

[20] S. Zhao and Q. Wang, "A contextual awareness-learning approach to multi-objective mobility management," in *2017 12th International Conference on Computer Science and Education (ICCSE)*, 2017, pp. 277–282.

[21] S. Adnan, Y. Fu, C. Zhen, N. Junejo, and H. Esmaiel, "Sparse detection with orthogonal matching pursuit in multiuser uplink quadrature spatial modulation mimo system," *IET Communications*, vol. 13, 09 2019.

[22] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Evaluation of adaptive active set management for multi-connectivity in intra-frequency 5g networks," in *2016 IEEE Wireless Communications and Networking Conference*, 2016, pp. 1–6.

[23] J. Stańczak, "Mobility enhancements to reduce service interruption time for lte and 5g," in *2016 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2016, pp. 1–5.

[24] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Mobility modeling and performance evaluation of multi-connectivity in 5g intra-frequency networks," in *2015 IEEE Globecom Workshops (GC Wkshps)*, 2015, pp. 1–6.

[25] D. Öhmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Achieving high availability in wireless networks by inter-frequency multi-connectivity," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–7.

[26] D. Ohmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Impact of mobility on the reliability performance of 5g multi-connectivity architectures," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 1–6.

[27] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Fast cell select for mobility robustness in intra-frequency 5g ultra dense networks," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016, pp. 1–7.

[28] M. Boujelben, S. Ben Rejeb, and S. Tabbane, "A novel mobility-based comp handover algorithm for lte-a / 5g hetnets," in *2015 23rd International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2015, pp. 143–147.

[29] S. Barbera, K. I. Pedersen, C. Rosa, P. H. Michaelsen, F. Frederiksen, E. Shah, and A. Baumgartner, "Synchronized rach-less handover solution for lte heterogeneous networks," in *2015 International Symposium on Wireless Communication Systems (ISWCS)*, 2015, pp. 755–759.

[30] Qualcomm, "Further details of rach-less handover, qualcomm 3gpp tsg ran2 meeting 94," Tech. Rep., 2016.

[31] H. Li and D. Hu, "Mobility prediction based seamless ran-cache handover in hetnet," in *2016 IEEE Wireless Communications and Networking Conference*, 2016, pp. 1–7.

[32] M. Chen, Y. Hao, L. Hu, K. Huang, and V. K. N. Lau, "Green and mobility-aware caching in 5g networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 8347–8361, 2017.

[33] P. Kela, X. Gelabert, J. Turkka, M. Costa, K. Heiska, K. Leppänen, and C. Qvarfordt, "Supporting mobility in 5g: A comparison between massive mimo and continuous ultra dense networks," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.

[34] L. L. Vy, L. Tung, and B. P. Lin, "Big data and machine learning driven handover management and forecasting," in *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2017, pp. 214–219.

[35] U. Karneyenka, K. Mohta, and M. Moh, "Location and mobility aware resource management for 5g cloud radio access networks," in *2017 International Conference on High Performance Computing Simulation (HPCS)*, 2017, pp. 168–175.

[36] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007. [Online]. Available: https://science.sciencemag.org/content/315/5814/972

[37] B. L. Dang, R. V. Prasad, I. Niemegeers, M. G. Larrode, and A. M. J. Koonen, "Toward a seamless communication architecture for in-building networks at the 60 ghz band," in *Proceedings. 2006 31st IEEE Conference on Local Computer Networks*, 2006, pp. 300–307.

[38] N. Pleros, K. Tsagkaris, and N. D. Tselikas, "A moving extended cell concept for seamless communication in 60 ghz radio-over-fiber networks," *IEEE Communications Letters*, vol. 12, no. 11, pp. 852–854, 2008.

[39] F. A. Hossain and A. M. Chowdhury, "User mobility prediction based handoff scheme for 60 ghz radio over fiber network," in *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, 2014, pp. 557–562.

[40] M. Bagaa, T. Taleb, and A. Ksentini, "Efficient tracking area management framework for 5g networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4117–4131, 2016.

[41] S. Hailu and M. Säily, "Hybrid paging and location tracking scheme for inactive 5g ues," in *2017 European Conference on Networks and Communications (EuCNC)*, 2017, pp. 1–6.

[42] Nokia, "Discussion of rrc states in nr, 3gpp r2 wg," Tech. Rep., 2016.

[43] S. Ikeda, N. Kami, and T. Yoshikawa, "Adaptive mobility management in cellular networks with multiple model-based prediction," in *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2013, pp. 473–478.

[44] N. B. Prajapati and D. R. Kathiriya, "Dynamic location area planning in cellular network using apriori algorithm," in *2015 International Conference on Industrial Instrumentation and Control (ICIC)*, 2015, pp. 660–662.

[45] "Preface," in *Data Mining (Third Edition)*, third edition ed., ser. The Morgan Kaufmann Series in Data Management Systems, J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. xxiii–xxix. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123814791000204

[46] O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "Energy efficient inter-frequency small cell discovery in heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7122–7135, 2016.

[47] D. Calabuig, S. Barmpounakis, S. Gimenez, A. Kousaridas, T. R. Lakshmana, J. Lorca, P. Lunden, Z. Ren, P. Sroka, E. Ternon, V. Venkatasubramanian, and M. Maternia, "Resource and mobility management in the network layer of 5g cellular ultra-dense networks," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 162–169, 2017.

[48] M. Boujelben, S. Ben Rejeb, and S. Tabbane, "A novel green handover self-optimization algorithm for lte-a / 5g hetnets," in *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2015, pp. 413–418.

[49] J. Ye, Y. He, X. Ge, and M. Chen, "Energy efficiency analysis of 5g ultra-dense networks based on random way point mobility models," in *2016 19th International*

Symposium on Wireless Personal Multimedia Communications (WPMC), 2016, pp. 177–182.

[50] A. S. Mubarak, O. A. Omer, H. Esmaiel, and U. S. Mohamed, "Geometry aware scheme for initial access and control of mmwave communications in dynamic environments," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019*, A. E. Hassanien, K. Shaalan, and M. F. Tolba, Eds. Cham: Springer International Publishing, 2020, pp. 760–769.

[51] M. Giordani and M. Zorzi, "Improved user tracking in 5g millimeter wave mobile networks via refinement operations," in *2017 16th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, 2017, pp. 1–8.

[52] A. Abdelreheem, E. M. Mohamed, and H. Esmaiel, "Location-based millimeter wave multi-level beamforming using compressive sensing," *IEEE Communications Letters*, vol. 22, no. 1, pp. 185–188, 2018.

[53] A. S. Mubarak, H. Esmaiel, and E. M. Mohamed, "Lte/wi-fi/mmwave ran-level interworking using 2c/u plane splitting for future 5g networks," *IEEE Access*, vol. 6, pp. 53 473–53 488, 2018.

[54] J. S. Kim, W. J. Lee, and M. Y. Chung, "A multiple beam management scheme on 5g mobile communication systems for supporting high mobility," in *2016 International Conference on Information Networking (ICOIN)*, 2016, pp. 260–264.

[55] B. Lannoo, D. Colle, M. Pickavet, and P. Demeester, "Radio-over-fiber-based solution to provide broadband internet access to train passengers [topics in optical communications]," *Communications Magazine, IEEE*, vol. 45, pp. 56 – 62, 03 2007.

[56] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010. [Online]. Available: https://science.sciencemag.org/content/327/5968/1018

[57] H. Farooq, A. Asghar, and A. Imran, "Mobility prediction-based autonomous proactive energy saving (aurora) framework for emerging ultra-dense networks," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 958–971, 2018.

[58] A. Mohamed, O. Onireti, S. A. Hoseinitabatabaei, M. Imran, A. Imran, and R. Tafazolli, "Mobility prediction for handover management in cellular networks

with control/data separation," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 3939–3944.

[59] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Predictive and core-network efficient rrc signalling for active state handover in rans with control/data separation," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1423–1436, 2017.

[60] D. Zhang, D. Zhang, H. Xiong, L. T. Yang, and V. Gauthier, "Nextcell: Predicting location using social interplay from cell phone traces," *IEEE Transactions on Computers*, vol. 64, no. 2, pp. 452–463, 2015.

[61] A. Mohamed, M. A. Imran, P. Xiao, and R. Tafazolli, "Memory-full context-aware predictive mobility management in dual connectivity 5g networks," *IEEE Access*, vol. 6, pp. 9655–9666, 2018.

[62] J. Yang, C. Dai, and Z. Ding, "A scheme of terminal mobility prediction of ultra dense network based on svm," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, 2017, pp. 837–842.

[63] X. Chen, F. Mériaux, and S. Valentin, "Predicting a user's next cell with supervised learning based on channel states," in *2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2013, pp. 36–40.

[64] S. Premchaisawatt and N. Ruangchaijatupon, "Enhancing indoor positioning based on partitioning cascade machine learning models," in *2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2014, pp. 1–5.

[65] N. Sinclair, D. Harle, I. A. Glover, J. Irvine, and R. C. Atkinson, "An advanced som algorithm applied to handover management within lte," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1883–1894, 2013.

[66] Libo Song, D. Kotz, Ravi Jain, and Xiaoning He, "Evaluating next-cell predictors with extensive wi-fi mobility data," *IEEE Transactions on Mobile Computing*, vol. 5, no. 12, pp. 1633–1649, 2006.

[67] Y. Cheng, Y. Qiao, and J. Yang, "An improved markov method for prediction of user mobility," in *2016 12th International Conference on Network and Service Management (CNSM)*, 2016, pp. 394–399.

[68] Y. Qiao, J. Yang, H. He, Y. Cheng, and Z. Ma, "User location prediction with energy efficiency model in the long term-evolution network," *International Journal of Communication Systems*, vol. 29, no. 14, pp. 2169–2187, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/dac.2909

[69] A. Li, Q. Lv, Y. Qiao, and J. Yang, "Improving mobility prediction performance with state based prediction method when the user departs from routine," in *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, 2016, pp. 1–7.

[70] G. Karagiannis, A. Jamakovic, K. Briggs, M. Karimzadeh, C. Parada, M. I. Corici, T. Taleb, A. Edmonds, and T. M. Bohnert, "Mobility and bandwidth prediction in virtualized lte systems: Architecture and challenges," in *2014 European Conference on Networks and Communications (EuCNC)*, 2014, pp. 1–5.

[71] M. Karimzadeh, Z. Zhao, L. Hendriks, R. de O. Schmidt, S. la Fleur, H. van den Berg, A. Pras, T. Braun, and M. J. Corici, "Mobility and bandwidth prediction as a service in virtualized lte systems," in *2015 IEEE 4th International Conference on Cloud Networking (CloudNet)*, 2015, pp. 132–138.

[72] P. Fazio, M. Tropea, F. D. Rango, and M. Voznak, "Pattern prediction and passive bandwidth management for hand-over optimization in qos cellular networks with vehicular mobility," *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2809–2824, 2016.

[73] A. Hadachi, O. Batrashev, A. Lind, G. Singer, and E. Vainikko, "Cell phone subscribers mobility prediction using enhanced markov chain algorithm," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 2014, pp. 1049–1054.

[74] H. Farooq and A. Imran, "Spatiotemporal mobility prediction in proactive self-organizing cellular networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 370–373, 2017.

[75] A. Ben Cheikh, M. Ayari, R. Langar, G. Pujolle, and L. A. Saidane, "Optimized handoff with mobility prediction scheme using hmm for femtocell networks," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 3448–3453.

[76] A. Ben Cheikh, M. Ayari, R. Langar, and L. A. Saidane, "Ohmp-cac: Optimized handoff scheme based on mobility prediction and qos constraints for femtocell networks," in *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2016, pp. 936–941.

[77] Wee-Seng Soh and H. S. Kim, "Dynamic bandwidth reservation in cellular networks using road topology based mobility predictions," in *IEEE INFOCOM 2004*, vol. 4, 2004, pp. 2766–2777 vol.4.

[78] M. R. Tabany and C. G. Guy, "A mobility prediction scheme of lte/lte-a femtocells under different velocity scenarios," in *2015 IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)*, 2015, pp. 318–323.

[79] N. Hoang, N. Nguyen, and K. Sripimanwat, "Cell selection schemes for femtocell-to-femtocell handover deploying mobility prediction and downlink capacity monitoring in cognitive femtocell networks," in *TENCON 2014 - 2014 IEEE Region 10 Conference*, 2014, pp. 1–5.

[80] L. Xia, L. . Jiang, and C. He, "A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method," in *2007 IEEE International Conference on Communications*, 2007, pp. 5665–5670.

[81] H. Song, X. Fang, and L. Yan, "Handover scheme for 5g c/u plane split heterogeneous network in high-speed railway," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 9, pp. 4633–4646, 2014.

[82] B. Li, H. Zhang, and H. Lu, "User mobility prediction based on lagrange's interpolation in ultra-dense networks," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016, pp. 1–6.

[83] A. Merwaday, Güvenç, W. Saad, A. Mehbodniya, and F. Adachi, "Sojourn time-based velocity estimation in small cell poisson networks," *IEEE Communications Letters*, vol. 20, no. 2, pp. 340–343, 2016.

[84] A. Merwaday and Güvenç, "Handover count based velocity estimation and mobility state detection in dense hetnets," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 4673–4688, 2016.

[85] A. Klein, A. Rauch, R. R. Sattiraju, and H. D. Schotten, "Achievable performance gains using movement prediction and advanced 3d system modeling," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, 2014, pp. 1–5.

[86] Wee-Seng Soh and H. S. Kim, "Qos provisioning in cellular networks based on mobility prediction techniques," *IEEE Communications Magazine*, vol. 41, no. 1,

pp. 86–92, 2003.

[87] W. . Soh and H. S. Kim, "A predictive bandwidth reservation scheme using mobile positioning and road topology information," *IEEE/ACM Transactions on Networking*, vol. 14, no. 5, pp. 1078–1091, 2006.

[88] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yacoub, "Planning wireless cellular networks of future: Outlook, challenges and opportunities," *IEEE Access*, vol. 5, pp. 4821–4845, 2017.

[89] H. Chung, J. Kim, Gosan Noh, Bing Hui, I. Kim, Youngmin Choi, Changseob Choi, Myongsik Lee, and Dongha KimDongha Kim, "From architecture to field trial: A millimeter wave based mhn system for hst communications toward 5g," in *2017 European Conference on Networks and Communications (EuCNC)*, 2017, pp. 1–5.

[90] J. Heinonen, P. Korja, T. Partti, H. Flinck, and P. Pöyhönen, "Mobility management enhancements for 5g low latency services," in *2016 IEEE International Conference on Communications Workshops (ICC)*, 2016, pp. 68–73.

[91] H. Wang, S. Chen, M. Ai, and H. Xu, "Localized mobility management for 5g ultra dense network," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8535–8552, 2017.

[92] A. S. Mubarak, O. A. Omer, H. Esmaiel, and U. S. Mohamed, "Backhaul overhead traffic reduction in dense mmwave heterogeneous networks towards 5g cellular systems," in *2019 36th National Radio Science Conference (NRSC)*, 2019, pp. 234–241.

[93] Y. Bi, G. Han, C. Lin, M. Guizani, and X. Wang, "Mobility management for intro/inter domain handover in software-defined networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1739–1754, 2019.

[94] M. Manalastas, H. Farooq, S. M. Asad Zaidi, and A. Imran, "Where to go next?: A realistic evaluation of ai-assisted mobility predictors for hetnets," in *2020 IEEE 17th Annual Consumer Communications Networking Conference (CCNC)*, 2020, pp. 1–6.

[95] F. B. Mismar and B. L. Evans, "Partially blind handovers for mmwave new radio aided by sub-6 ghz lte signaling," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2018, pp. 1–5.

[96] H. Ryden, J. Berglund, M. Isaksson, R. Cöster, and F. Gunnarsson, "Predicting strongest cell on secondary carrier using primary carrier data," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2018, pp. 137–142.

[97] S. Zang, W. Bao, P. L. Yeoh, B. Vucetic, and Y. Li, "Managing vertical handovers in millimeter wave heterogeneous networks," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1629–1644, 2019.

[98] S. Khunteta and A. K. R. Chavva, "Deep learning based link failure mitigation," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 806–811.

[99] S. M. Asad Zaidi, M. Manalastas, A. Abu-Dayya, and A. Imran, "Ai-assisted rlf avoidance for smart en-dc activation," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[100] Z. Ali, N. Baldo, J. Mangues-Bafalluy, and L. Giupponi, "Machine learning based handover management for improved qoe in lte," in *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, 2016, pp. 794–798.

[101] S. A. R. Zaidi, A. Imran, D. C. McLernon, and M. Ghogho, "Characterizing coverage and downlink throughput of cloud empowered hetnets," *IEEE Communications Letters*, vol. 19, no. 6, pp. 1013–1016, June 2015.

[102] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Predictive and core-network efficient rrc signalling for active state handover in rans with control/data separation," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1423–1436, March 2017.

[103] O. Onireti, A. Imran, and M. A. Imran, "Coverage, capacity, and energy efficiency analysis in the uplink of mmwave cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 3982–3997, May 2018.

[104] O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "Energy efficient inter-frequency small cell discovery in heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7122–7135, Sep. 2016.

[105] H. Farooq, A. Asghar, and A. Imran, "Mobility prediction-based autonomous proactive energy saving (aurora) framework for emerging ultra-dense networks,"

IEEE Transactions on Green Communications and Networking, vol. 2, no. 4, pp. 958–971, Dec 2018.

[106] A. Asghar, H. Farooq, and A. Imran, "Concurrent optimization of coverage, capacity, and load balance in hetnets through soft and hard cell association parameters," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8781–8795, Sep. 2018.

[107] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to Empower SON with Big Data for Enabling 5G," *IEEE Network*, 2014.

[108] A. Asghar, H. Farooq, and A. Imran, "Self-healing in emerging cellular networks: Review, challenges, and research directions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 1682–1709, thirdquarter 2018.

[109] Syed Muhammad Asad Zaidi, Marvin Manalastas, Hasan Farooq, and A. Imran, "Mobility Management in 5G and Beyond: A Survey and Outlook," *IEEE ACCESS*, 2020.

[110] 3GPP, "Technical Report 21.915 Release 15 Description," Tech. Rep., 2019.

[111] Matlab, "Why Use MATLAB and Simulink for 5G?" Tech. Rep. [Online]. Available: https://www.mathworks.com/solutions/wireless-communications/5g.html

[112] ns 3, "mmWave Cellular Network Simulator," Tech. Rep. [Online]. Available: https://omnetpp.org/

[113] OMNeT++, "OMNeT++: Discrete Event Simulator," Tech. Rep. [Online]. Available: https://apps.nsnam.org/app/mmwave/

[114] OPNET, "OPNET: Optimum Network Performance," Tech. Rep. [Online]. Available: https://www.openairinterface.org/

[115] OpenAirInterface, "OpenAirInterface: 5G Software Alliance for Democratising Wireless Innovation," Tech. Rep. [Online]. Available: http://opnetprojects.com/opnet-simulator/

[116] J. Baek, J. Bae, Y. Kim, J. Lim, E. Park, J. Lee, G. Lee, S. I. Han, C. Chu,

and Y. Han, "5G K-Simulator of Flexible, Open, Modular (FOM) Structure and Web-based 5G K-SimPlatform," in *IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2019.

[117] M. K. Muller, F. Ademaj, T. Dittrich, A. Fastenbauer, B. R. Elbal, A. Nabavi, L. Nagel, S. Schwarz, and M. Rupp, "Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, 2018.

[118] S. Sun, G. R. MacCartney, and T. S. Rappaport, "A novel millimeter-wave channel simulator and applications for 5G wireless communications," in *IEEE International Conference on Communications (ICC)*, 2017.

[119] N. Mohsen and K. S. Hassan, "C-RAN simulator: A tool for evaluating 5G cloud-based networks system-level performance," in *IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2015.

[120] T. Dominguez-Bolano, J. Rodriguez-Pineiro, J. A. Garcia-Naya, and L. Castedo, "The GTEC 5G link-level simulator," in *International Workshop on Link- and System Level Simulations (IWSLS)*, 2016.

[121] X. Wang, Y. Chen, and Z. Mai, "A Novel Design of System Level Simulator for Heterogeneous Networks," in *IEEE Globecom Workshops (GC Wkshps)*, 2017.

[122] V. V. Diaz and D. Marcano Aviles, "A Path Loss Simulator for the 3GPP 5G Channel Models," in *IEEE International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, 2018.

[123] Ke Guan, Bo Ai, Thomas Kumer, Ruisi He, Andreas Moller, and Zhangdui Zhong, "Integrating composite urban furniture into ray-tracing simulator for 5G small cells and outdoor device-to-device communications," in *European Conference on Antennas and Propagation (EuCAP)*, 2016.

[124] Y. Wang, J. Xu, and L. Jiang, "Challenges of System-Level Simulations and Performance Evaluation for 5G Wireless Networks," *IEEE Access*, vol. 2, pp. 1553–1561, 2014.

[125] S. M. A. Usama Masood, H. Farooq and A. Imran, "A Machine Learning based 3D Propagation Model for Intelligent Future Cellular Networks," in *IEEE Globecomm*, 2019.

[126] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic Traffic Simulation using SUMO," in *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. [Online]. Available: https://elib.dlr.de/124092/

[127] O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "Impact of positioning error on achievable spectral efficiency in database-aided networks," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[128] H. Farooq and A. Imran, "Spatiotemporal Mobility Prediction in Proactive Self-Organizing Cellular Networks," *IEEE Communications Letters*, vol. 21, no. 2, pp. 370–373, Feb 2017.

[129] S. M. et al., "Leveraging mobility and content caching for proactive load balancing in heterogeneous cellular networks," *Transaction on Emerging Telecommunications Technologies*, September 2019.

[130] A. Zoha, A. Saeed, H. Farooq, A. Rizwan, A. Imran, and M. A. Imran, "Leveraging intelligence from network cdr data for interference aware energy consumption minimization," *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1569–1582, July 2018.

[131] S. M. A. Marvin Manalastas, H. Farooq and A. Imran, "Where to Go Next? : A Realistic Evaluation of AI-Assisted Mobility Predictors for Heterogeneous Networks (HetNets)," in *IEEE CCNC*, 2019.

[132] Ericsson, "Ericsson Mobility Report: 5G subscriptions to top 2.6 billion by end of 2025," Tech. Rep., 2019.

[133] 3GPP, "37.863 - E-UTRA (Evolved Universal Terrestrial Radio Access) - NR Dual Connectivity (EN-DC) of LTE 1 Down Link (DL) / 1 Up Link (UL) and 1 NR band," Tech. Rep., 2019.

[134] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Mobility modeling and performance evaluation of multi-connectivity in 5g intra-frequency networks," in *2015 IEEE Globecom Workshops (GC Wkshps)*, 2015, pp. 1–6.

[135] D. Öhmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Achieving high availability in wireless networks by inter-frequency multi-connectivity," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–7.

[136] D. S. Wickramasuriya, C. A. Perumalla, K. Davaslioglu, and R. D. Gitlin, "Base station prediction and proactive mobility management in virtual cells using recurrent neural networks," in *2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON)*, 2017, pp. 1–6.

[137] S. Kang and S. Bahk, "Analysis of dual connectivity gain in terms of delay and throughput," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 2018, pp. 1218–1220.

[138] C. Wang, Z. Zhao, Q. Sun, and H. Zhang, "Deep learning-based intelligent dual connectivity for mobility management in dense network," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–5.

[139] Y. Wu, X. Yang, L. P. Qian, H. Zhou, X. Shen, and M. K. Awad, "Optimal dual-connectivity traffic offloading in energy-harvesting small-cell networks," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 1041–1058, 2018.

[140] N. H. Mahmood, M. Lopez, D. Laselva, K. Pedersen, and G. Berardinelli, "Reliability oriented dual connectivity for urllc services in 5g new radio," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, 2018, pp. 1–6.

[141] M. Centenaro, D. Laselva, J. Steiner, K. Pedersen, and P. Mogensen, "Resource-efficient dual connectivity for ultra-reliable low-latency communication," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–5.

[142] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, and S. Alraih, "Dynamic handover control parameters for lte-a/5g mobile communications," in *2018 Advances in Wireless and Optical Communications (RTUWO)*, 2018, pp. 39–44.

[143] Y. Mal, J. Chen, and H. Lin, "Mobility robustness optimization based on radio link failure prediction," in *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2018, pp. 454–457.

[144] M. T. Nguyen, S. Kwon, and H. Kim, "Mobility robustness optimization for handover failure reduction in lte small-cell networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4672–4676, 2018.

[145] M.-h. Song, S.-H. Moon, and S.-J. Han, "Self-optimization of handover parameters for dynamic small-cell networks," *Wireless Communications and*

Mobile Computing, vol. 15, no. 11, pp. 1497–1517, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcm.2439

[146] J. Puttonen, J. Kurjenniemi, and O. Alanen, "Radio problem detection assisted rescue handover for lte," in *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2010, pp. 1752–1757.

[147] S. K. Srivastava, M. R. Kanagarathinam, S. Diggi, and H. Natarajan, "Cleh — cross layer enhanced handover for ims sessions," in *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2018, pp. 1–4.

[148] A. Łukowa and V. Venkatasubramanian, "Performance of strong interference cancellation in flexible ul/dl tdd systems using coordinated muting, scheduling and rate allocation," in *2016 IEEE Wireless Communications and Networking Conference*, 2016, pp. 1–7.

[149] S. M. A. Zaidi, M. Manalastas, A. Abu-Dayya, and A. Imran, "Ai-assisted rlf avoidance for smart en-dc activation," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[150] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) Protocol Specification," Tech. Rep., 2016.

[151] 3GPP, "3GPP TS 23.501 - System Architecture for the 5G System," Tech. Rep., 2020.

[152] S. M. A. Zaidi, M. Manalastas, H. Farooq, and A. Imran, "SyntheticNET: A 3GPP Compliant Simulator for AI Enabled 5G and Beyond," *IEEE Access*, pp. 1–1, 2020.

[153] Cisco, "Cisco Annual Internet Report (2018–2023) White Paper," Tech. Rep., 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[154] Ericsson, "Ericsson Mobility Report November 2020," Tech. Rep., 2020. [Online]. Available: https://www.ericsson.com/4adc87/assets/local/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf

[155] 3GPP, "38.717 Technical Report Rel-17 NR inter-band Carrier Aggregation for 5 bands DL with x bands UL (x=1, 2)," Tech. Rep., 2020.

[156] A. Mazin, M. Elkourdi, and R. D. Gitlin, "Comparative performance analysis of beam sweeping using a deep neural net and random starting point in mmwave 5g new radio," in *2018 9th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2018, pp. 451–456.

[157] J. Fan, L. Han, X. Luo, Y. Zhang, and J. Joung, "Beamwidth design for beam scanning in millimeter-wave cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 1111–1116, 2020.

[158] B. Yin, Y. Chen, Z. Zhang, M. Wang, and S. Sun, "Beam discovery signal-based beam selection in millimeter wave heterogeneous networks," *IEEE Access*, vol. 6, pp. 16 314–16 323, 2018.

[159] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, S. S. Panwar, and T. S. Rappaport, "Directional cell discovery in millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6664–6678, 2015.

[160] Y. Kim, H. Lee, and Y. S. Cho, "Fast initial access technique for millimeter-wave systems using retrodirective arrays," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 834–836.

[161] C. Liu, M. Li, S. V. Hanly, P. Whiting, and I. B. Collings, "Millimeter-wave small cells: Base station discovery, beam alignment, and system design challenges," *IEEE Wireless Communications*, vol. 25, no. 4, pp. 40–46, 2018.

[162] Y. Yang, H. S. Ghadikolaei, C. Fischione, M. Petrova, and K. W. Sung, "Reducing initial cell-search latency in mmwave networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018, pp. 686–691.

[163] S. Habib, S. A. Hassan, A. A. Nasir, and H. Mehrpouyan, "Millimeter wave cell search for initial access: Analysis, design, and implementation," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2017, pp. 922–927.

[164] M. Jasim and N. Ghani, "Fast beam discovery for mmwave cellular networks," in *2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON)*, 2017, pp. 1–6.

[165] A. Mazin, M. Elkourdi, and R. D. Gitlin, "Accelerating beam sweeping in mmwave standalone 5g new radios using recurrent neural networks," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–4.

[166] I. Filippini, V. Sciancalepore, F. Devoti, and A. Capone, "Fast cell discovery in mm-wave 5g networks with context information," *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1538–1552, 2018.

[167] R. Parada and M. Zorzi, "Cell discovery based on historical user's location in mmwave 5g," in *European Wireless 2017; 23th European Wireless Conference*, 2017, pp. 1–6.

[168] H. Soleimani, R. Parada, S. Tomasin, and M. Zorzi, "Statistical approaches for initial access in mmwave 5g systems," in *European Wireless 2018; 24th European Wireless Conference*, 2018, pp. 1–6.

[169] A. S. Marcano and H. L. Christiansen, "Macro cell assisted cell discovery method for 5g mobile networks," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016, pp. 1–5.

[170] A. S. A. Mubarak, E. M. Mohamed, and H. Esmaiel, "Millimeter wave beamforming training, discovery and association using wifi positioning in outdoor urban environment," in *2016 28th International Conference on Microelectronics (ICM)*, 2016, pp. 221–224.

[171] C. Sun, J. Zhang, X. Zhang, and W. Wang, "Macro-assisted millimeter-wave small cell discovery in ultra dense wireless network," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2017, pp. 545–551.

[172] A. Prasad, P. Lunden, K. Valkealahti, M. Moisio, and M. A. Uusitalo, "Network assisted small cell discovery in multi-layer and mmwave networks," in *2015 IEEE Conference on Standards for Communications and Networking (CSCN)*, 2015, pp. 118–123.

[173] R. Zia-ul-Mustafa and S. A. Hassan, "Machine learning-based context aware sequential initial access in 5g mmwave systems," in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.

[174] 3GPP, "3GPP TS 36.304 - Evolved Universal Terrestrial Radio Access (E-UTRA)

User Equipment (UE) procedures in idle mode," Tech. Rep., 2021.

[175] "everythingRF ksf410.a," https://www.everythingrf.com/products/all-antennas/taoglas/741-318-ksf410-a, accessed: 2021-04-08.

[176] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[177] H. N. Qureshi, A. Imran, and A. Abu-Dayya, "Enhanced mdt-based performance estimation for ai driven optimization in future cellular networks," *IEEE Access*, vol. 8, pp. 161 406–161 426, 2020.

[178] National Transportation Safety Board (NSTB), "Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report," Tech. Rep., 2017. [Online]. Available: https://www.nstb.tc.faa.gov/reports/PAN96_0117.pdf

[179] S. M. Asad Zaidi and A. Imran, "Effect of mobility and cell density on sinr in emerging ultra-dense cellular networks," *IEEE Sensors (Submitted)*, 2021.

[180] L. Simic, S. Panda, J. Riihijarvi, and P. Mahonen, "Coverage and robustness of mm-wave urban cellular networks: Multi-frequency hetnets are the 5g future," in *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2017, pp. 1–9.

[181] I. Akbari, O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "How reliable is mdt-based autonomous coverage estimation in the presence of user and bs positioning error?" *IEEE Wireless Communications Letters*, vol. 5, no. 2, pp. 196–199, 2016.