“Development and Utilisation of Predictive Modelling Tools for Optimising the Operations of Future Cellular Networks”

Ali Rizwan

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

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

December 2024

Abstract

With the advent of ultra-densification and the proliferation of diverse underlying technologies, 5G and beyond networks promise unprecedented capabilities but also introduce significant challenges in network management due to the manifold increase in complexity. Two such key challenges are network resources optimization and network maintenance. It is challenging to perform these tasks manually, with 5G and beyond networks it does not remain viable at all. Self-aware solutions like Self-Organizing Networks (SON) have, therefore, been proposed and well explored in research to address these challenges.

However, the legacy SON, still relies on the predefined instruction sets making them inelastic to network changes. On the other hand, they use the results from field tests or customers complaints for the identification of network issues which cause unnecessary delays and make existing SON reactive. But, to cater for the exponential increase in network complexity and diverse use cases in 5G and beyond networks, SON needs to be more intelligent, adaptable and autonomous. They are required to be proactive rather than reactive. Artificial intelligence can play a crucial role here and machine learning can equip SON with this intelligence by exploiting the hidden patterns in the real network data. Another advantage of machine learning is that it also brings prediction capacity important for making SON proactive.

Artificial intelligence, particularly machine learning, offers the potential to transform SON into proactive systems by extracting actionable insights from network data and enabling predictive capabilities. This research focuses on leveraging machine learning to enhance two key SON functions: self-optimization and self-healing.

This study begins by exploring and classifying various data types generated within the network, highlighting their potential roles in wireless cellular networks (WCN). A review of existing and potential use cases for SON functions and machine learning-based approaches provides the foundation for this work. For self-optimization, a Support Vector Machine (SVM) model is developed to predict internet traffic loads using Call Detail Records (CDR), achieving up to 91% prediction accuracy. Additionally, a novel machine learning model leveraging Geohash global indexing predicts users' next locations with approximately 95% accuracy, marking a significant contribution to mobility prediction.

To enable self-healing, a hybrid machine learning scheme is proposed. Using CDR data, network cells are grouped based on performance via K-means clustering. Subsequently, an SVM classifier is employed to categorize cell performance with 98% accuracy. The traffic prediction model for self-optimization is further refined through a Support Vector Regression (SVR) approach, achieving 97% prediction accuracy. The predictive capabilities of the model contribute to energy savings of up to eightfold, underscoring its practical impact.

By integrating prognostics and self-aware systems into SON, this research demonstrates a pathway to achieve self-optimization and self-healing in 5G and beyond networks, laying the groundwork for sustainable, intelligent network management.

Statement of Originality

University of Glasgow
College of Science & Engineering

Name: Ali Rizwan
Registration Number: XXXXXXXXX

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

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

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

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

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

Signature: ....

Date:..................

List of Publications

1. Zoha, A., Saeed, A., Farooq, H., Rizwan, A., Imran, A. and Imran, M.A., “Leveraging Intelligence from Network CDR Data for Interference aware Energy Consumption Minimization” in IEEE Transactions on Mobile Computing, 17(7), pp.1569-1582. 2018

2. Rizwan, A., Arshad, K., Fioranelli, F., Imran, A. and Imran, M.A., “Mobile Internet Activity Estimation and Analysis at High Granularity: SVR Model Approach” in proceedings of IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2018

3. Rizwan, A., Nadas, J.P.B., Imran, M.A. and Jaber, M. “Performance Based Cells Classification in Cellular Network using CDR Data”. In proceeding of IEEE International Conference on Communications (ICC), 2019

4. Taufique, A., Rizwan, A., Imran, A., Arshad, K., Zoha, A., Abbasi, Q.H. and Imran, M.A., 2019. Big Data Analytics for 5G Networks: Utilities, Frameworks, Challenges, and Opportunities. Wiley 5G Ref: The Essential 5G Reference Online, pp.1-38.

5. Rizwan, A., Jaber, M., Filali, F., Imran, A. and Abu-Dayya, A., 2021. A zero-touch network service management approach using AI-enabled CDR analysis. IEEE Access, 9, pp.157699-157714.

6. Rizwan, A., Abubakar, A.I., Ozturk, M., Zoha, A., Hussain, S., Imran, M.A. and Abbasi, Q.H., “Machine Learning Based Predictive Mobility Management for Mobile Communication Networks” (in progress)

Contents

Abstract **i**

Statement of Originality **iii**

List of Publications **v**

List of Acronyms **xv**

Acknowledgements **xix**

Declaration **xxi**

1 Introduction **1**

1.1 Motivation 2

1.1.1 Progress in Enabling Technologies for 5G and Beyond Networks 4

1.1.2 Opportunities and Challenges in SON 4

1.1.3 Abundance of Real Network Data 4

1.1.4 Machine Learning Prediction Models for Enabling Proactivity in SON 5

1.2 Problem Statement 6

1.3 Research Contributions 8

1.4 Thesis Outline 10

2 Literature Review **13**

2.1 Self Organising Networks (SON) 15

2.1.1 Self-configuration 15

2.1.2 Self-optimisation 16

2.1.3 Self-healing 17

2.2 Wireless Cellular Network Data 17

2.2.1 WCN Data and its Classification 17

2.3 Machine Learning New Utilities 20

2.3.1 End-to-End Network Visibility 21
2.3.2 Self-Coordination 21
2.3.3 Long Term Dynamics of the Network 22
2.3.4 Faster and Proactive Network 22
2.3.5 Smart Caching 22
2.3.6 Energy Efficiency 23
2.4 Existing Applications of Data Driven Machine Learning 23
2.5 Further Unleashing the Potential of Data Driven Machine Leaning in WCN 28
2.5.1 End User KPIs 28
2.5.2 Measurement Composition of KPIs 29
2.5.3 Conventional Approach of KPI Estimation 33
2.5.4 Machine Learning Approach for KPI Estimation (User Centred Approach) 34
2.6 Opportunities and Challenges 36
2.6.1 Opportunities 37
2.6.2 Challenges 38
2.7 Conclusion 40

3 Machine Learning for Traffic Load Estimation 43
3.1 Data Set Description 48
3.2 Methodology 49
3.3 Proposed Model and Performance Metrics 50
3.3.1 Support Vector Regression (SVR) 50
3.3.2 Performance Metrics 52
3.4 Analysis and Results 52
3.4.1 Spatiotemporal Analysis of Maximum Internet Activity 52
3.4.2 Models Performance Analysis 54
3.5 Conclusion 57

4 Mobility Pattern Prediction for Wireless Cellular Networks using a Global Indexing System 59
4.1 Relevant work 60
4.2 Enabling Network Optimization via Mobility Prediction 61
4.2.1 Proactive Caching and Pre-buffering 61
4.2.2 Energy-Efficient Cell Switching 62
4.2.3 Base Station Skipping 62
4.2.4 Pop-up Networking 62

4.2.5 Secondary User Mobility Predictions in Cognitive Radio Networks . . . 62
4.3 Machine Learning Driven Mobility Prediction . . . 63
4.3.1 Methods of Locating Users . . . 63
4.3.2 ML Approaches . . . 66
4.3.3 Predicates . . . 67
4.3.4 Feature Space . . . 67
4.3.5 Prediction Schemes . . . 67
4.3.6 Predicates . . . 68
4.4 Comparative Study of Machine Learning Approaches for the Prediction of next Location . . . 68
4.4.1 Data Set Description . . . 68
4.4.2 Methodology . . . 69
4.4.3 Data Processing . . . 70
4.4.4 Results and Discussion . . . 71
4.5 Conclusion . . . 75

5 Performance Based Cells Classification in Cellular Network Using CDR Data 77
5.1 Machine Learning for Mobile Networks . . . 79
5.2 Methodology . . . 82
5.2.1 Clustering Model . . . 82
5.2.2 Clustering Validation Metrics . . . 83
5.2.3 Clustering Implementation Scheme . . . 84
5.2.4 Classification . . . 87
5.3 Data-set Description . . . 88
5.4 Results . . . 89
5.4.1 Clustering . . . 89
5.4.2 Classification . . . 92
5.5 Conclusion . . . 94

6 CDR Analysis for Energy Consumption Minimization 97
6.1 Leveraging CDR Data for Proactive ICI–EE Solutions . . . 99
6.1.1 CDR Analysis and Traffic Insights . . . 99
6.1.2 System Model and Problem Formulation . . . 101
6.1.3 Proposed Energy Consumption Aware Resource Allocation (ECA) Scheme . . . 105
6.2 Results and Analysis . . . 108
6.2.1 Traffic Predictability and Analysis . . . 108

6.2.2 Simulation-Based Energy Efficiency Gains 108
6.2.3 Service Quality and Scalability 110
6.3 Conclusion 110

7 Conclusion and Future Directions 113
7.1 Conclusion 113
7.2 Future Direction 114
7.2.1 Prognostics in SON 115
7.2.2 Reinforcement Learning in SON 115
7.2.3 Machine Learning on Big Data 116
7.2.4 Traffic Load Estimation and Applications 118
7.2.5 Mobility Pattern Predictions and Applications 119
7.2.6 Fault Diagnosis 119
7.2.7 Energy Efficiency by Proactive Instead of Cell Switching . . . 120
7.2.8 Integrated Model 121
7.2.9 Light Efficient Models 121
7.2.10 Control and Data Plane Separated Architecture 122
7.2.11 Bridging Gap between QoS and QoE 122

List of Tables

2.1 Various types of data generated from different layers of the wireless cellular networks 19
2.2 List of studies with key SON issues they addressed using data-driven machine learning 24

3.1 Results for the machine learning models against evaluation metrics . . 55

4.1 Different global geo-coding systems 66
4.2 Geohash grid specifications 66
4.3 Set of features exploited in enriched data with the help of KNN 69
4.4 Results of classification algorithms at different resolutions 72

5.1 Performance of K-means against clsuter validation metrics for different values of K on both tiers of clustering 85
5.2 Classification cross validation results 92

List of Figures

2.1 Main functions and sub-functions of Self-organising Networks (SON) [1,2] 16
2.2 Basic architecture of LTE [3] 17
2.3 Different approaches to categorise the network data. 18
2.4 Standard KPIs and their immediate dependencies. 29
2.5 Accessibility KPI related PIs and measurements 30
2.6 Utilisation KPI related PIs and measurements 30
2.7 Mobility KPI related PIs and measurements 31
2.8 Integrity KPI related PIs and measurements 31
2.9 Retainability KPI related PIs and measurements 32
2.10 Availability KPI related PIs and measurements 32
2.11 Machine learning for user-centric KPI hierarchy and dependencies . . 34
2.12 User-centric Accessibility KPI computation analysis using data driven machine learning 35
2.13 User-centric Integrity KPI computation analysis using data driven machine learning 36

3.1 Cells distribution with respect to maximum traffic generated on hourly basis in a week 45
3.2 Heat map of maximum hourly internet activity in 225 cells over a week 46
3.3 Grid over Milan and area under observation 49
3.4 Layout of SVR model implementation 50
3.5 CDF of maximum hourly Internet activity: Actual VS Predicted . . . 56
3.6 CDF of mean hourly Internet activity: Actual VS Predicted 56
3.7 CDF of minimum hourly Internet activity: Actual VS Predicted 56

4.1 Machine learning driven mobility prediction components and cellular network optimization use cases. 63
4.2 Geohash cells at resolution levels V to VIII. 65
4.3 Number of samples for each cell at different resolution levels of Geohash 70

4.4 Performance comparison of regression and classification approaches . 71
4.5 Performance of KNN for next location predilection 74
4.6 Correlation of features and predicate 75

5.1 Classification process flow chart 81
5.2 Selection of optimal K, at Tier I using elbow methods. 84
5.3 Four clusters generated at Tier I 86
5.4 Average normalized traffic 88
5.5 Sample segments in their respective clusters. 90
5.6 Confusion matrix 93

6.1 Call activity level for POI and non-POI 99
6.2 Internet activity level for POI and non-POI 99
6.3 Milan activity level at 6 am, 6 pm, 12 pm 100
6.4 POI on Milan map 101
6.5 Activity level for non-POI and POI macro cells 101
6.6 Weekly call activity 101
6.7 Weekly internet activity 101
6.8 System Model with Heterogeneous Cells 102
6.9 Flow diagram for ECA scheme 105
6.10 EE (bits/Joule) performance for non-POI cells(left), and POI cells (right) 109
6.11 Energy reduction gain (ERG) performance for non-POI (left) and POI case (right) 109

List of Acronyms

1G	First Generation of Wireless Cellular Systems
2G	Second Generation of Wireless Cellular Systems
3D	Three-Dimensional
3G	Third Generation of Wireless Cellular Systems
3GPP	3rd Generation Partnership Project
4G	Fourth Generation of Wireless Cellular Systems
5G	Fifth Generation of Wireless Cellular Systems
ANN	Artificial Neural Networks
AP	Access Points
ARIMA	Auto-regressive Integrated Moving Average
ARMA	Auto-regressive moving Average
ATSSS	Access Traffic Steering, Switch And Splitting
BS	Base Station
BSC	Base Station Controller
BSSAP	Base Station Subsystem Application Part
CAC	Call Admission Control
CAPEX	Capital Expenditure
CBS	Control Base Station
CCO	Coverage And Capacity Optimization
CDF	Cumulative Distribution Function
CDR	Call Detail Record
CDSA	Control And Data Plane Separated Architecture
CH	Calinski-Harabasz Index
CN	Core Network
CNN	Convolutional Neural Network
CQI	Channel Quality Indicator
CR	Cognitive Radio
C-RAN	Centralized Radio Access Networks
CRM	Customer Relationship Management
D2D	Device To Device

D2I	Drone To Infrastructure
D2X	Drone To User
DBS	Data Base Station
DES	Double Exponential Smoothing
DT	Decision Trees
DTR	Decision Tree Regression
ECA	Energy Consumption Aware
ECG	Energy Consumption Gain
ECR	Energy Consumption Ratio
EE	Energy Efficiency
eMBB	Enhanced Mobile Broadband Communications
E-RAB	E-Radio Access Bearer
ERG	Energy Reduction Gain
FARIMA	Fractionally Differenced Auto Regressive Integrated Moving Average
GPS	Global Positioning System
GSM	Global System For Mobile Communication
HMM	Hidden Markov Model
HO	Handover
HSPA	High Speed Packet Access
ICI	Inter-Cell-Interference
Inter-RAT	Inter-Radio Access Technology
IoT	Internet Of Things
KNN	K Nearest Neighbours
KPIs	Key Performance Indicators
LASSO	Least Absolute Shrinkage And Selection Operator
LM	Levenberg-Marquardt
LOF	Local Outlier Factor
LR	Linear Regression
LTE	Long Term Evolution
LTE-A	LTE-Advanced
M2M	Machine To Machine
MA	Mean Accuracy
MAE	Mean Absolute Error
MBMS	Multimedia Broadcast Multicast Service
MCS	Modulation And Coding Scheme
MDT	Minimization Data Tests
MIMO	Multiple Input Multiple Output

ML	Machine Learning
MME	Mobility Management Entity
mMTC	Massive Machine Type Communications
mm-Waves	Millimeter-Waves
MNO	Mobile Network Operator
MSC	Mobile Switching Center
MUE	Macro-Users
NB	Naive Bayse
NCL	Neighbor Cell List
NFV	Network Function Virtualization
NN	Neural Networks
NNR	Nearest Neighbour Based Regression
OBOs	Operational And Business Objectives
OCSVMD	One Class Support Vector Machine Based Detector
OPEX	Operational Expenditure
PGW	Packet Data Network Gateway
PI	Performance Indicators
PM	Performance Management
POI	Point Of Interest
PUs	Primary Users
QoE	Quality Of Experience
QoS	Quality Of Service
RACH	Random Access Channel
RAN	Radio Access Network
RAT	Radio Access Technology
RBF	Radial Basis Function
RBs	Resource Blocks
RLF	Radio Link Failure
RMSE	Root Mean Square Error
RMSSTD	Root-Mean-Square Standard Deviation
RNN	Recurrent Neural Network
RRC	Radio Resource Control
RRM	Radio Resource Management
RS	R-Squared
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
SCN	Satellite Communication Network
SCs	Small Cells

SDN	Software Defined Networking
SGW	Serving Gateway
SINR	Signal-To-Interference-Noise Ratio
SMS	Short Message Service
SON	Self-Organizing Network
SRM	Structural Risk Minimization
SUE	Small Cell-Users
SUs	Secondary Users
SVC	Support Vector Machine Based Classifier
SVM	Support Vector Machine
SVR	Support Vector Regression
TAU	Tracking Area Update
TCO	Total Cost Of Operation
TCP	Transmission Control Protocol
TTI	Transmission Time Interval
UE	User Equipment
URLLC	Ultra-Reliable Low Latency Communications
V2V	Vehicle To Vehicle
V2X	Vehicle To User
VoIP	Voice Over Internet Protocol
VoLTE	Voice Over LTE
WCN	Wireless Cellular Networks

Acknowledgements

First and foremost, I am grateful to my main supervisor, Prof Muhammad Ali Imran for his commendable mentorship, throughout this journey of research training and personal development in my PhD. I would like to express my sincere gratitude to him for always being there with a patiently listening ear, kind advice, and willingness to facilitate whenever some academic or personal challenges were hindering my progress and growth. I would also like to thank him for giving me the opportunity to work under his laissez-faire supervision and providing funding for my PhD. Further to that, I want to thank him for being kind to involve me in multiple research projects which not only gave me opportunities to have first-hand experience of working on market-driven research under his inspirational leadership but also helped me to pay off the balance of my tuition fees. I must commend, he knows the art of adding value to others.

Secondly, I would like to thank my second supervisor Dr Qammer Abbasi who always inspired me to go out of my comfort zone and tap my true potential. I want to thank him for his support, valuable inputs and feedback on a regular basis that helped me to convert my research work into tangible results. I want to thank him especially for making himself available as the first point of contact for any help or guidance I needed.

I would also like to thank Dr Ahmed Zoha for sharing his knowledge and providing guidance in technical matters particularly in machine learning. Though he was not my official supervisor, he always spared time out of his busy schedule to provide detailed feedback and inputs on my research work throughout my PhD.

I would also like to thank my third supervisor Dr Francesco Fioranelli, who ensured his support and availability for help, particularly at the start of my PhD when I needed it the most. I am also grateful to Dr Sajjad Hussain and Dr Wasim Ahmad for all the discussions and feedback on my research in general. I would also like to thank my friends and colleagues who helped me keep going during my PhD with their moral support and valuable company.

I must acknowledge and appreciate the support of my wife Attia Batool who also took care of many of my responsibilities to help me focus on my PhD, without her ubiquitous support this journey could not be possible. I would also like to thank her for building our family with a beautiful son Muhammad Hasan Ali, whose presence eased out this challenging last year of my PhD, particularly the pandemic phase. I would also like to thank my family, parents, sisters, brothers for their unconditional support, particularly my elder brother Dr Ali Imran for having confidence in me and always being there to help, support and uplift whenever I went down. Without him, I would have not trod this path.

Aligned with academic integrity and as part of the post-defense refinement process, I utilized AI tools to proofread and in some cases rephrase selected content, focusing solely on enhancing the presentation of this thesis. In the end, I would like to dedicate my work presented in this thesis to my beloved sister late Erum Shafqat who gave me the courage to take this path. Even though she was battling with cancer with a known end but she always showed me positivity and hope to go this long path. She was the happiest person when I got this opportunity and for reasons, I believe, she would have been the happiest person for me completing it.

Declaration

With the exception of chapters 1 and 2 , which contain introductory material and literature review and partially chapter 7, where the empirical model of implementation is a work of my colleague Dr Ahmed Zoha and it is included to maintain the flow of the thesis, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Introduction

5G has transformed telecommunications, opening new opportunities across diverse domains. For end-users, it offers seamless connectivity, unmatched mobile speeds, and reliable service even in congested environments, significantly enhancing the quality of experience (QoE). For network providers, 5G accommodates a vast array of devices and services while improving quality of service (QoS) and reducing operational costs.

According to Ericson's report of 2024 [4], till 2024, over 300 operators globally have launched commercial 5G services, with approximately 2.3 billion subscriptions accounting for 25% of all mobile users. Projections suggest this will grow to 6.3 billion subscriptions, or 67% of all mobile users, by 2030 Beyond these figures, 5G's true potential lies in its transformative use cases, spanning healthcare, autonomous vehicles, industrial automation, and augmented reality. These applications are fostering a smart, interconnected digital ecosystem that reshapes communication, commerce, and daily life .

5G networks are meeting and exceeding expectations, laying a solid foundation for future cellular advancements. These networks deliver improvements in data rates, reliability, latency, connectivity, and coverage, while also supporting diverse applications, Internet of Things (IoT) integration, and enhanced data security and privacy. By 2024, cellular IoT connections have reached 4.2 billion, with forecasts exceeding 6 billion by 2030 [4].

5G's success is built on enabling technologies such as ultra-densification, massive multiple-input multiple-output (MIMO), millimeter-wave (mmWave) spectrum, software-defined networking (SDN), network function virtualization (NFV), big data analytics, and cloud computing etc. [5–8], These technologies collectively address growing demands for higher data rates, capacity, and coverage, enabling data-intensive applications like autonomous vehicles and smart cities.

Ultra-densification involves deploying numerous heterogeneous cells (macro, micro, and small cells) in an area to enhance coverage. High-frequency mmWave spectrum (24 GHz to 100 GHz) provides greater bandwidth and higher data rates but has limited range. This limitation is mitigated by small cells and massive MIMO. Massive MIMO, with its large antenna arrays, increases throughput and spectral efficiency. Together, these technologies enable high performance while improving energy efficiency. While these advancements enable 5G's capabilities, they also introduce complexities. Modern base stations (BSs) are intricate systems with thousands of parameters. Managing nationwide networks with hundreds of thousands of BSs presents challenges in configuration, optimization, and maintenance. Currently, these tasks rely heavily on human expertise, resulting in high operational expenditures (OPEX) and delays in issue resolution.

In 5G and beyond, manual management is neither practical nor sustainable. Autonomous solutions capable of handling network operations efficiently are critical. To address this, the concept of self-organizing networks (SONs) was introduced. SONs reduce human dependency by automating tasks such as resource optimization and fault detection.

Self-organizing networks use machine learning (ML) to perform network operations intelligently. ML enables SONs to analyze real network data, predict traffic loads, and forecast user mobility patterns. These predictions allow SONs to optimize network resources proactively, ensuring seamless service and enhanced QoS.

Additionally, ML identifies abnormalities in network performance, aiding in timely maintenance. This automated fault detection is essential for managing complex 5G networks. By leveraging ML, SONs can transition from reactive to proactive management, addressing challenges and improving efficiency.

Detailed discussions on SON functions, machine learning applications, and network data are presented in Chapter 2.

1.1 Motivation

With its journey from the First Generation (1G) of connecting people to people, to the Fifth-Generation (5G) connecting everything, wireless telephony has already become part and parcel of our socioeconomic fabric. Second-Generation (2G) brought a cultural revolution by introducing sms and mms services, and Third-Generation (3G) with voice over IP (VoIP) paved path for internet-based services that lead to Fourth-Generation (4G) of live streaming. 4G brought a tremendous increase in data rate and mobile usage for the users and profits for the mobile network operators. Building on this momentum, 5G offers unprecedented improvements in data rates, latency, connectivity, and capacity. Its transformative potential lies in diverse use cases across industries, enabling advancements in healthcare, autonomous systems, industrial automation, and smart cities. Beyond facilitation, 5G and beyond networks aim to revolutionize production, services, and lifestyles.

Nevertheless, characteristics of 5G and beyond networks are not just limited to ultra-reliability, low latency, faster data transportation and instant access enabled by greater bandwidth, dense networks, increased number of channels, and massive MIMO. A key desired attribute of future mobile networks is being cognisant. It means it is capable of context-aware autonomous decision making, enabled by artificial intelligence. On the network end, this intelligence is supposed to help network adapt to dynamic ecosystem of 5G and beyond networks. Role of artificial intelligence in 5G and beyond networks is expected to exponentially increase the slowing down linear growth of financial worth of the mobile industry. Besides financial growth, artificial intelligence can increase the impact of future mobile technologies (5G and beyond) manifolds in all walks of life.

The defining characteristics of 5G and beyond extend beyond ultra-reliability, low latency, and high-speed data transmission. A key aspiration is to make networks intelligent and context-aware, capable of autonomous decision-making through artificial intelligence (AI). This intelligence enables networks to adapt to the dynamic and complex ecosystems of 5G and beyond. By integrating AI, these networks can unlock exponential growth in both financial value and societal impact.

Despite substantial progress in enabling technologies such as ultra-dense small cell deployments, massive MIMO, and millimeter-wave (mmWave) spectrum,network complexity has grown significantly. This complexity makes manual management and optimization increasingly infeasible. Network operators, burdened by high operational expenditure (OPEX) and reliance on domain expertise, face challenges in maintaining quality of service and promptly addressing network issues. As 5G networks expand, manual operation becomes neither practical nor economically viable, emphasizing the need for autonomous solutions. This realization has driven research into self-organizing networks (SONs), which aim to reduce human dependency while enhancing efficiency.

This research draws inspiration from the pivotal role of 5G and beyond networks in driving socioeconomic progress. It addresses the challenges and leverages opportunities presented by the enabling technologies critical to realizing the vision of 5G and beyond.

1.1.1 Progress in Enabling Technologies for 5G and Beyond Networks

As mentioned earlier that 5G and beyond is attributed with some key enabling technologies like network ultra-densification with heterogeneous cells, extended bandwidth of spectrum, SDN, massive MIMO, NFV, big data, cloud computing, etc. A remarkable progress is already made in these disciplines independently and under the umbrella of wireless cellular technologies. Maverick technologies like cloud computing and big data are revolutionising every field of life, SDN and NFV have emerged as the new norm in networking, whereas high-frequency radio waves like mm-Wave are the new carriers for the fast data transportation. Modern mobile network infrastructures, already in place, are taking advantages from these technologies along with the mobile network specialised technologies like deployment of heterogeneous cells, massive MIMO, and splitting and reusablity of extended bandwidth. Progress in these fields and their pragmatic role in the realisation of 5G and beyond networks are very promising.

1.1.2 Opportunities and Challenges in SON

The plethora of enabling technologies mentioned above increase the complexity of the network manifolds making the manual resource management and network maintain impossible. Though, SON has been applied as an alternative for the auto-management of resources. But still, to date, legacy SON functions commonly rely on the predefined instruction set developed with the help of domain knowledge and results from the field tests. It leads to too much operation cost and unnecessary delays neither affordable operationally nor viable commercially. Therefore, there is a dire need for making SON proactive rather than reactive. To achieve this goal developments on two ends can really help. First, the use of alternative data to understand the network, for this purpose 3GPP 10 has proposed minimisation data tests (MDT). In this approach instead of conducting field test, network operators can exploit user equipment (UE) data and network data containing very useful information. Second, the use of machine learning for the prediction of possible future scenario based on the historical data collected from network and UE.

1.1.3 Abundance of Real Network Data

The network behaviour related intelligence can be gathered by applying machine learning on the data generated from the different ends of the network. Network traffic behaviour is the existence of certain patterns in the traffic over the network. This traffic can be generated by human-to-human, human to machines and machines to machines interaction in the context of network facilities. The continuous interaction leads to certain patterns of services consumption, users' mobility, network operations, resources requirement etc. that combined can be labelled as overall network behaviour. Some data generated from different ends of the network are discussed in chapter 2.

One such very important data set from wireless cellular networks is called detailed record (CDR). It is very comprehensive data which contains footprints of SMS, Calls and Internet usage activity over the network. It comprises timestamp for every-time a service is accessed and also the location of users, which can help to extract information about the network traffic behaviour in spatiotemporal dimension. Machine learning on the real network data like CDRs is getting popular in research and commercial models for gathering intelligence about different aspects of network behaviour like traffic congestion, handovers, network failure etc. One main reason for using real data instead of synthetic data is that the synthetic data is generated based on multiple assumptions for a predefined scenario in a simulation. But real scenarios may differ and model trained on synthetic data will not stand valid. This becomes even more crucial in the 5G and beyond networks which are so complex, involving so many parameters that they all can not be considered in a simulation. Therefore, the use of real data can help there as it reflects the actual status of the network influenced or tuned by all the parameters involved.

1.1.4 Machine Learning Prediction Models for Enabling Proactivity in SON

The discussion above clearly reflects machine learning has a crucial role in making SON proactive. Machine learning is a well-known application domain of artificial intelligence. It is a data-driven approach in which data and possible outcomes are fed to machine learning algorithm which as a result builds a model that can make choices and decision in an automated manner based on the patterns in the data. But the real beauty of machine learning algorithms is there capacity to predict future possibilities or outcomes for certain scenarios based on the historical data. This capacity of machine learning to predict can play a crucial role in making SON proactive rather than reactive. Because of the availability of real-time real network data like CDR, SON does not need to depend on the results of field tests. Machine learning can leverage from this data to predict future possible outcomes and provide it to SON functions so it can manage and optimise it's resources in advance to ensure seamless services.

Making SON proactive is not the only requirement, but developing simple efficient models is also very important. Deep learning applied on data from complex data sets, e.g., raw network data, has been trending recently to make SON proactive. Machine learning is meant to address challenges created due to increased complexities in 5G and beyond network, and not meant to add to network complexities and OPEX. But deep learning models itself or complex and data hungry. They require loads of data and more computational resources as compared to classical machine learning algorithms. It makes deep learning not the optimal choice in many prediction scenarios for the SONs, particularly where the computational resources are limited like in edge or fog computing. Secondly, where the classical machine learning models can easily exploit the hidden patterns in a well known easily available data set to outperform deep learning, there is no point of using deep learning in such scenarios. For example, CDR data can be exploited with the classical machine learning algorithms for the development of the models for the prediction of traffic load, mobility pattern and fault detection.

1.2 Problem Statement

Where network ultra-densification with a plethora of technologies is making 5G and beyond networks a reality, there it brings additional challenges mainly because of the complexity increased manifolds. It is no longer operationally possible or commercially viable to manually optimise numerous network resources, thousands of network parameters and maintain hundreds of thousands of network nodes. In a setup of deployment of small cells in network ultra-densification, there are two key challenging areas. First is the proactive optimisation of network resources like the optimisation of the radio resources, energy consumption, and network parameters for example antenna tilts. Second is the management and maintenance of the numerous cells in the network.

Though SON is proposed to address these challenges to avoid dependency on human involvement. But legacy SON still depend mostly on predefined instruction sets which can not adapt to continuously changing environment of the network in terms of traffic load, users location and services consumed. Besides that, SON commonly depend on data collected from field tests which cause unnecessary delays and make SON reactive. Such approach is not viable for 5G and future networks which are expected to adapt to continuous network changes in a proactive manner. Motivated by the opportunities mentioned in the section above, to address this main challenge of making SON proactive rather than reactive we try to find answers for the following questions in this research:

1. Does there exist some patterns in real network data that can be exploited to predict future scenarios helpful for SON to make decisions in a proactive

manner? Finding patterns in real network data can be very helpful as it provides a realistic picture of the network or user behaviour.

2. Is it possible to identify a single comprehensive data set that can be used to exploit the patterns in the data and predict future scenarios? Having a single comprehensive data can help to reduce the overhead cost of data processing

3. What are the potential machine learning algorithms that can produce accurate traffic prediction models and out perform other state of the art prediction models for the same scenarios? Traffic load estimation is very important for making SON proactive and it is a well researched topic. Though the use of deep learning and data from multiple sources is trending, but develop simple prediction models with simple algorithms, data and features is very important. It is so because deep learning itself is very complex and data hungry which makes it not feasible for simple tasks or tasks performed on the edge of the network. Nevertheless, if simple models can do the same job there is no point of using complex models to overburden an already complex network.

4. Is it possible to develop user location prediction model using a global hierarchical grid indexing system? Network comprises heterogeneous cells overlaying each other. Understanding of users mobility pattern across cells and capacity to predict users next cell can play crucial role in the provision of services in seamless manner. Contemporary models rely on synthetic data and grid system defined by the researchers for specific scenario. They neither reflect the real mobility pattern of users nor the study perfomed can be transferred or scaled. It is therefore important to study and develop prediction model based on real mobility data of human and a global grid system that can be used by any other entity.

5. Can real network data like CDR help to develop a solution for the the auto diagnosis of the sub-optimally performing nodes in the network? Faulty nodes are commonly identified by the field tests and heavily rely on human involvement and domain knowledge causing delays in the identification and diagnosis of the fault. For 5G and beyond networks it is inevitable to have an auto-diagnosis solution which not only detects a fault on real time but also identify the root cause of the problem.

6. How can real network traffic load prediction model help in Self-optimisation function of SON for energy consumption minimisation? Can one such prediction model be used to address multiple SON challenges like inter-cell-interference (ICI) and energy consumption? ICI and energy consumption are twin key

challenges faced by 5G and beyond networks. Traffic load prediction model have been researched and used to address these challenges separately using synthetic data. Addressing the simultaneously exploiting real network data is an ideal scenario.

1.3 Research Contributions

The main goal of this research has been to address the key challenge of making SON proactive rather than reactive. It is further narrowed down by focusing on adding machine learning based predictability in two key SON functions, Self-optimisation and Self-healing. For Self-optimisation, underlying key enabling mannerisms e.g., traffic density patterns and mobility patterns are identified along with a relevant data set CDR for the study of these patterns. The CDR data is exploited with the help of classical machine learning algorithms to develop models for the prediction of traffic load and mobility patterns. The models developed can be used for making SON functions like Self-optimisation proactive rather than reactive for multiple use cases, e.g, radio resource management, energy efficiency, data offloading etc. Besides that, CDR data is also exploited for developing machine learning based scheme for the auto diagnosis of network issues. A machine learning based solution is developed that categorise network cells based on their performance. Further to that, sub-performing cells are also identified and classified into different categories. Study on network traffic prediction is extended and model built on CDR data is used in a use case of the minimising inter-cell-interference (ICI) and energy consumption.

In this research, classical machine learning is used to develop generic high-resolution network traffic load prediction models, global mobility pattern prediction model helpful towards SON use cases of Self-optimisation. But the analysis and models developed in this research can also be used by other public & commercial entities like advertising agencies, road infrastructure and urban planning organisation etc. to optimise their resources and improve their services. For example, traffic density prediction models can provide insight about population density in an area at a certain time and potential users of a service if users profiling is also performed. Similarly, mobility pattern prediction models can help to identify influx and outflux of people in a certain area at a particular time of the day. It can also help to identify popular routes adopted by mobile users at a certain time of the day. Different service providers can optimise their services proactively based on this information.

The major contributions of this research include:

1. **Identification of Fundamental Patterns for Generic Predictive Models**

   Towards the goal of enabling proactive SON functions for diverse use cases, two key patterns—traffic load and user mobility—are identified. Prediction models for these patterns, presented in Chapters 3 and 4, are designed for applications across multiple SON use cases, such as proactive radio resource management, energy efficiency, data offloading, handover management, and content caching. A review of existing and potential use cases leveraging such models is provided in Chapter 2.

2. **Identification of a Single Comprehensive Dataset**

   Following MDT guidelines, this research identifies and classifies network-generated data from various sources, highlighting their roles and importance in Chapter 2. End-user KPIs are linked with contributing PIs and network measurements. Call Detail Record (CDR) data is identified as a comprehensive dataset capable of supporting diverse predictive models, including traffic load and user mobility prediction. Furthermore, the same CDR data is used to classify network cells/nodes based on performance.

3. **Development of Efficient and Accurate Predictive Models**

   To achieve efficient and accurate predictive models that enable proactive SON functions, CDR data is utilized with classical machine learning algorithms. Traffic prediction models developed in Chapter 3 demonstrate high-resolution internet traffic prediction, outperforming state-of-the-art statistical and deep learning models. In Chapter 4, user location prediction models are developed using a global indexing system, ensuring transferability across different networks. These models use classical machine learning techniques to strike a balance between simplicity and accuracy.

4. **Proactive Self-Optimization via Traffic and User Location Prediction Models**

   Two types of predictive models are developed to enhance the SON self-optimization function proactively.

   Traffic Prediction: A high-resolution traffic prediction model is developed using Support Vector Machines (SVM) to predict minimum, mean, and maximum internet traffic hourly. This efficient and simple model outperforms state-of-the-art algorithms and is detailed in Chapter 3.

User Location Prediction: A series of regression and classification models are developed to predict users' next multiple locations based on a global indexing system (e.g., latitude and longitude). These models, presented in Chapter 4, use basic features to maintain simplicity while providing accurate predictions. These insights allow SON to optimize resources proactively by forecasting traffic loads and user mobility patterns.

5. **Robust Self-Healing via a Hybrid Scheme** A hybrid scheme comprising clustering and classification techniques is proposed to enhance SON's self-healing function. CDR data is utilized to identify patterns in cell performance in real wireless networks at an hourly resolution.

   Clustering: Patterns are identified using the K-means clustering algorithm, with cells labeled based on performance using domain knowledge.

   Classification: A classification model based on the SVM algorithm is developed, achieving an accuracy of approximately 98% in classifying cell performance. This method ensures a robust and efficient auto-diagnosis system for sub-optimal network nodes, as detailed in Chapter 4.

6. **Deploying Traffic Prediction Models for Multiple SON Challenges**

   A traffic prediction model based on SVM is developed using real CDR data and deployed to address multiple SON challenges simultaneously. In Chapter 4, this model is implemented in a scheme to minimize inter-cell interference (ICI) and energy consumption concurrently demonstrated in Chapter 6: .

   A mathematical model is formulated for the joint optimization of ICI and energy consumption. The model is evaluated through system-level simulations using NS-3, demonstrating its effectiveness in addressing these critical challenges.

1.4 Thesis Outline

The rest of the thesis is organised as follows. Chapter 2 presents a comprehensive literature review, beginning with a brief introduction to Self-Organizing Networks (SON) and their use cases. It classifies data generated from various parts of an LTE network under different schemes. Emerging and existing applications of data-driven machine learning in cellular networks are also discussed. Additionally, this chapter highlights important Key Performance Indicators (KPIs) and their contributing Performance Indicators (PIs), explaining how machine learning applied to network data can aid in realizing these KPIs and PIs.

Chapter 3 focuses on an empirical study using the conventional machine learning algorithm, Support Vector Regression (SVR), to predict internet traffic at a high temporal resolution. The performance of SVR-based models is compared against state-of-the-art statistical and deep learning algorithms. Chapter 4 extends the research to mobility patterns, this chapter presents the machine learning prediction models developed for user mobility and their outcomes. It discusses emerging applications and popular schemes, detailing the critical steps involved in mobility pattern prediction for 5G and beyond networks.

Chapter 5 presents a study on another important task for the SON in wireless cellular networks, called fault detection. It presents a hybrid approach based on the classical machine learning clustering and classification algorithms.

Chapter 6 presents a solution for two key issues expected in 5G and beyond networks, namely inter-cell interference and energy consumption. It presents a mathematical model of joint optimisation for these two issues, then proposes an algorithm to address this issue and evaluate the proposed algorithm in a scenario developed with the help of machine learning-based traffic prediction model. The same solution is also evaluated with frequency reuse-1 scheme, through system level simulations in NS-3.

In the end, Chapter 7 concludes how the research presented in this thesis has achieved the main objective set for this research. Besides that, it also lists down future direction for the extension and application of the research and prediction models presented in this thesis.

Chapter 2

Literature Review

The evolution and ubiquitous Wireless Cellular Networks (WCN) have revolutionised the world by enabling human to human, human to machine and Machine to Machine (M2M) communication. These communications, a range of spatiotemporal data is generated. This data is related to signal quality, traffic flow, subscriber profile, Quality of Service (QoS), Quality of Experience (QoE), customer complaints and feedback, wireless environment, network configuration, network operations and management etc., amongst others. Additionally, with the network evolution towards 5G, a multitude of technologies such as network densification, miscellany of node types, split of control and data plane, network virtualisation, heavy and localised cache, infrastructure sharing, concurrent operation at multiple frequency bands, simultaneous use of different medium access control, physical layers and flexible spectrum allocations are expected to increase the size of data generated from the network and about the network, exponentially.

Data driven machine learning has transformed the way we think, deal, and benefit from the world of data and information. It is a significant leap forward from conventional sample-based, hypothesis-driven studies to robust analyses of entire datasets. This transformation has been possible by overcoming the limitations of data storage and processing regardless of the size and nature of data. Part of this revolution lends itself to Moore's-law-backed growth in hardware storage and processing power. Other part stems from advances in software including parallel processing tools such as MapReduce. Leveraging both, the hardware and software advances, data driven machine learning tools can help us picture such a past, present and future of facts, information and predictions with the help of actual data that might not be seen otherwise.

In legacy WCNs, only a fraction of the data generated about the network state and health is retained and analysed. Even this limited analysis is mostly carried for business intelligence and/or network performance evaluation purposes, while most of the data is discarded or remain unused, trapped in silos. Due to analysis of only partial data, the information extracted are mostly incomplete and inaccurate. Thus, the decisions made are often guesstimates leading to sub-optimal performance. Leveraging data driven machine learning can unleash the potential to combine and analyse the whole data from all segments, unveiling end-to-end picture of the network and customers and uncovering the hidden potentials and patterns within the data. WCN operators and vendors can leverage from data driven machine learning to extract key information that can be crucial for optimising and maintaining the network resources efficiently.

This chapter presents the literature review on the data generated from different ends of the network, applications and potential use cases of data driven machine learning, and important Key Performance Indicators (KPIs) with their subsequent Performance Indicators (PI) that can be fulfilled by machine learning applied on real network data. There has been existing surveys on wireless network data from application, network, transmission, data layers and mobile big data perspective discussed in [9] and [10] respectively. However, they do not cover industry professionals' and practitioners point of view specifically how wireless cellular network performance can be improved using real network data and what challenges and opportunities exist for the 5G and beyond networks. Therefore, this chapter discusses the opportunities and challenges that stem from the crossroads of data-driven machine learning and wireless cellular networks..

Salient Features of Literature Review

The salient characteristics of the literature review performed here are summarised as follows:

1. A wide range of WCN data types are identified that can be collected from User Equipment (UE) and from various sources within Radio Access Network (RAN) and Core Network (CN). Besides that taxonomies are listed to classify the WCN data into various categories (Section 2.2).

2. Several new utilities are identified and discussed for machine learning applied to WCN data to improve WCN performance in various aspects. It is explained how these data can be leveraged to overcome technical challenges that affect network performance. It also discusses how data driven machine learning can help achieve some of the envisioned 5G and beyond requirements, which otherwise are very challenging (Section 2.3).

3. A brief review of the literature on the data driven machine learning for WCN along with a comparative summary of various studies in tabular format are presented in (Section 2.4)

4. A new perspective is present on how the resolution of the data, particularly in case of measurements, can drastically change the insights that can be drawn. Lower level measurements are identified which usually remain hidden as dark data but has potential to be utilised by machine learning for enabling 5G and beyond wireless networks. This perspective is explained with the help of several examples of measurements and data streams from Long-Term-Evolution (LTE) networks (Section 2.5).

5. Some emerging opportunities and challenges in WCN data driven machine learning for the next generation of cellular networks aka 5G and beyond are presented in (Section 2.6).

2.1 Self Organising Networks (SON)

In contemporary networks human input and intervention is required to accomplish network operations from network deployment to network maintenance. 5G and Beyond networks are envisioned as cognisant networks capable to perform end-to-end network functions seamlessly in an autonomous manner without any external intervention. To meet this goal, researchers and standardisation bodies like Third Generation Partnership Project (3GPP) has proposed SON. SON are characterised with intelligence to make important decisions for the network management on its own with least possible human intervention. This intelligence can be gather by exploiting the network with the help of machine learning. The data generated from different ends of the network provides contextual information about the network and users' behaviour in different circumstances. Whereas machine learning exploit hidden patterns in the that data and based on those patterns can differentiate among different scenarios and make decisions for important network operations.

The main functions expected from the SON are commonly grouped into three broad categories: Self-configuration, Self-optimisation and Self-healing [1,2]. Key functions expected from SON under these three headings are listed in Fig 2.1. 5G and beyond networks characterised with SON features are expected to configure network nodes and parameters, optimise network resources and deal with network performance issues in an automated manner.

2.1.1 Self-configuration

Whenever a new network is deployed, or extended with the addition of new terminal or upgraded with new facilities, there are many network parameters that need to be configured. These parameters include, operational and radio access parameters

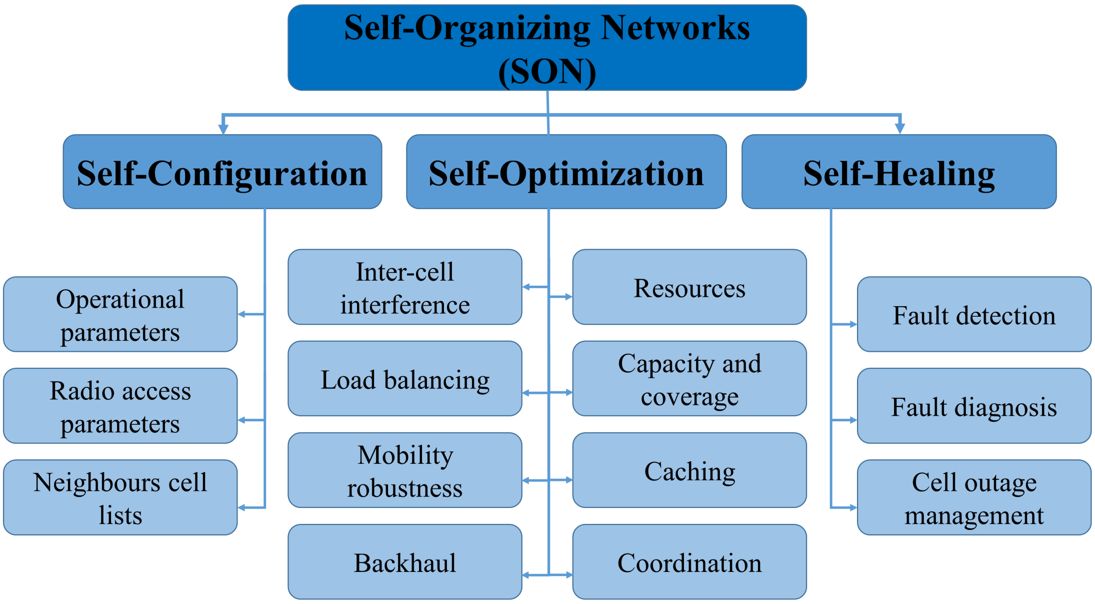

Figure 2.1: Main functions and sub-functions of Self-organising Networks (SON) [1,2]

such as IP address and backhaul, besides the neighbour cells list configuration needed for enabling communication between the newly installed cell and the existing neighbouring cells. In contemporary networks, these parameters are configured at base station with input from domain experts. It is aimed that Self-configuration feature in SON will configure these parameters autonomously such that the newly added nodes work as plug and play.

2.1.2 Self-optimisation

Another key role expected from SON is Self-optimisation. After network configuration, Self-optimisation aim to optimise the operational parameters and utility of network resources in an efficient and autonomous manner. Network with Self-optimisation function is expected to automatically adapt to changing circumstances and schedule the resources and operational parameters such that the best possible services are ensured with the minimal operational cost. 5G and beyond networks are expected to have thousands and thousands of base stations with numerous configuration parameters and huge requirement of overall resources, energy and bandwidth to name few. It is very important function to make future cellular networks viable rather than profitable. Main aspects of the network that SON need to optimise in automated manner are network traffic distribution aka load balancing, inter-cell interference, network robustness towards the mobility of connected devices, network resource, backhaul and caching.

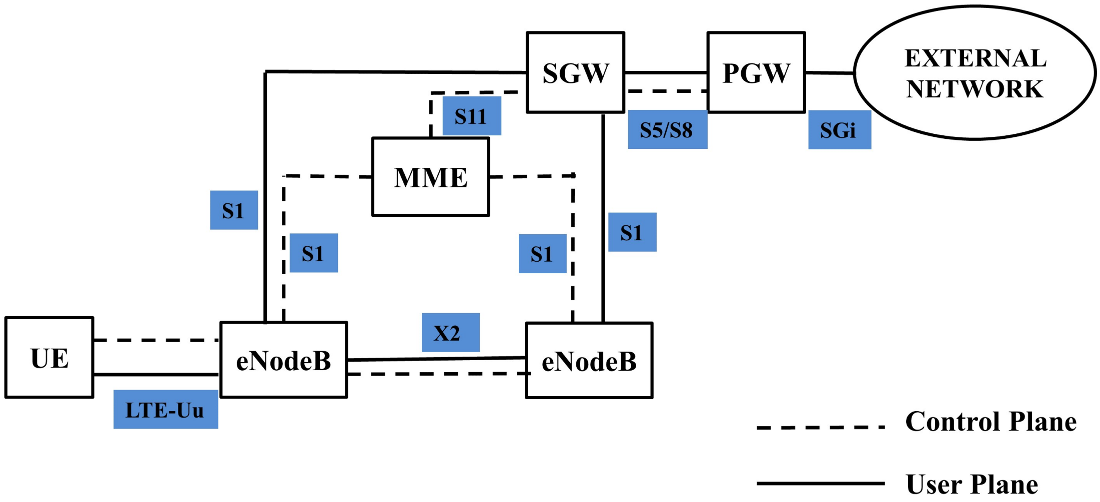

Figure 2.2: Basic architecture of LTE [3]

2.1.3 Self-healing

Contemporary cellular network comprise hundreds of thousands of base-stations and they need numerous skilled people for on site troubleshooting and maintenance which costs network operators a fortune. In 5G and beyond networks the number of these sites are expected to increase many-folds and it will be non-viable for the network operators to maintain them manually. Here self-haling function of SON can play it's key role. Self healing refers to the auto detection of a non functional or sub performing network facility due to some issue. It is not limited to that, it is further expected to identify the root cause of the issue and also try to resolve that issue autonomously. The main goal of Self-healing function is to eliminate or reduce the network performance inefficiencies autonomously.

2.2 Wireless Cellular Network Data

In this section, data collected from UE and from various nodes of RAN and CN of the LTE-Advanced (LTE-A) network are presented. Initially data is labelled corresponding to measurements, events and network operation then it is classified into different categories.

2.2.1 WCN Data and its Classification

A typical cellular network architecture with network elements and the standardised interfaces is shown in Fig. 2.2 (using LTE terminology). At a high level, the network has three layers; the user end, the access network and the core network. There are various sources of data in the network that can be classified based upon 1) their

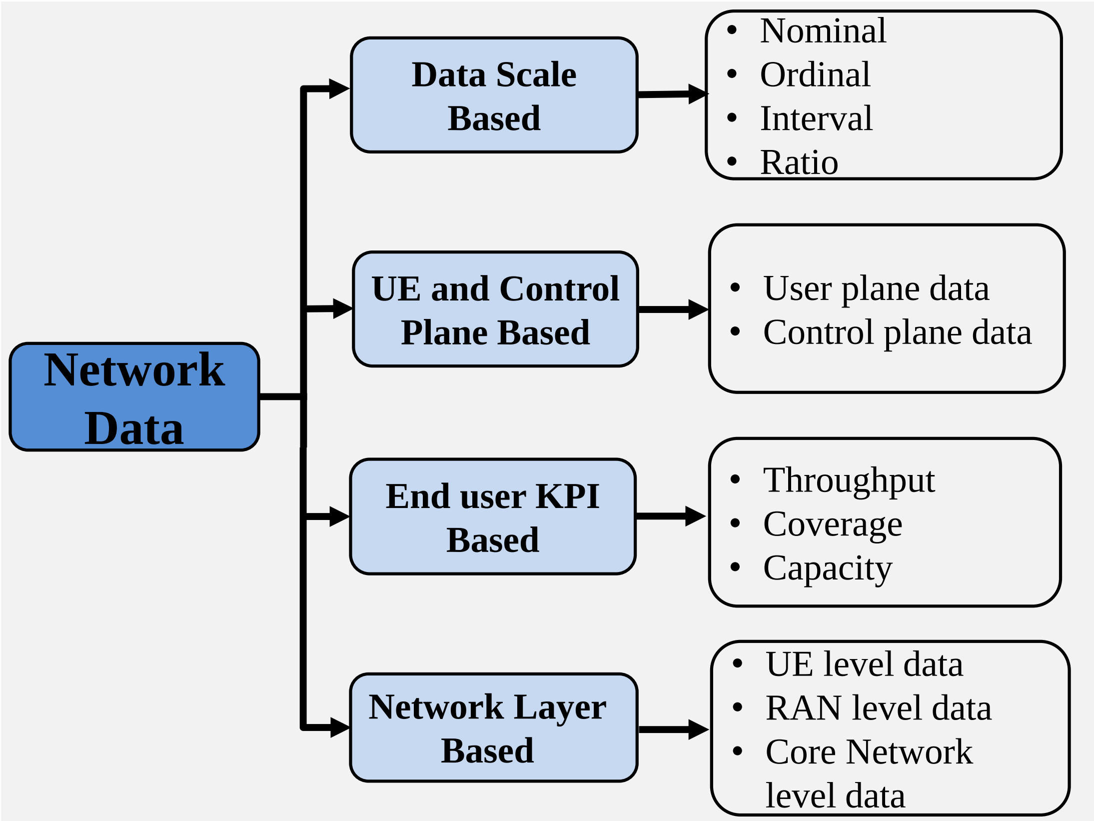

Figure 2.3: Different approaches to categorise the network data.

association to layers of the network (i.e. network layer based) 2) corresponding plane: user plane (UE data) or control plane (control data) 3) their impact on end user KPIs (relevant KPI based) and 4) their scale and measurement as shown in Fig. 2.3

1. *Data and Measurement Scale Based Classification:*

   Under this category, data is classified into four types Nominal, Ordinal, Interval and Ratio. Nominal scale are used to label data without any quantitative measure, e.g. Radio lLink Failure (RLF) and Reference Signal Received Power (RSRP). It is important to note that nominal data are mutually exclusive and they do not possess any numerical significance. Ordinal scale provides order to the data, however, the actual difference between them is not really known, e.g., A is better than B or C but it can not be quantified how much better it is. Thus, Ordinal scales are typically measures of non-numeric concepts like customer happiness, discomfort and satisfaction. Interval type of scaling in addition to knowledge of the order, allows to quantify the actual difference between the variables. Time is a good example to describe Interval type of data, in which the increments are known, consistent and measurable. Ratio scale tells about not only the orders of the variables, but also exact value between units (i.e. Intervals) and allows for a wide range of descriptive and inferential statistics to be applied.

Table 2.1: Various types of data generated from different layers of the wireless cellular networks

Network Layer Based Data						
RAN Data			Core Network Data			
Subscriber Data		eNodeB Data	Subscriber Data		MME Data	SGW & PGW Data
Voice Session	Data Session		App Data*	Business Data		
Call setup time	IP traffic flow	Received interference power	Indoor/ outdoor status	Subscription type	Service request failure ratio	Total no. of packets
Call success rate	Data session success rate	Thermal noise power	Mobility status	CRM data	Dedicated bearer activation failure ratio	Session failure ratio
Access failure	Service setup time	Channel base band power	Contextual information	Customer complaint	Attached users	Dedicated bearer creation failure ratio
HO failure rate	IP throughput	Channel quality indicator	Event Data/Detail record	Spectrum utility map	Supported subscribers	Modify bearer failure ratio
HO Success rate	Application throughput	Reference signal received power		Customer satisfaction (churn)	Attach failure ratio	Down link data drop ratio
Call drop ratio	Session drop rate	Minimisation of drive test reports			Bearers Established	
Speech quality	Data streaming quality	PRB usage per cell**			Paging failure ratio	
Packet jitter		No. of connected users**			X2 handover failure ratio	
Delay		Received random access**			S1 handover failure ratio	
		Preamble per cell**			Intra MME Tracking Area Update(TAU)	
					Intra MME TAU failure ratio	

2. *User and Control Plane Based Classification:* This data belong to either user plane or control plane. Data in the network is either utilised by the UE or used for providing signalling and control information to the UE. Recently there has been a new proposed Control and Data Plane Separated Architecture (CDSA) that separates the control plane from data plane to avoid coupling between data and control access points [11]. In addition, user and control plane data facilitates leveraging Software Defined Network (SDN) approaches that softwarise the control plane of the network and decouples it from data plane. Thus identification of user plane and control plane data becomes essential in realising futuristic network operations.

3. *End-user KPI Based Classification:* The end-level KPI in the network is cumulative effect of a number of PIs. For example, LTE end user throughput depends upon a number of factors e.g. session establishment time, performance of the multi-antenna technology at the UE and base station (BS), block error rate at air interface, delay of cell transition during handover, available bandwidth per user, end-to-end performance of the CN,the delay associated and TCP/IP issues at the end server. The identification of affect of a number of PIs over individual KPI helps to observe not only the involved parameters and their strength but also correlation among them. Self organising networks (SON) [12, 13] can utilise the correlation information to design self-coordination solutions avoiding conflicts among the network functions.

4. *Network Layer Based Classification:* This classification involves data generated at various layers of the network like UE layer, RAN layer and CN layer, detailed network layer data is presented in Table 2.1. For more detailed description please see [13]. Authors in [14] characterises cellular data (such as signalling data, traffic data, location data, waveform data and heterogeneous data) as big data and provide example platform for their analysis. All the above data can be utilised in a number of ways to optimise the network performance from various perspective.

As the network complexity grows with advent of 5G and beyond, the applications of the wireless cellular data will unfold limitless. In the following some exemplary new utilities of wireless cellular data are discussed in the context of wireless network optimisation and management.

2.3 Machine Learning New Utilities

The major capacity gain in 5G is expected through base station densification and improvement in network-wide efficiency. To achieve that, one of the solution is to equip the network with SON functions for almost all functionalities [12, 13]. Such as configuration of new nodes, optimisation of network parameters and resources, recovery of faults, coordination among network parameters and functions etc. These SON functions can be potentially enhanced through data driven machine learning as recently proposed in [13]. In the following some of the potential utilities of data driven machine learning in cellular network are described in the context of 5G.

2.3.1 End-to-End Network Visibility

As discussed, now, data can be collected, combined and analysed from each source in the network. Since machine learning and analysis can be performed on entire data, it enables to visualise connections and correlations among all parameters and KPIs from different departments. For example, optimisation department might require to add bandwidth to increase throughput. This change will have impact on many other departments such as finance department (in terms of Operational Expense (OPEX)) and marketing department (in terms of number of users and required services). Thus, analysis of all data connects each dot throughout the network enabling a transparent end-to-end view of the network. This helps the network planner with intelligent and efficient long term planning..

2.3.2 Self-Coordination

As discussed earlier, WCN data can be exploited to find the correlation among the KPIs and the PIs and network parameters that affect these KPIs and PIs. Although third generation 3GPP has already highlighted the need for self-coordination among SON functions (Self-configuration, Self-optimisation and Self-healing), SON function coordination has been overlooked in the networks up to 4G. However, recent studies [15, 16] due to the increasing number of parameters and functions in the current and emerging 5G architectures, the likelihood of conflicts among the functions are extremely high. By revealing the functional correlations among parameters PIs and KPIs through data driven machine learning, intelligent self-coordination models can be developed to avoid conflicts for smooth and efficient network operation. Self-coordination will become more crucial in ultra-dense deployment where thousands of nodes will be operating with overlapping and conflicting interests where data driven machine learning will play the role of the negotiator.

2.3.3 Long Term Dynamics of the Network

The SON solutions in the current networks respond to the short-term dynamicity of the network such as slow fading, user mobility, etc. although the gain has been smaller. Data driven machine learning will incorporate long term dynamicity of the wireless environment, e.g., dynamic changes in user density in a particular region, prediction of user mobility patterns on long-time scale, data usage patterns over days etc. This will add to the network efficiency achieved through current SON techniques.

2.3.4 Faster and Proactive Network

In the current network, the SON solution comes into play after the problem has occurred. It takes time in diagnosing, detecting and compensating which can range from the order of minutes to days. With this reactive approach, it is difficult to achieve the desired level of low latency in 5G networks. Instead, the network should behave proactively, in the sense that it can predict the future network behaviour (possible fault, coverage holes, network congestion etc.) and user behaviour (mobility pattern, next cell, data rate and QoS requirements) by learning from the historical data. In this way, proper action can be taken beforehand to avoid the faults occurrence. Through analysis of past data and building intelligent prediction/forecasting models, cellular networks can be made proactive and quicker in response. Apart from network behaviour, users future behaviour (mobility pattern, data usage pattern) can also be predicted for resource allocation optimisation etc. Such prediction/forecasting through machine learning is also referred to as predictive modelling.

2.3.5 Smart Caching

To meet the unprecedented growth in traffic, the available network resources as well as the innovation in PHY/MAC layer technologies fall far behind the target 1000x capacity gain. Caching of popular contents at the edge of cellular network can help bridge the traffic vs. capacity gap. Caching is identified as one of the indispensable approach towards a successful 5G. On one hand, data driven machine learning will reveal users context information such as location, network usage pattern, social circle etc. On the other hand, data driven machine learning again can be utilised to evaluate spatiotemporal popularity of contents for individual as well as group of users. Having known users context and content popularity information, Mobile Network Operators (MNOs) can proactively cache popular contents at the required edge nodes (Base Station (BS) or user devices) to provide services with improved QoE. This proactive caching will have two-fold benefits. First, it will reduce the peak traffic demand reducing back-haul bandwidth requirement. Second, it will reduce Total Cost of Operation (TCO) by avoiding access of contents from the origin server (that requires Internet) every time the content is demanded.

2.3.6 Energy Efficiency

If the current strategy of keeping all BSs always ON is continued, it would be impossible for the networks to meet the energy saving target in 5G, or even be economically viable due to substantially increased BS density. In the dense deployment scenario, data driven machine learning can enable the network to adjust the configuration of cells that are most appropriate for serving given distribution of users in time and space. Additionally, by predicting users mobility pattern and network utilisation, MNOs can activate a dormant cell whereas send the unused active cell into dormant state. Thus, using data driven machine learning, cell switching strategy can be changed from always-on to available on demand, preventing wastage of energy.

2.4 Existing Applications of Data Driven Machine Learning

As identified in section III, there are numerous parameters in the network which are reported and gathered at/from the UE, RAN and CN layer of the network. By analysing them, the network performance can be continuously monitored, instantly optimised and proactively protected from possible fault occurrence. The Table 2.2 provides the summary of the literature on research various works that involve data driven machine learning approach for network optimisation, fault detection / prediction, Self-healing and mobility prediction. Different network measurements, parameters and statistics used for network analysis are also shown.

1. Network Optimisation

   The conventional way of estimating cellular coverage is through sophisticated planning tools that utilise building and terrain data combined with appropriate propagation models. However, those estimates are not accurate, resulting in coverage holes. To fill the coverage gaps, network operators perform drive tests which are costly, or they perform network simulation which is partly unreliable, or they use KPI approach which is not precise enough. In order to overcome these inadequacies, study of Minimisation of Drive Tests (MDT) was started as a 3GPP work item in release 10 for 3G, High Speed Packet Access (HSPA) and LTE [17] [18]. The MDT report is proved to be extremely useful data

Table 2.2: List of studies with key SON issues they addressed using data-driven machine learning

Parameters/ References	Coverage optimisation	Cell Outage Detection	Use MDT reports	Use RLF reports	Use RSRP statistics	Use NCL statistics	Use RSRQ statistics	Use HO statistics	Use call trace data	Use TX power	Detect Sleeping cell	Use Antenna Tilt	Planning optimisation	Resource Allocation	Mobility Prediction	User data location	User Density	Fault Prediction	Self healing
[14, 17]	✓		✓										✓						
[18]	✓																		✓
[19]	✓			✓															
[20]		✓			✓														
[21]			✓			✓													
[22]						✓	✓	✓											
[23]					✓		✓	✓	✓										
[24]									✓										
[25–28]		✓																	✓
[25, 26]			✓																✓
[29, 30]											✓								✓
[30]						✓		✓											✓
[27]																		✓	
[16]		✓				✓													
[31]		✓																	✓
[20]		✓		✓															
[32]		✓	✓							✓		✓						✓	
[33–35]		✓			✓			✓				✓						✓	
[28, 36]									✓				✓	✓	✓				
[37, 38]													✓	✓	✓	✓			
[15, 16]								✓					✓	✓	✓				
[36]									✓						✓				
[28]															✓				
[39, 40]															✓		✓		
[15,41–47]									✓						✓	✓			
[48]																		✓	
[49]								✓										✓	

for coverage analysis. The authors in [19] [20] enhance the coverage estimated through MDT reports using Spatial Bayesian prediction scheme. A data driven machine learning approach for improving coverage analysis has been discussed in [21]. The authors utilise BSSAP mobility and Radio Resource Management (RRM) messages collected between the Base Station Controller (BSC) and Mobile Switching Centre (MSC) nodes to identify inter-technology handovers from 3G to 2G. The point of handover indicates 3G coverage discontinuity (holes). The coverage gap found through above experiments are found to be more accurate with more detailed coverage map than those obtained through drive test and tracking BS KPIs. Coverage hole detection and optimisation is also performed using RLF report generated from UE [22]. Reference Signal Received Power (RSRP) statistics data has been analysed in [23] to determine cell outage in a control-data plane separated heterogeneous network. In addition to coverage, machine learning has also been applied to ensure transparent handover. Article [45] proposed a model that utilises detected cell set to improve automated Neighbour Cell List (NCL) performance. All the cells in the detected set are considered as part of NCL and thus they predict all those cells' performance. This helps in identifying missing neighbours and thus avoid walk/drive testing. Authors in [24] applied machine learning on Performance Management (PM) data such as handover success rate, received signal quality to self-optimise NCL. The optimised NCL is further utilised for building handover Self-optimisation model. Usually, the intra-frequency handovers in LTE are performed based upon RSRP measurement. Article [50] discusses suitability of handover based on Reference Signal Received Quality (RSRQ) measurement in a situation when severe interference causes RSRP to go below threshold.

2. Fault Detection and Self-healing

Recently, big data analysis combined with machine learning techniques have drawn attention for development of Self-healing mechanisms. Researchers have proposed clustering [30] [31] [32], classification [33] and knowledge based Machine Learning (ML) algorithms [31], for cell outage detection. MDT measurement reports from the network have been utilised in [30] [31] from which KPIs relevant to network performance are gleaned. Due to large dimension of KPIs, authors perform dimensionality reduction. Apply different machine learning algorithms such as k-nearest neighbour, Local Outlier Factor (LOF), One Class Support Vector Machine based Detector (OCSVMD) etc., to automate the detection process. Data mining frameworks have been developed to detect Random Access Channel (RACH) to further detect sleeping cell in [35]

[37] which is hardly traceable by network operators. The authors used n-gram analysis technique that also includes dimensionality reduction and classification of data, except in [37] where they use k-nearest neighbour technique to identify abnormal behaviour in the sequence of network events collected through MDT report from the UE. The key idea of the system is that it provides a set of abnormal users which eventually reveals location of problematic cell. The article in [38] also proposed UE measurement based cell outage detection combined with location information.

For multi-tier networks, the authors in [39] present a cooperative architecture in a self-organised femto-macro cellular networks. The objective is to investigate intra-cell correlations of RSRP statistics in space domain in order to make a trigger decision and then extracts correlations of inter-cell RSRP statistics in both space and time domains to improve the detection accuracy. In [23], k-nearest neighbour is utilised to detect and locate cell outage in macro-pico heterogeneous networks. Further, Onireti et. al. in [15] [16] proposed a complete cell outage management framework that includes both detection and compensation mechanism in control and data plane separated architecture. For control plane, MDT reports are analysed applying k-nearest neighbour and local outlier factor algorithm to detect the outage. For data plane, grey prediction model has been proposed. For compensation, an actor-critic reinforcement learning algorithm that optimises Coverage and Capacity Optimisation (CCO) of the identified outage zone by adaptive adjustment of the antenna gain and transmission power of the BSs in that plane. Other cell outage compensation frameworks have been developed in [51] [41] [44]. For compensation, the authors' utilisation of control parameters such as RSRP, antenna tilt, uplink target received power can be effectively utilised to compensate the cell outage.

3. Mobility Prediction

In the literature, researchers heavily utilise wireless cellular network data such as Call Detail Record (CDR) [46] [52], Global Positioning System (GPS) location data [53] [54], handover information data [55] for analysis and prediction of user and network behaviour for further planning and resource allocation. Article [46] analyses CDR to find the distance travelled by hundreds of thousands of population every day in three major cities of USA. The article [52], on the other hand reveals how many people travel from one place (origin) to another (destination) in a given time. Users density in a particular area and their activities can be monitored using CDR as proposed in [49] [56]. Similarly, city block's properties have been analysed utilising cellular data [48]. These predictions have been useful in smart city planning from various perspectives including communication network, transport, ecology and environment. Cellular data has been extremely useful in predicting mobility. Authors in [57] build cluster of cell sequences where users were present in to represent physical routes and the destination probabilities is derived from this sequence. Article [58] classify users into residents, commuters and visitors by analysing their Global System for Mobile communication (GSM) call profiles, and the categorisation information is further utilised to predict their mobility pattern. The authors in [59] utilise stay places and trip pattern of the users to find the spatial route, whereas they find the temporal pattern by analysis of duration of stay in a cell coverage region. The spatial information based prediction model utilises sequential models to predict next stages. Article [60] utilises CDR for extracting movement information and predict the future movement using Hidden Markov Model (HMM). Authors in [27] also utilised HMM to predict user's next location. Authors in [61] utilise handover information using Markov chain to predict the next handover cell of the user. Markov chain based model is proved to be more accurate for predicting whether a user will move from one region to another in next period in [62]. Methods like recursive least squares lattice adaptive filter on mobility traces have also been utilised for future location prediction [63].

4. Traffic Prediction

   An analysis was performed on data collected from devices using Android, iOS and Windows platform, to study the traffic and user application behaviour dynamics. The study reveals that all of the platforms incur high traffic during night. Windows have highest traffic variance while browser and instant messengers attracts more users thus achieving large user ratio. Another analysis performed on 3G traffic also reveals that the traffic in night consumes the highest bandwidth. Additionally, their finding shows user number distribution, user request distribution and location characteristics [64]. The study performed in [28] extract human mobility and traffic utilisation pattern from a MNO's perspective. Importantly, their analysis reveals that consumed traffic is strongly correlated with mobility and therefore, they exhibit distinct temporal pattern. A framework for rating important mobility features that influence application usage behaviour is presented in [65]. The rating (correlation) is further utilised to forecast application utilisation behaviour of individuals as well as crowds. Data from Twitter and 3G networks have been exploited to forecast traffic requirement in Satellite Communication Network (SCN). In addition to users, MNOs' traffic usage pattern and characteristics of data services can also be revealed by using machine learning [66]. The authors in [67] extract the distribution and temporal patterns of different services in wireless cellular data in terms of service request times and service duration.

2.5 Further Unleashing the Potential of Data Driven Machine Leaning in WCN

Managing KPI complexity in a single-technology environment is difficult enough. In 4G the number of parameters per node rose to around fifteen hundred from five hundred in 2G [13]. Even if the same trend continues, the number of parameters in 5G and beyond will be significantly high. Additionally, in a cellular ecosystem encompassing all technologies up to 5G, the level of complexity will increase dramatically. Therefore, in this section, first, the end user KPIs are defined and then it is explained how these are computed conventionally. Then it is discussed how data driven machine learning can be leveraged to gain insights by digging down to the granular level.

2.5.1 End User KPIs

The service experienced by the end users is measured using a number of PIs. Multiple PIs are aggregated to form one KPI. According to 3GPP standard, there are six main KPIs which directly influence network performance and end user experience. These are *Accessibility, Retainability, Integrity, Mobility, Availability and utilisation*.

- *Accessibility* defines how easy it is for the user to obtain a requested service within a specified tolerances and other given conditions. Session setup time, call success rate etc. are example of accessibility KPIs.

- *Retainability* defines the capability of the network to continuously provide requested service under given conditions for the desired period of time. An example KPI in this context is session abnormal release rate which means poor retainability for a network.

- *Integrity* means the degree to which a service is provided without excessive impairments, for example uplink/downlink throughput, latency, packet loss etc.

- When user is moving from one cell to another, handover performance and other mobility related performances are measured through *Mobility* KPI. LTE Handover success rate and Inter-Radio Access Technology (Inter-RAT) handover success rate are examples of Mobility KPIs.

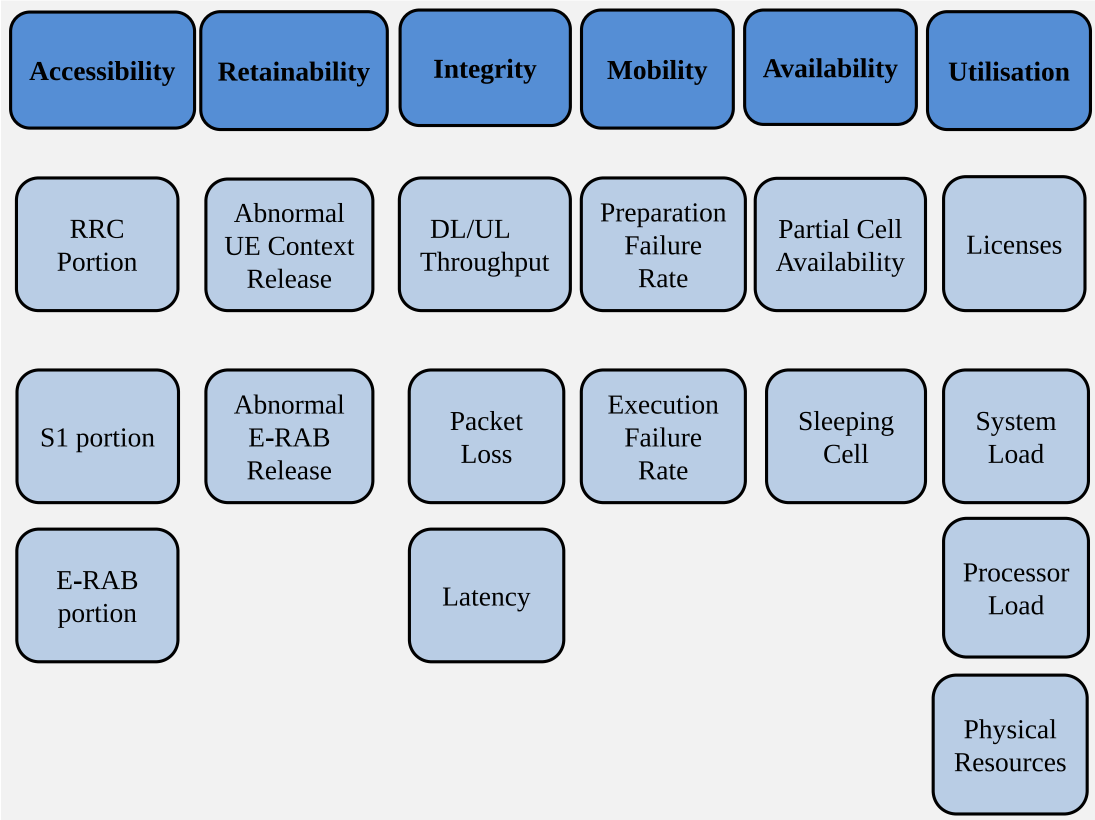

Figure 2.4: Standard KPIs and their immediate dependencies.

- As the name suggests, *Availability* means the percentage of time a cell is considered to be available. In some cases, although the cell may seem to be up and running, it can be sleeping or in outage condition due to some anomaly or disabled manually because of operator's maintenance procedure. In such a case, poor network availability or complete outage can occur.

- Whereas *utilisation* refers to the simultaneous usage of the network resources without affecting the end user experience.

2.5.2 Measurement Composition of KPIs

These end user KPIs discussed in previous subsection further depend upon various others PIs. The high level blue print for six standardised KPIs and their immediate PIs dependencies is shown in Fig. 2.4. The insights obtained by dissection of the various other KPIs help to better couple users' perceived QoS and QoE with the operator's Operational and Business Objectives (OBOs). The PIs which comprise the KPIs, themselves depend upon plethora of lower level measurements. The measurements are shown for accessibility, utilisation, mobility, integrity, retainability and availability

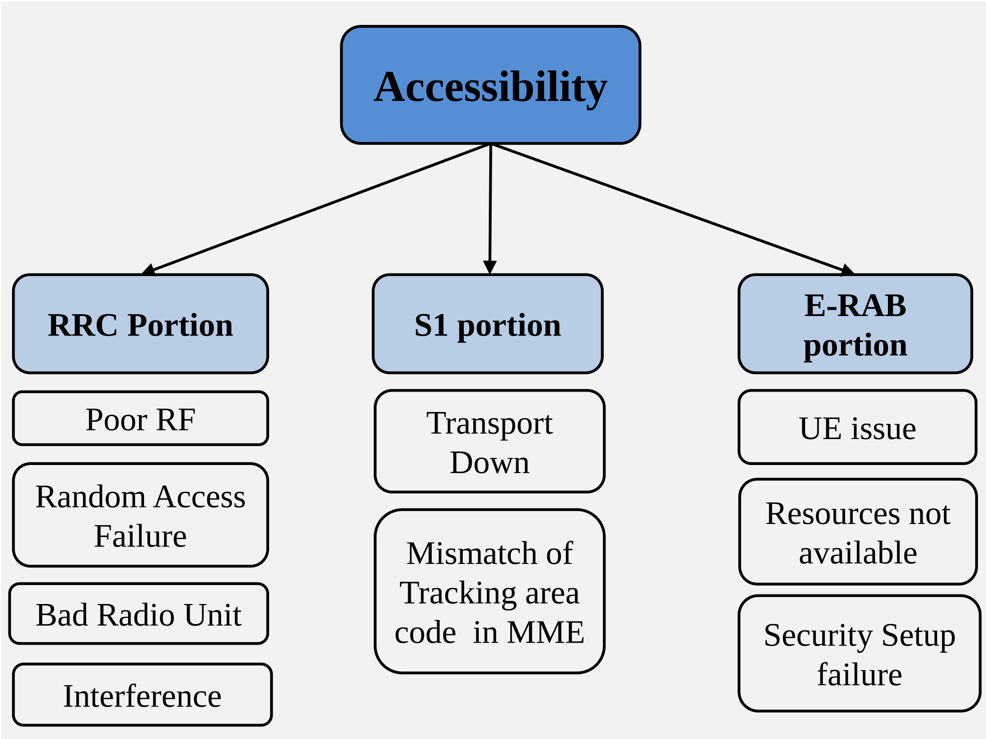

Figure 2.5: Accessibility KPI related PIs and measurements

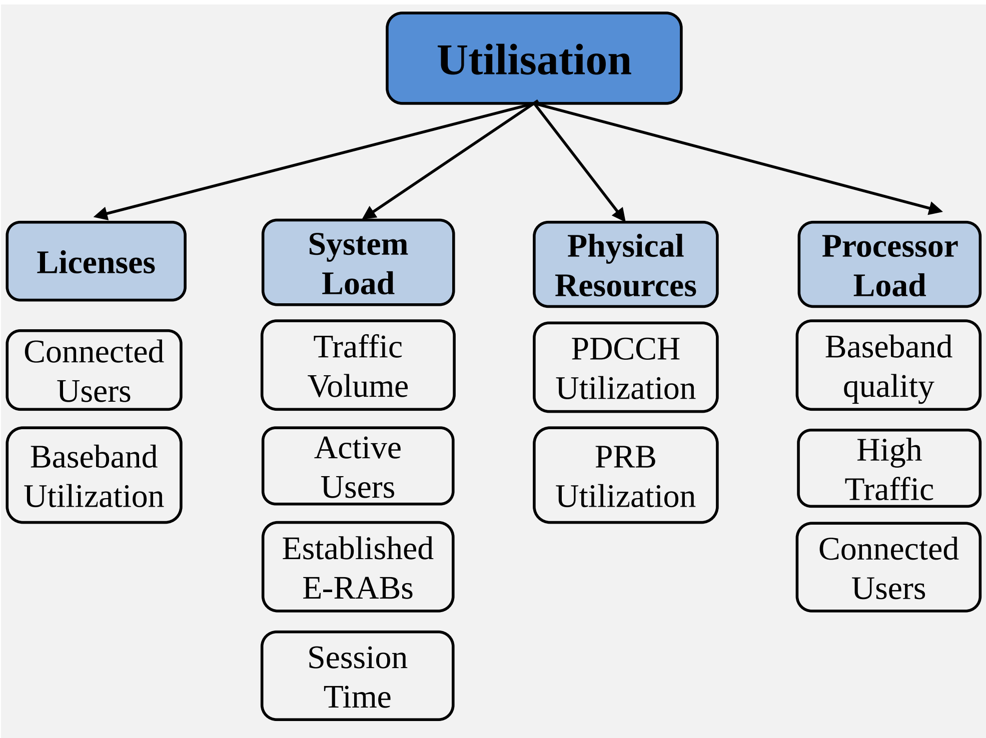

Figure 2.6: Utilisation KPI related PIs and measurements

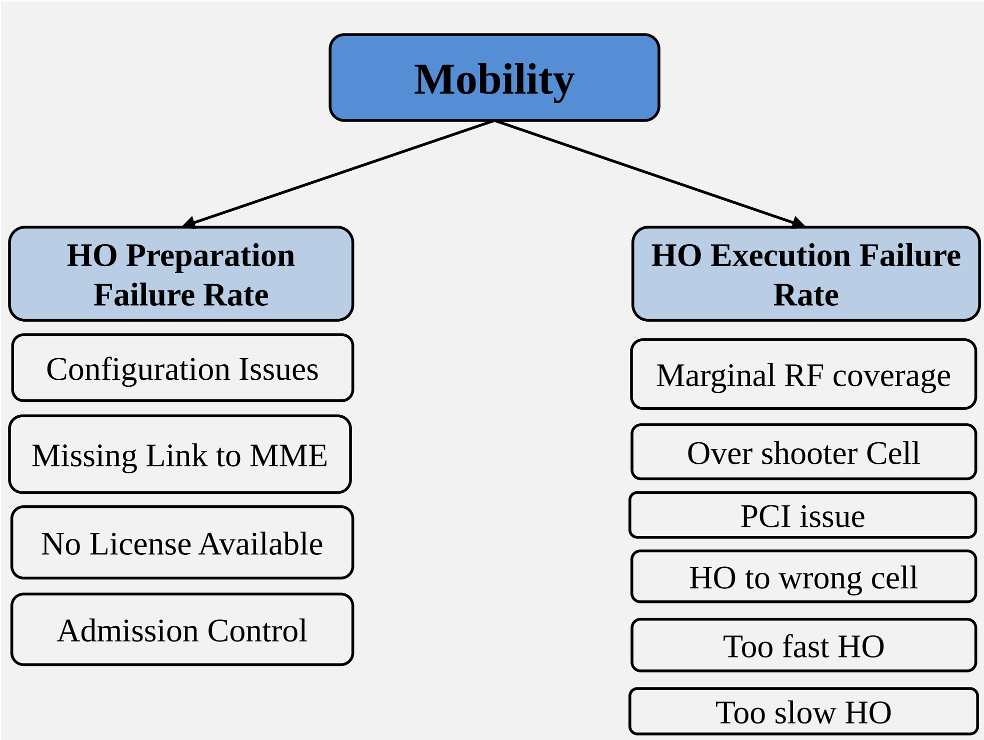

Figure 2.7: Mobility KPI related PIs and measurements

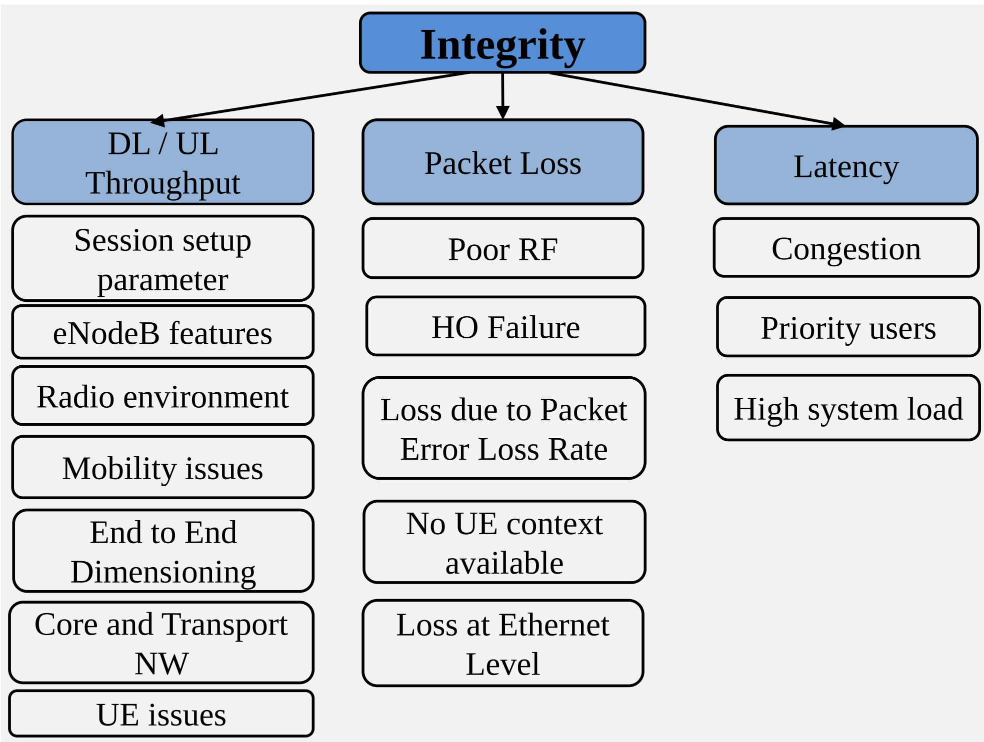

Figure 2.8: Integrity KPI related PIs and measurements

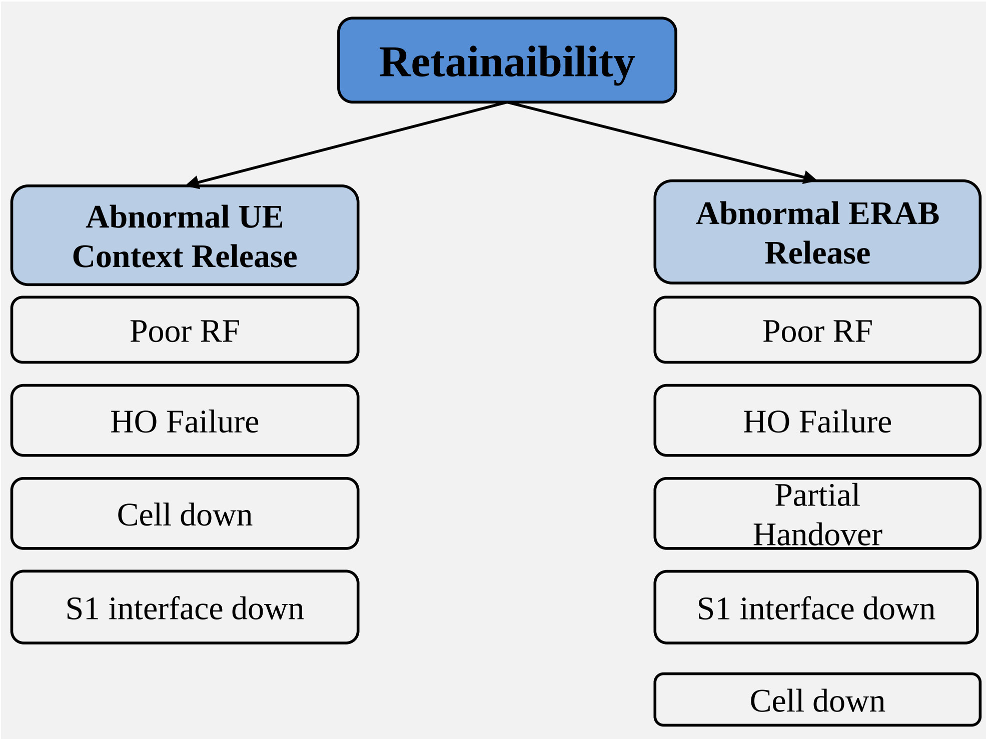

Figure 2.9: Retainability KPI related PIs and measurements

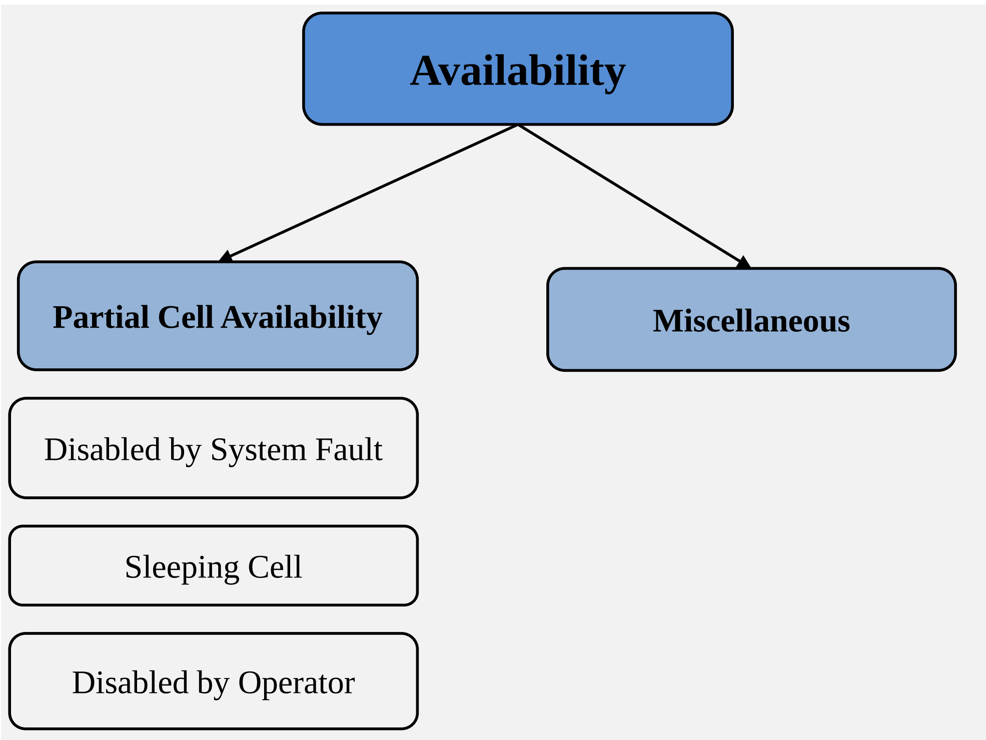

Figure 2.10: Availability KPI related PIs and measurements

KPIs in Fig. 2.5, Fig. 2.6, Fig. 2.7, Fig. 2.8, Fig. 2.9 and Fig. 2.10 respectively. It must be noted that all these grass root level measurements are by no means representing an exhaustive list of all the measurements needed, in order to compute a specific KPI. It represents what major specific measurements comprise a given KPI. In addition, these tree diagrams informs us that how deeper resolutions in performance measurements can help in better diagnosis of the real root cause affecting a specific OBO or KPI. Having extensive measurements can provide the correct decision to make for a specific parameter in order to improve respective KPI. These measurements are the building blocks of KPIs and end users' experience. In order to realise self organised networks enabled by real data driven machine learning for 5G and beyond wireless networks, these measurements should be available and standardised for user centred KPIs estimation. The true value of such self organised wireless cellular networks of the future can only be realised when all the measurements are recorded, mined and analysed to perform a KPI computation analysis. In order to demonstrate this approach, as an example a user centred data driven machine learning approach for the KPI estimation is presented in subsection 2.5.4.

2.5.3 Conventional Approach of KPI Estimation

In conventional approach, a range of KPIs are evaluated to ensure the target network performance. Normally, the selected KPIs are uncoordinated and do not reflect the user experience. It requires significant effort to improve one set of KPIs at the expense of others. The correlation among different KPIs is not taken into account in conventional approach. As a result it ends up disturbing one KPI while improving the other. In addition, reaching the set targets for these poorly selected KPIs may have little impact on overall system efficiency, performance and user experience.

In conventional KPIs estimation approach, multiple PIs are integrated into a single KPI to effectively indicate the end user performance. Even though, the performance statistics for all PIs are recorded for investigative and troubleshooting purposes. But, they are not utilised in real-time for performance improvement on regular basis. This approach has been adopted mainly to reduce computational overheads of storage, processing and analysis of the data generated by myriad of PIs based measurements. Nevertheless, in such aggregation based approach, the underlying individual PIs are not effectively utilised to identify the exact root cause of the performance degradation.

Conventional machine learning schemes operate on user data about user's decisions, actions and social interactions. These data are highly dynamic for the global optimisation at macro level. However, it ignore the user's' local interactions and contextual information at micro level. In short, the conventional approaches are network-centred. In conventional approach, a good KPI value may not reflect a great

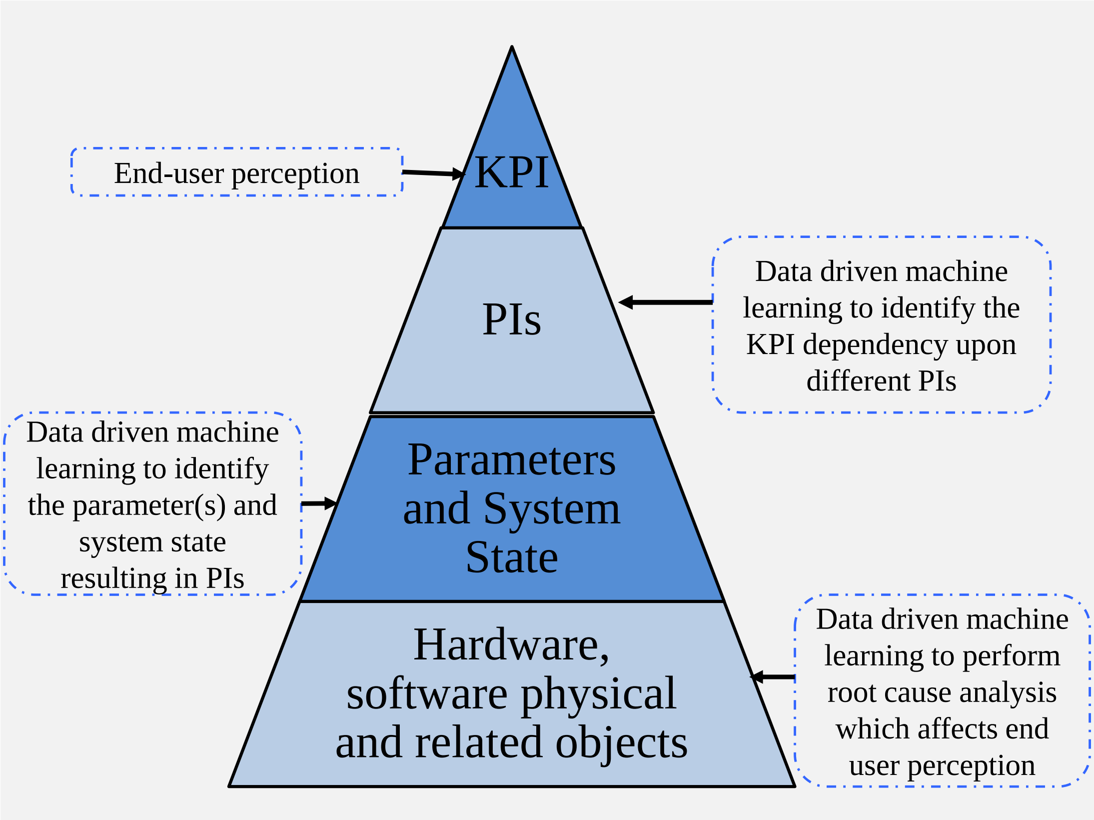

Figure 2.11: Machine learning for user-centric KPI hierarchy and dependencies

user experience in reality. For example, an LTE radio bearer can be dropped when an always-on LTE device is not transmitting data. Measuring such a KPI does not offer any real understanding of user experience, as sessions are quickly reestablished when needed and any delay incurred goes unnoticed by the user.

2.5.4 Machine Learning Approach for KPI Estimation (User Centred Approach)

Data driven machine learning approach can help to go to a much deeper level to identify root causes that affects individual PI. It allows to go into the granularity of the parameters, observe their individual and compound effect on end user KPIs. It also allows to identify the correlation among parameters to establish relation of each individual parameter with multiple KPIs.

Data driven machine learning shifts the network centred approach of conventional KPI estimation to user-centred approach for 5G and beyond networks. In a user centred approach, all the factors affecting the end level KPI are analysed from the user and OBO perspective. The data generated by each event is categorised into different classes of data as previously discussed in section III. The data category identification helps in

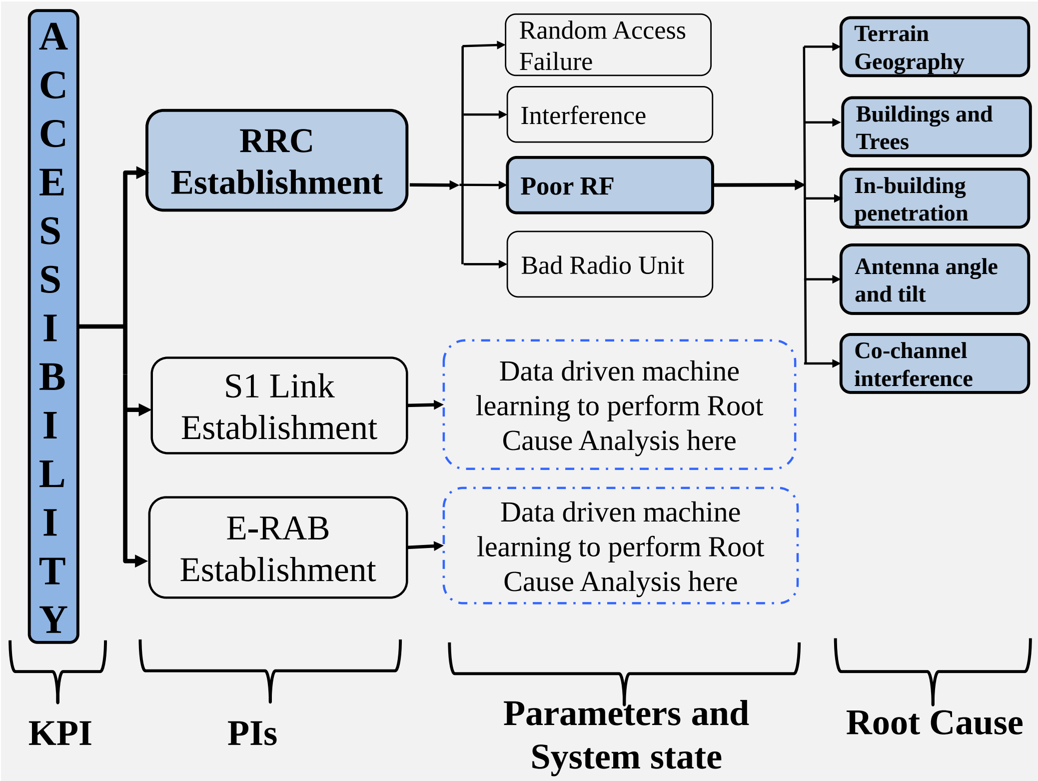

Figure 2.12: User-centric Accessibility KPI computation analysis using data driven machine learning

deciding the relevant data analysis tools and procedure to be applied for each event. The high level procedure of data driven machine learning for the user centred KPI estimation is shown in Fig. 2.11

As an example for demonstration, Fig. 2.12 describes the accessibility KPI being estimated by machine learning as a user centred approach. Accessibility KPI depends upon three major PIs: RRC establishment, S1 portion and E-Radio Access Bearer (E-RAB) portion. If accessibility KPI is poor on account of RRC phase establishment. Data driven machine learning helps us further identify, what is the cause of poor accessibility, what parameters affect it, even so it ultimately leads to the real root cause of the problem causing poor accessibility KPI, as depicted in Fig. 2.12.

Let's say poor accessibility KPI is due to bad RF conditions. Fig. 2.12 further breaks down the cause of poor RF in various categories which may cause it. It can be seen from Fig. 2.12 that poor RF condition further depends upon many other factors. These include terrain geography, non-terrain obstacles like building structures and foliage, building penetration and antenna angles. The evidence is clear that one end-user KPI is dependent upon a number of the other parameters which are further correlated as show in Fig. 2.12. Using data driven machine learning approach, the dependency of top parameter can be traced back to the intermediate and root cause

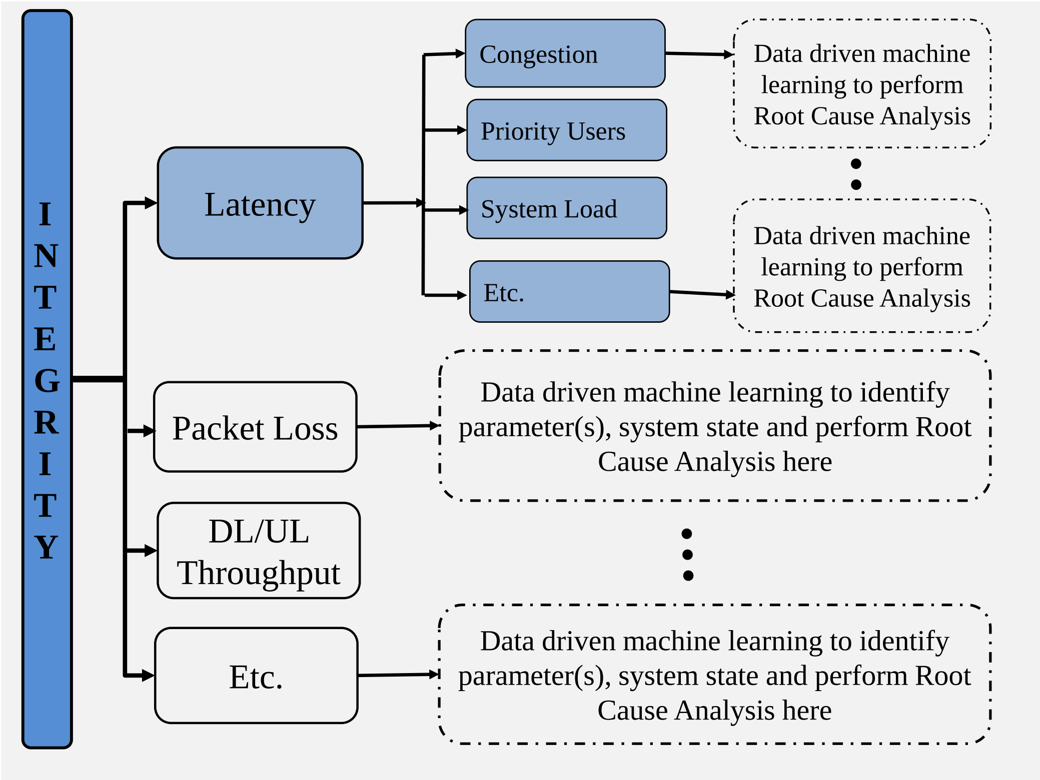

Figure 2.13: User-centric Integrity KPI computation analysis using data driven machine learning

parameters. Hence, the power of network analytics can be used to identify and cross correlate the inter dependencies of one PI over the other PIs. In addition using data driven machine learning real time effect of a single parameter and /or root cause on the end user experience in the form of a KPI can be predicted, therefore it is termed as user centred approach.

Similar to Fig. 2.12 the hierarchical KPIs estimation for Integrity KPIs is identified in Fig. 2.13. Using data driven machine learning methodology described in Fig. 2.12 and Fig. 2.13, real time anomaly identification and KPI improvement can be done in a user centring way. Similarly the blue print for user centred hierarchical KPIs for Retainability, Mobility and Availability can be identified With the help of data driven machine learning.

2.6 Opportunities and Challenges

The key common concept behind the data driven machine learning is the analysis of entire data space which further enables to find inter-connection, correlation, conflict and coordination among the parameters and the associated entities in the network. These relations are exploited to leverage the massive potential hidden within and behind the massive data generated in the wireless cellular network. In fact, the researchers have realised only those potentials that they have been able to think and relate with information so far extracted from the data. As the necessities will unfold with time, so will the potentials hidden behind the data. However, to achieve those benefits, many challenges have to be overcome. In the following,the opportunities that can be availed and the associated challenges especially with regard to 5G and beyond networks are discussed.

2.6.1 Opportunities

Flexible Network Functionality and Deployment

Network function virtualisation and software defined networking have enabled decoupling of software and hardware functionalities in the network. Through data driven machine learning, regional user and service requirements can be predicted, based on which network functions can be deployed dynamically. For example, by analysing user and traffic information in an area, operators can deploy Multimedia Broadcast Multicast Service (MBMS) functionality module where there is much broadcasting requirement. The analysis of the traffic and user attributes data can help predict the traffic characteristics in a region based upon which hot spot Access Points (AP) can be deployed or dynamically switched to meet the traffic requirement. This would be extremely useful for 5G dense heterogeneous network. Standardisation initiatives of networking hardware and protocols across the industry can also enable flexible sharing of networking facilities among multiple operators.

Internet of Things (IoT)

Global telecommunications companies are seizing the opportunity to provide end-to-end communication over internet for IoT environments. IoT domain requires a range of data rate and accordingly requires the processing power. For example, energy meter at every house does not require high data rate and also they are less sensitive to delay. On the other hand, medical appliances connecting to internet require prompt actions without compromising delay. Through data mining and machine learning, the data generated by Internet of Things (IoT) devices can be processed to facilitate sufficient resource and also clean data to provide the concise information. For a machine in America communicating with a machine in Asia, machine learning will facilitate sufficient bandwidth and clean stream of data that conveys the essence of what is to be communicated.

2.6.2 Challenges

Data driven machine learning in cellular network has its own idiosyncratic challenges. Some of these challenges are briefly outlined in the following.

Need for Paradigm shift from conventional data analytic to big data analytic

In conventional data analytics, analyses are performed on sample data selected from whole data or population using different statistical techniques, commonly random selection. A statistical inference is made towards the behaviour of overall population based on results of the sample analysis. Conventional data analytic may help to infer some useful information about the whole data or phenomenon under observation but it fails to provide the real and exact picture of the problem under consideration or an insight to set of possible causes and effects. Whereas, modern day data driven machine learning can take the whole data into consideration for analysis and intelligence gathering that helps to have comprehensive and real information about complete population. The analysis using entire data makes us worry-free about some individual data points that may bias the overall analysis. Data driven machine learning also allows testing of new hypothesis at many levels of granularity, which were not possible by using conventional data analytics tools. Correlation is another very important statistical aspect of data analytics which help to extract valuable information about the relationship among the various factors and features of the data. In case of small-data, correlations are useful but in the context of big data they outshine. Through correlation, we can glean insights more easily, faster, and clearer than before. In contrast to conventional data analytics, big data analytics enabled by machine learning at scale can identify and explain multidimensional correlation across the various features, horizontally and vertically, which is the backbone of predictive capacity of a data driven machine learning tool. Information about correlation among the data features play a crucial role to establish coordination among the network functions in a self-organising network.

Standardising procedures for data acquisitions and transportation

In the world of cellular networks, multiple vendors are contributing through network devices corresponding to different layers of the network. Thus inter-operability becomes necessary. In such a scenarios, standards must be developed to specify how real network data will be transferred among different platforms. The real network data comes in structured and unstructured forms such as picture, videos, audio, click streams etc. Standard procedures are, therefore, needed to encode and transport data within and across different networks. This will enable not only multi-vendor operability but also enhance QoS management with low latency and high fidelity.

Enhancing speed of the data processing

In cellular networks, the requirement for high speed processing can be viewed from two different perspectives. Firstly, the hyper competitive business environments, companies not only have to find and analyse the relevant data they need but also they must find it quickly, however, the challenge is going through the sheer volumes of data and assessing the level of detail needed all at high speed. Secondly, to pro-actively meet the various demand by users and rapidly changing network environment, the data collected at different layers of the network need to be promptly analysed and necessary resource allocated to avoid delay and to avoid fault occurrences.

Integrating of inter-department data

Although data driven machine learning offers big opportunity, the integration of data from disparate sources has remained a huge challenge in cellular networks. With the increasing volume and variety, it is difficult to integrate inter-department data and extract useful insights from them. For example, the data obtained from the network optimisation department may need to be complemented with data from customer CRM. CRM data further needs to include all the information related to customers. It is challenging to manage and control data quality so that it could meaningfully connect well-understood data from the data warehouse with the data that is less understood. Thus data integration requires further research to leverage network analytics.

Challenges Pertaining to Data in data driven machine learning

- *Data Cleaning*

  Although it is assumed that data will be processed quickly, the value of data for network performance optimisation process will be jeopardised if the data is not accurate and timely. This is a challenge pertaining to machine learning in any field and it is more pronounced with rise in volume. To address this issue, there need to have a data governance or information management process in place to ensure the data is clean. Additionally, 5G networks prefer to have a proactive method to address data quality issues to avoid incorrect interpretation of data and model using the data.

- *Ensuring privacy of data*

  Machine learning usually involves data related to users' location from which their movement behaviour, their shopping interests as well as their future locations can be predicted. Consider that some retailers have used data analysis to predict intimate personal details such as the due dates of pregnant shoppers. In such cases subsequent marketing activities resulted in having members of the household discover a family member was pregnant before she had told anyone, resulting in an uncomfortable and damaging family situation. Moreover, with the sufficient amount of data and powerful machine learning tools, it could become impossible to completely remove the ability to identify an individual if there are no rules established for the use of anonymised data files.

- *Enhancement of security of data*

  With the emergence of small cell technologies and due to ultra-dense BSs requirement in 5G, cloud RAN has become important for effective control and utilisation of resources among the small BSs. Security along with the privacy issues are magnified by the 4Vs of big data, such as large scale cloud infrastructures, diversity of data sources formats, streaming nature of data acquisition, and high volume of data integration etc. The traditional security mechanism which are tailored to securing small scale static data are inadequate. For example, machine learning for anomaly detection would generate too many outliers. Thus enhancement in security mechanism considering the challenges imposed by nature of large scale data.

2.7 Conclusion

Data-driven machine learning holds immense potential to unlock the hidden value within wireless cellular network data. By incorporating data from across the network into the analysis, it becomes possible to uncover patterns, connections, and correlations that are not evident through conventional methods. These insights, when combined with intelligent machine learning algorithms, can automate wireless network functionalities, revolutionizing the way cellular networks are operated and maintained.

In this chapter, key data types generated from various network segments have been identified and their potential roles in enabling SON functions have been described. Additionally, it has been demonstrated how these data types can be leveraged to enhance SON functionalities in the context of 5G and beyond networks. Existing literature efforts toward this goal have also been summarized.

Through an exemplary case study, this chapter has highlighted the necessity of decomposing cellular network KPIs into their contributing PIs to realize the envisioned utilities of wireless cellular network data. Actionable insights derived from data-driven machine learning can enable automation of network functionalities, improve resource efficiency, and boost performance. These advancements are critical for meeting the demands of emerging 5G cellular networks, paving the way for a new era of intelligent and autonomous wireless communication systems.

Chapter 3

Machine Learning for Traffic Load Estimation

During the last decade, mobile services have sharply evolved from only cellular network-based services like messages and calls to internet-based services like mobile apps and web surfing on mobiles. On one end such services demand a different set of bandwidths, network protocols and resources for data transmission to cater diversified data types, there on the other end, they have also raised user-network interaction to the highest level ever and this trend is increasing [68]. Data consumption, user interaction with the network and the time spent by users on the cellular network to access Internet-based services, has surpassed the conventional cellular services such as call and Short Messaging Service (SMS). StatCounter, a research company that tracks internet activity globally, concluded that the number of web pages accessed using mobile devices already exceeds the number of web pages accessed from desktop computers and laptops since October 2016, and this trend is increasing [69]. For instance, subscribers from the USA spent almost 90% of their mobile phone time on the mobile internet in 2015.

These statistics clearly show an increasing demand of a huge range of Internet-based services on the cellular network and require the network to be capable to cater a variety of data types with efficiency and better latency [70]. So it becomes crucial for the network to learn user's internet usage behavior and preferences in terms of contents, timings, and vicinity for the provision of user-specific seamless services. Further to it, the future network must be able to predict demand for internet services at different spatiotemporal granularity for better Radio Resources Management (RRM) and pre-emptive measures against key challenges like admission control, traffic congestion etc. To meet that objective, there exists a need to design most efficient and optimal (RRM) algorithms. The analysis of spatiotemporal patterns of internet consumption at higher granularity is also important for the understanding and information management of varying communication level expected among numerous devices locally in future networks [71].

The increased frequency of user network interaction has also led to an activity level of very high granularity over the network with fine footprints of respective activity records, e.g call data records (CDRs). Such CDRs, also provide an opportunity to gather intelligence about users' behaviors and preferences towards different on and off network services. This intelligence can be accumulated by the identification of patterns and correlations in the existing data with the application of data analytics. Such data analytics based cognizance can help to improve overall network performance via Radio Resources Management (RRM) strategies at shorter intervals by making timely autonomous decisions [72]. This kind of network intelligence is a driving factor to make future networks more pre-emptive, autonomous and self-organizing, some of the key features expected in the future cellular (i.e. 5G and beyond) and IoT based networks [73].

Spatiotemporal understanding and prediction of traffic can help optimize resources like switching off certain eNodeB for possible energy conservation. Similarly, timely and accurate traffic prediction can also play an important role in managing operational and quality of services related problems e.g. congestion control, admission control, network bandwidth allocations etc. [74]. In chapter 6 a small cell sleep cycle centered approach [75] is proposed that leverages from spatiotemporal prediction based on same CDR to pro-actively schedule radio resources. Results for this approach show substantial energy savings and reduced inter-cell-interference (ICI), without compromising the users Quality of Service (QoS). Besides the future traffic prediction, understanding of high granularity spatiotemporal traffic distribution and patterns are also important for network planning and configuration for future networks where network densification is seen as a mean to meet the diversified high data demands.

In this chapter, the mobile users' internet usage behavior with respect to time and location is studied with the help of real network data. For the purpose of analysis, actual two months cellular internet activity data for Milan city, released by Telecom Italia, is used [76]. Here, the real internet traffic variance over a network in the spatiotemporal domain is explored at a high granularity level particularly in term of time where the the variance of traffic even within an hour is studied. The detailed study of the users' preferences in terms of the data contents is out of the scope of this study.

Cellular data is a rich source of information for multidisciplinary research and multifaceted decision-making processes. There exists enormous research on cellular network architecture, functionalities, and services. Now plenty of work also exists on the utilization of data analytics for network improvement as it can be seen from

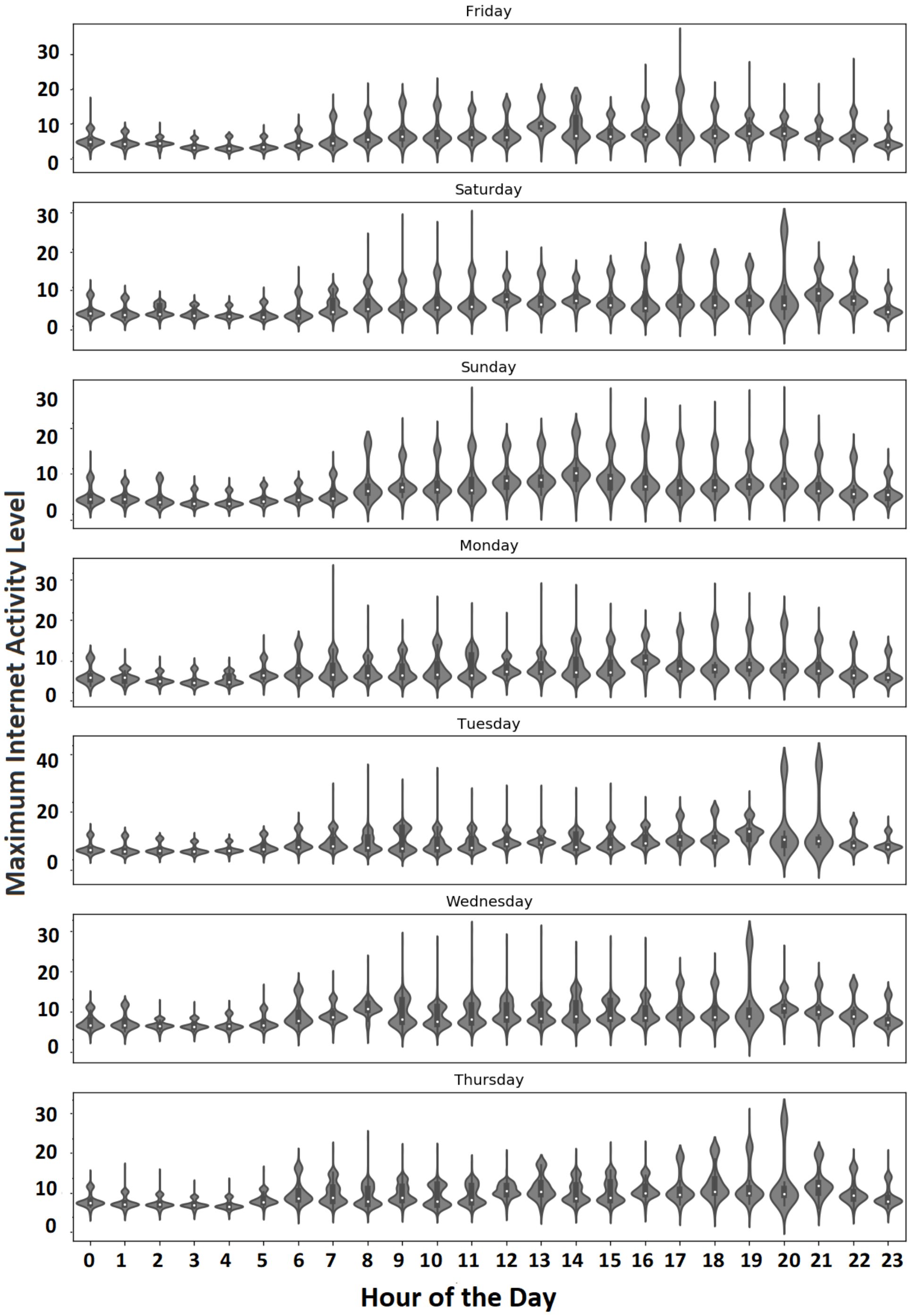

Figure 3.1: Cells distribution with respect to maximum traffic generated on hourly basis in a week

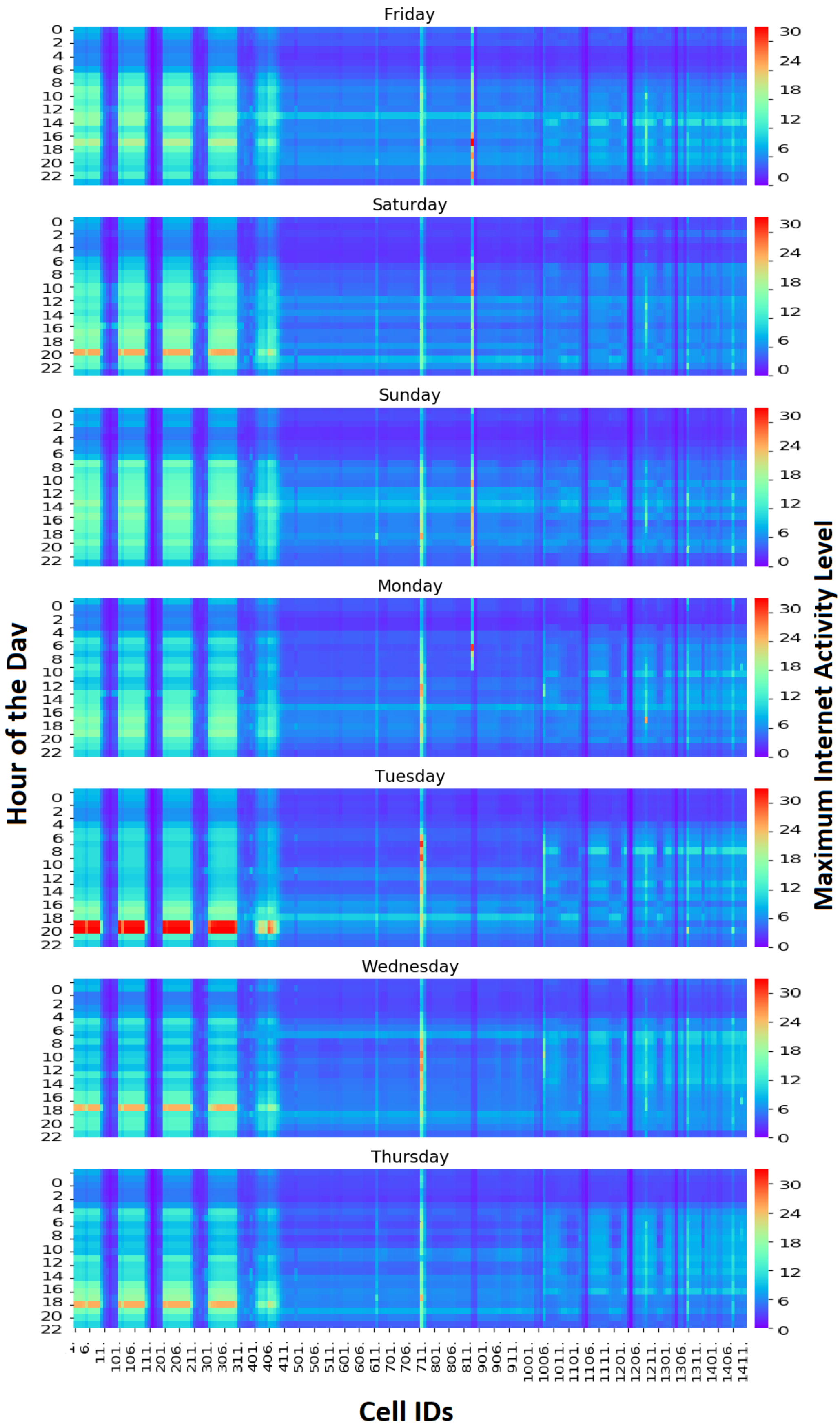

Figure 3.2: Heat map of maximum hourly internet activity in 225 cells over a week

[75, 77, 78], where [78] specifically discuss predictability in networks. Significance of the use of real mobile data in analytics is very high as it captures and exhibits the true feel of actual network behaviour. But research on spatiotemporal analysis and predictability of cellular network traffic based on real network data is very limited and research on spatiotemporal analysis and predictability at high granularity is even rare especially for the internet traffic. Authors in [68] emphasize the need of models to predict traffic demand on short (e.g., minute to hours) and medium intervals of time (e.g., days to weeks) after presenting a detailed review of the literature published in the last decade on the topic.

Another drawback of the existing research is that aggregated hourly activity level is taken into account whereas maximum traffic is more significant for estimation of demand and resource allocation [79–81]. The aggregated or mean traffic is more stable as compared to maximum traffic which has high spatiotemporal variance as it can be seen from Fig. 3.1 and Fig. 3.2. Further, most of the traffic prediction models proposed in the literature are for call and SMS data only and were separately trained and tested at different locations [79–81]. In this research, the proposed model is platform and location independent. The model is trained simply by providing six data points of aggregated activity level, each for ten minutes in an hour, for all cells. In the literature, Artificial Neural Networks (ANN) is one of the most popular non-linear models to forecast complex network traffic and outperform traditional time-series models like ARMA and FARIMA [82]. Studies focusing internet traffic on the cellular network for spatiotemporal analysis and short-term predictability are very rare. In [83], authors have applied deep learning methods for the prediction of internet traffic and results are used as a benchmark in the study presented in this chapter.

In this chapter, a Support Vector Regression (SVR) model is used for the prediction of future internet activity for three different levels, minimum, maximum and mean at high granularity. These levels help to have a basic idea about the activity level in the different cells for a shorter period of time. The performance of the proposed method is compared with the SOTA deep learning methods available in the literature. One objective of this study is to prove that a classical, comparatively simple, SVR model can perform much better than the complex deep learning models for cellular network problems like the one under study here, internet activity estimation at high granularity. Deep learning models are not the optimal solution in all cases, therefore in this study focus has been on these three to compare the performance of SVR with that of the deep learning models for the same data at same granularity and same predicates This is a timely research as future internet activity estimation using data mining and machine learning over a cellular network at high granularity is one of the most important problems for the research community in order to design efficient and intelligent 5G and beyond 5G cellular networks [84]. In this chapter, it is concluded that the proposed SVR based method outperforms SOTA approaches used in the recent literature.

Main contributions of this study include:

- Performed detailed spatiotemporal analysis of real-world cellular internet traffic data, identifying significant patterns and variations across time and location at high granularity.
- Demonstrated the suitability of classical machine learning models like SVR model to accurately predict minimum, mean, and maximum internet activity levels for the next hour using historical data.
- Conducted a rigorous evaluation of the SVR model against other machine learning techniques (e.g., ARIMA, CNN-RNN) and demonstrated its superior performance in prediction accuracy and efficiency.
- Leveraged real-world Call Detail Records (CDRs) from Telecom Italia to validate the proposed model, ensuring practical relevance and reliability.
- Showcased the groundwork for the application of traffic load prediction in enabling energy-efficient cellular network operations, such as proactive small cell sleeping and resource optimization.

3.1 Data Set Description

To study the internet activity dynamics on a cellular network it is of paramount importance to use actual data from a cellular network operator. The internet activity data used in this chapter is obtained from a comprehensive big dataset released by Telecom Italia as part of Big Data Challenge 2014 [76]. The dataset includes CDRs (i.e. SMS, call, and internet activity), precipitation data, electricity consumption data, weather station data and website data for the city of Milan, Italy and provides it for November and December 2013. In this chapter, internet activity data from CDRs is used for the spatiotemporal analysis of the behavior of users using smartphones.

For the data collection and aggregation, the city of Milan is geographically mapped as a 100 by 100 grid of 10,000 rectangular cells as shown in Fig. 3.3. Internet activity level is represented by an imitated rational number for confidentiality. Each number refers to an activity level aggregated for each cell separately for a time interval of 10 minutes. These numbers do not represent actual internet data consumption but refer to activity level in a cell. So they can help for the comparative study of internet activity in

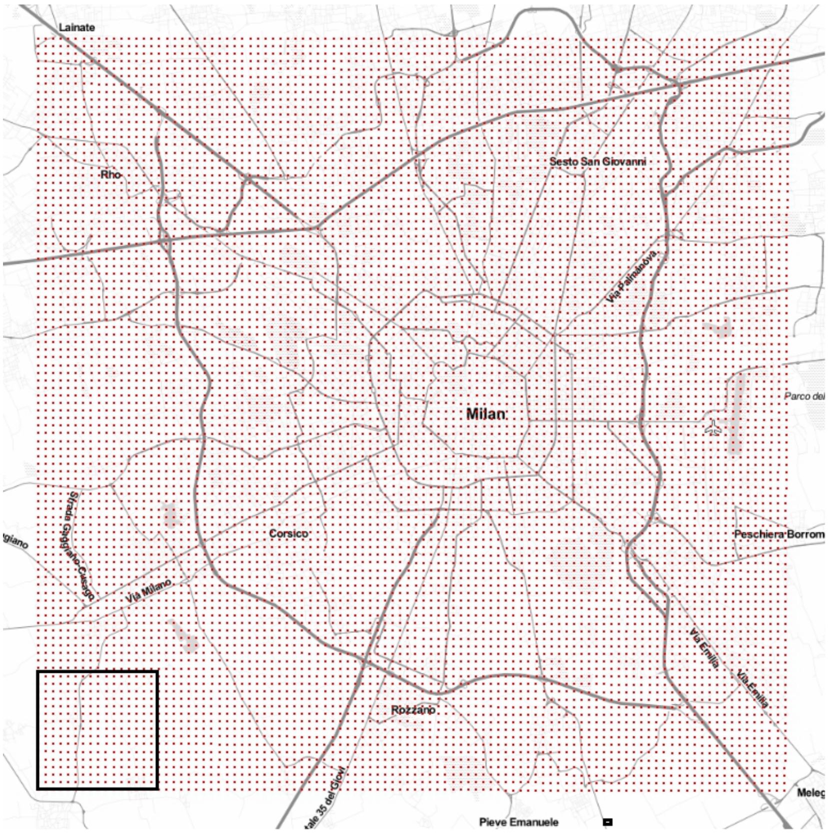

Figure 3.3: Grid over Milan and area under observation

various cells at different time slots. They can give an idea which cells or time slot have more or less active as compared to other cells or time slots and how much the difference is. In this study, data of nine weeks is used, for the months of November and December 2013, and 225 cells covering the left bottom corner of the city as highlighted in Fig. 3.3, a grid of 15 by 15 cells. First three weeks data is used for cross-validation and later more data is added for training purposes. Finally, data for the first eight weeks is used for the training purposes and the ninth week data is used for testing and performance evaluation.

3.2 Methodology

First, three basic levels of internet activity are calculated for each hour of the day for all 225 cells i.e. Minimum, Average and Maximum level of activity. This was followed by the study of the spatiotemporal changes in maximum internet activity level. A density candle plot, Fig. 3.1 and a heat map Fig. 3.2 are plotted for the maximum internet activity level which helps to understand maximum internet activity distribution and variance over each day of a week for 225 cells and over a day across the 225 cells respectively. For visualization purposes here focus has been on maximum internet

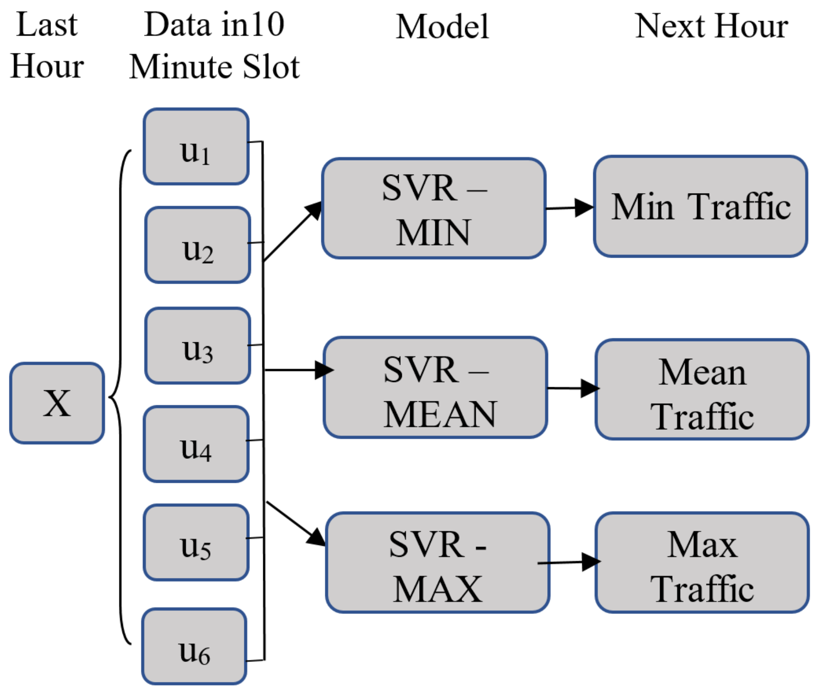

Figure 3.4: Layout of SVR model implementation

activity as it is commonly the most sought-after feature for network resources planning and allocation.

Three SVR models, independent from each other, are trained and tested for the prediction of three activities level Maximum, Mean, and Minimum. Initially, SVR models are separately trained using two weeks data and trained models are validated using the data of the third week which was kept separate. In the validation step, SVR exhibited an accuracy of 85%, 87% and 83% for Minimum, Mean and Maximum tasks respectively. In the end, the model was trained on eight weeks data and tested against unseen data of week nine which was kept separate at the very start of the experiment. The performance of the models was evaluated using the performance metrics defined in section IV.B. The internet activity over an hour, aggregated in six slots of ten minutes each, works as an input to the SVR model for the prediction of each internet activity level for the next hour.

3.3 Proposed Model and Performance Metrics

3.3.1 Support Vector Regression (SVR)

A statistical learning theory based epsilon-insensitive nonlinear SVM (ε-SVM) regression is implemented here. The basic idea of SVR is Structural Risk Minimization (SRM), which uses the sum of empirical risk and model complexity as the minimization objective. In practice, the time series of base stations' traffic show non-linear behavior. Hence, non-linear SVR is used in the proposed base station activity forecasting scheme. To formulate the problem, let's define training data as $x_i, y_i, i = 1, 2, 3, \ldots, n = 24$ where x_i is the input vector representing an hour of the day (e.g. X_1 represents the first hour in the morning 00:00 to 1:00 am) comprising six scalar values u each representing internet activity for a ten-minute time slot as shown in Fig. 3.4. Similarly, y_i represents the maximum, minimum or mean value of the corresponding hour, depending on task the model is trained for. First, the input is mapped on a multidimensional nonlinear feature space using a non-linear transformation function [85] represented as $\Phi(\prod)$. In the high dimensional feature space, regression function can be expressed as follows:

$$f(x) = \prod \omega\Phi(X) + b \tag{3.1}$$

Such that $\omega \in R^d$ and $b \in R$ where d represents the dimensions or number of columns in data as it is six in this case and b represents the bias. And outcome of $\Phi(X)$ represents the input features space. The quality of estimation is measured by the loss function. Here, the ε-insensitive loss function is used, which ignores errors that are within ε distance of the observed values. For training samples outside ε–insensitive zone, the slack variable ζ_i, ζ_i^* is introduced that allows the errors to exist up to ζ_i, ζ_i^* beyond ε-insensitive zone. So SVR model is trained by solving the minimisation problem defined as (1.2):

$$\min \frac{1}{2}||\omega||^2 + C \cdot \sum_{i=1}^{n} (\zeta_i + \zeta_i) \tag{3.2}$$

s.t

$$\begin{cases} y_i - \prod \omega\Phi(x_i) + b \leq \varepsilon + \zeta_i \\ -y_i - \prod \omega\Phi(x_i) + b \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \\ i = 1, 2, 3, ..., n \end{cases}$$

The constant C is a positive numeric value that regularise the function for flatness and over fitting. It imposes a penalty on the values beyond ε-insensitive zone and determine the level of tolerance for deviation of values beyond ε-insensitive zone. An heuristic method is used here for the selection of C and ε. By extensive iterations, using the values of $C = 1$ and $\varepsilon = .02$, the loss function is minimum.

3.3.2 Performance Metrics

In order to evaluate the performance of the proposed SVR model, the following performance metrics were used: Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and Mean Accuracy (MA) [83]. The MAE, RMSE, and MA were calculated for each task, Minimum, Mean and Maximum separately. Let y_i represents the actual hourly minimum, mean and maximum internet activity in test data and $\hat{y}_i$ represents corresponding minimum, mean or maximum hourly internet activity predicted by the relevant model. Hence, performance metrics can be written as (3.3)-(3.6):

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{3.3}$$

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2} \tag{3.4}$$

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n} |y_i - \widehat{y}_i| / y_i \tag{3.5}$$

where n represents the number of instances in the test data. Mean accuracy (MA) is measured using MAPE metric as follow:

$$MA = (1 - MAPE) \times 100 \tag{3.6}$$

where y_i and $\hat{y}_i$ respectively represent actual and estimated values of minimum, mean and maximum hourly internet activity.

3.4 Analysis and Results

3.4.1 Spatiotemporal Analysis of Maximum Internet Activity

The maximum internet activity level is critical for network resource planning and optimization, as it represents peak usage demands in the network. Observing both the heat maps and violin plots reveals distinct temporal and spatial activity patterns, varying significantly by time of day and across network cells.

Spatial Analysis:

The heat map (Fig. 3.2) provides a spatial perspective of maximum internet activity across the cellular grid for a week. Each grid cell represents a unique spatial location in the network, while the color gradient indicates varying activity levels.

High-Activity Zones: Specific clusters in the bottom-left corner (e.g., cells 1–11, 101–111, 201–211) consistently exhibit high activity levels across the week. These clusters appear as yellow to red regions, indicating maximum activity levels of 24–30 or above. The spatial proximity of these cells suggests these areas are critical for resource allocation, possibly due to dense user populations or significant commercial activity.

Moderate-Activity Zones: The blueish-green shades (activity levels 6–18) dominate the grid, representing moderately active cells. These cells are distributed throughout the network and may correspond to residential or mixed-use zones.

Low-Activity Zones: Blue cells with activity levels between 0 and 6 are sparsely distributed, reflecting areas with limited network usage. These regions may correspond to rural areas, underpopulated zones, or areas with limited mobile penetration.

Spatial Cluster Identification: The grid highlights clear clusters of activity, which can inform network planning. For example, consistent high-activity zones may require enhanced capacity through additional small cells or spectrum reallocation.

Temporal Analysis

The violin plots (Fig. 3.1) offer a temporal perspective, showing hourly traffic distribution across days of the week.

Weekday Patterns: Maximum traffic levels rise significantly between 7 a.m. and 9 p.m., with a pronounced peak during evening hours (6 p.m.–9 p.m.). This reflects typical working hours and post-work usage spikes, likely driven by streaming, gaming, and social media activities. Higher variance in traffic levels during these hours indicates the presence of both highly active and moderately active cells.

Weekend Patterns: Traffic remains relatively stable but exhibits a broader distribution during weekends, particularly in the afternoon and evening hours. This aligns with increased leisure and recreational use.

Low-Activity Periods: Across all days, traffic levels dip significantly from 11 p.m. to 6 a.m., reflecting reduced user activity during nighttime. However, certain cells maintain activity, likely reflecting areas with nightlife or essential services.

Key Spatiotemporal Clusters

Clusters identified in both the heat map and violin plots reveal areas requiring differentiated network management strategies.
High-Activity Clusters: Require advanced resource optimization techniques such as traffic offloading, load balancing, and dynamic spectrum allocation. These clusters also highlight potential areas for deploying additional infrastructure like small cells or massive MIMO.
Moderate-Activity Clusters: Serve as transition zones and require periodic monitoring to handle peak surges effectively. Moderate activity zones may benefit from pre-emptive measures, such as capacity boosting during events or holidays.
Low-Activity Clusters: Highlight potential opportunities for energy conservation, e.g., dynamically switching off underutilized base stations during non-peak hours.

Implications for Network Resource Management

Dynamic Allocation: Insights from both the heat map and violin plots emphasize the need for dynamic resource allocation tailored to spatiotemporal traffic patterns. High-activity clusters demand consistent monitoring and pre-emptive load balancing.
Energy Efficiency: Moderate and low-activity clusters present opportunities for energy-efficient measures such as small cell sleep cycles and energy-aware routing.
Proactive Management with SON: These visualizations underscore the value of self-organizing networks (SON) equipped with predictive algorithms for real-time traffic forecasting and proactive optimization.

3.4.2 Models Performance Analysis

The Support Vector Regression (SVR) model demonstrated, as represented in Table 3.1, exceptional performance across all three prediction tasks, namely minimum, mean, and maximum internet activity levels as compared to other state-of-the-art solutions for this problem. When compared to both classical models like ARIMA and Levenberg-Marquardt (LM) algorithm-based neural networks (NN) and deep learning-based models such as CNN-RNN, the SVR model consistently outperformed these alternatives. As highlighted by the authors in [83], deep learning models like recurrent neural networks (RNN), three-dimensional convolutional neural networks (3D CNN), and the combination of CNN and RNN (CNN-RNN) are often considered state-of-the-art for spatiotemporal data analysis. However, the findings in this study challenge that narrative, demonstrating that a well-trained classical SVR model can provide superior results for specific tasks, particularly those with high spatiotemporal granularity.

Table 3.1: Results for the machine learning models against evaluation metrics

Task	Metric	SVR	ARIMA	LM	CNN-RNN [83]
Min.	**MA**	90.25%	67%	61%	69%
	MAE	0.42	22.85	32.86	21.35
	RMSE	0.61	58.18	90.96	50.4
Mean	**MA**	91%	75%	68%	72%
	MAE	0.45	24.34	33.08	26.85
	RMSE	0.65	52.25	81.46	58.36
Max.	**MA**	89.72%	63%	63%	67%
	MAE	0.6	49.78	56.01	44.74
	RMSE	0.88	100.85	126.36	92.32

For accuracy, the SVR model achieved mean accuracy scores of 90.25% for minimum activity prediction, 91% for mean activity prediction, and 89.72% for maximum activity prediction, as shown in Table 3.1. These figures are significantly higher than those of CNN-RNN, which ranged from 67% to 72% for similar tasks, as reported in [83]. This indicates that SVR captures underlying patterns more effectively, offering a robust solution for predicting internet activity levels. In comparison, deep learning models like CNN-RNN, while powerful, achieved predictability of 70% to 80% for all activity levels, with slightly higher accuracy in multitask learning setups [83]. The superior performance of SVR underscores its ability to handle specific nuances in spatiotemporal data that deep learning models might oversimplify or overfit.

In terms of mean absolute error (MAE), the SVR model further demonstrated its reliability, achieving values of 0.42, 0.45, and 0.6 for minimum, mean, and maximum activity predictions, respectively, as detailed in Table 3.1. These results are markedly better than those of CNN-RNN, which exhibited MAE values of 21.35, 26.85, and 44.74 for the same tasks, as reported in [83]. The reduced MAE highlights the precision of the SVR model in generating predictions close to actual values. Similarly, root mean square error (RMSE) values provide additional evidence of SVR's effectiveness. The SVR model achieved RMSE values of 0.61, 0.65, and 0.88 for minimum, mean, and maximum predictions, respectively, compared to higher values for CNN-RNN, such as 50.4 for minimum and 92.32 for maximum predictions [83]. These results demonstrate that SVR minimizes large prediction errors, making it highly reliable for network planning and resource allocation.

The cumulative distribution function (CDF) plots for SVR predictions further validate its alignment with actual data. For minimum activity levels, the CDF in Fig. 3.7 indicates that approximately 60% of cells exhibit activity levels below 5, with fewer than 10% exceeding an activity level of 10. The upper bound aligns closely with the actual maximum of approximately 17. Similarly, for mean activity levels, as depicted in Fig. 3.6, around 50% of cells have activity levels below 5, with only 10%

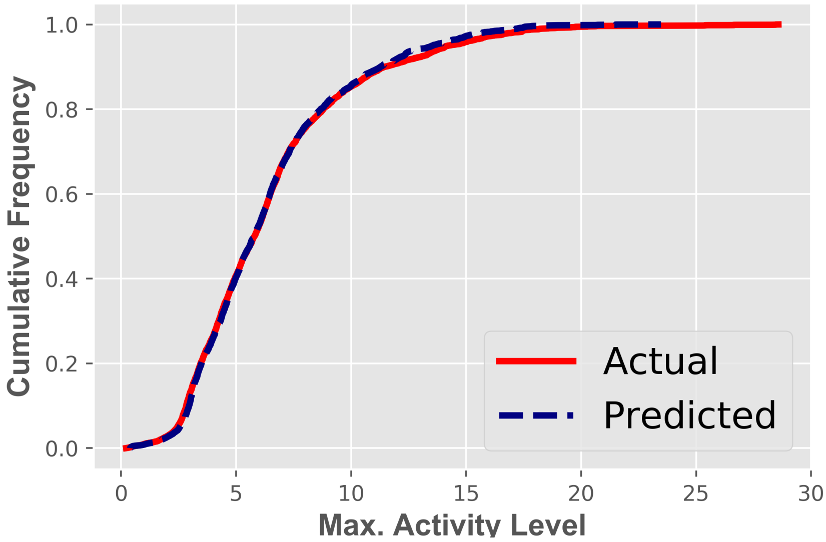

Figure 3.5: CDF of maximum hourly Internet activity: Actual VS Predicted

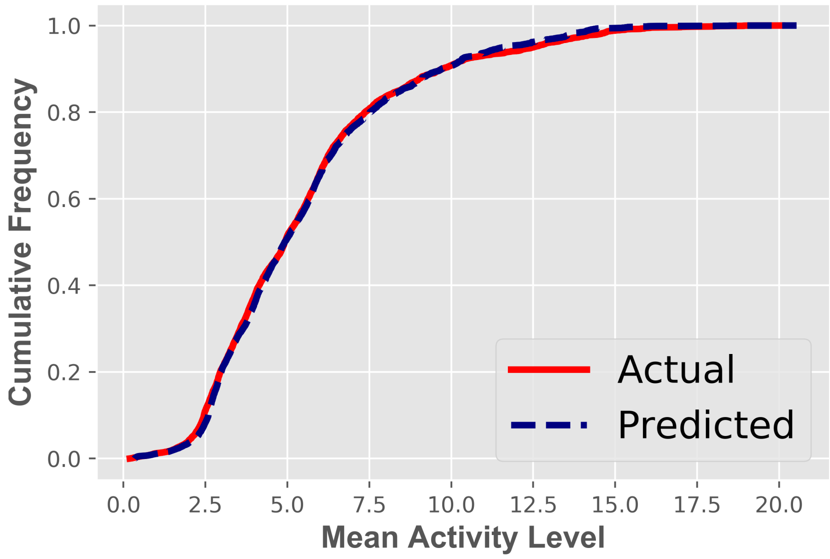

Figure 3.6: CDF of mean hourly Internet activity: Actual VS Predicted

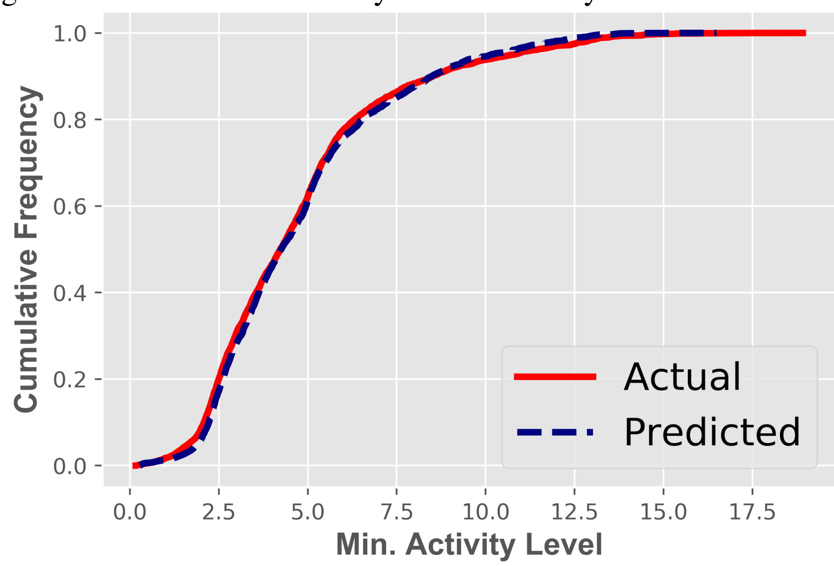

Figure 3.7: CDF of minimum hourly Internet activity: Actual VS Predicted

exceeding 10. The close correspondence between actual and predicted values reflects SVR's ability to accurately capture mean activity patterns. For maximum activity levels, the CDF in Fig. 3.5 reveals that 40% of cells generate activity levels below 5, with nearly 20% exceeding 10 and an upper bound approaching 25. This close alignment between actual and predicted CDFs indicates SVR's superior predictive fidelity, particularly for high activity levels critical for network optimization.

Authors in [83] emphasized the importance of multitask learning in CNN-RNN models, achieving slightly higher accuracy in such configurations. However, the SVR model, trained and tested separately for each task, proved to be more effective in this study. This challenges the assumption that multitask learning is always superior and highlights the adaptability of classical machine learning techniques for targeted tasks. Furthermore, the SVR model's simplicity and computational efficiency make it a viable alternative to deep learning models, especially in scenarios with limited computational resources or where interpretability is crucial.

These findings have significant implications for network management. Accurate predictions of maximum activity levels are critical for resource allocation, congestion management, and ensuring seamless user experiences. Predictions of minimum and mean activity levels provide valuable insights for long-term planning, such as optimizing infrastructure and implementing energy-saving strategies. Additionally, the SVR model's scalability and minimal computational overhead make it suitable for larger networks and more complex scenarios, including edge computing and real-time decision-making.

In conclusion, the SVR model emerges as a robust and efficient tool for predictive modeling in 5G and beyond networks. By outperforming deep learning and classical time-series models across various metrics, as presented in Table 3.1, and by demonstrating close alignment with actual data in CDF plots (Figures 3.5, 3.6, and 3.7), SVR bridges the gap between classical and modern machine learning approaches. It offers a reliable and computationally efficient solution for spatiotemporal analysis and network resource optimization. Its consistent performance across different activity levels underscores its potential to meet the growing demands of modern cellular networks while maintaining accuracy and reliability.

3.5 Conclusion

This chapter has addressed the critical challenge of accurately predicting internet activity levels, a cornerstone for proactive self-organizing networks (SON) in 5G and beyond. By leveraging a nonlinear Support Vector Regression (SVR) model, the research demonstrated that significant spatiotemporal variance exists in the maximum, mean, and minimum activity levels over cellular networks. Using real network data, the SVR model achieved superior predictive accuracy compared to advanced deep learning models and traditional statistical methods, while maintaining computational efficiency and simplicity. The results highlight the capability of classical machine learning models to deliver high-resolution predictions, supporting the design of proactive resource management and optimization strategies. This work underscores the importance of data-driven insights for enabling intelligent, adaptive, and energy-efficient networks, aligning with the broader goal of realizing fully autonomous and efficient 5G and beyond network ecosystems.

Chapter 4

Mobility Pattern Prediction for Wireless Cellular Networks using a Global Indexing System

Data-driven mobility pattern prediction is essential for achieving zero-touch automation and closed-loop performance optimization in 5G and 6G networks. While several studies have proposed methods for indexing user locations and predicting mobility patterns, most rely on localized grid systems, which lack generalizability. This chapter addresses this gap by conducting a comparative study of mobility pattern prediction using a global indexing system, specifically Geohash.

The study reviews promising use cases enabled by mobility pattern prediction and explores the main methods of tracking users' movements and predicting their mobility. A comparative evaluation of regression and classification machine learning approaches is presented, using Geohash as the hierarchical grid system. Additionally, a KNN-based model is developed, which incorporates Geohash ID and user speed as features, achieving an accuracy of 94% at resolution level VIII.

The rapid evolution of mobile networks, including 5G, aims to meet diverse requirements such as 1000x capacity increases, ultra-low latency, and energy efficiency for enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-type communications (mMTC). However, these advancements exacerbate challenges like seamless user mobility management, especially in ultra-dense deployments, where frequent handovers can degrade Quality of Experience (QoE). Accurate mobility prediction is critical to addressing these challenges. This chapter proposes a high-accuracy mobility prediction model based on real-world data while avoiding computationally expensive methods like deep learning.

4.1 Relevant work

Several techniques for modeling and predicting user mobility patterns have been proposed in literature [86,87] and they can broadly be grouped into three categories: 1) Traces based models which are derived from measurements obtained from deployment systems such as connection logs or user location update information; 2) Random synthetic models which employ mathematical models to characterize the movement of mobile devices; and 3) Data-driven machine learning (ML) models which use historical data containing contextual information about the users mobility behaviour. Data-driven ML models would be most relevant in 5G networks because the ultra-dense and heterogeneous nature of network deployment would make it difficult to develop analytical models that can accurately fit the mobility of users in order to analyze system metrics; e.g., HO rate, throughput, coverage, and sojourn time. In addition, ML techniques would play a crucial role, because there are certain hidden patterns that exist in human behaviour which cannot be fully described or modeled mathematically. However, with sufficient data, these hidden patterns or behaviours can be intelligently learned using proper data analytics and ML algorithms.

The focus of this article is to identify potential applications of mobility pattern prediction for 5G networks, and to investigate different data analytics and ML techniques for the diverse applications. First, some futuristic applications, such as content caching and pop-up networking, that utilize mobility pattern prediction to improve the performance of 5G networks are identified and discussed. Second, different methods of locating users are discussed followed by highlighting the significance of a global indexing grids. Third, ML based mobility prediction algorithms and implementation schemes are presented briefly. Lastly, using Geohash grid system to locate users with reference cells, the performances of regression and classification based ML algorithms for the prediction of future locations of the users, a very popular task in mobility prediction, are compared thoroughly. Research on mobility pattern prediction in cellular networks highly rely on synthetic data or uses localized grid system. As presented in a recent study [88], authors exploit synthetic data in a user defined environment to predict the future location of mobile users and they achieve an accuracy of prediction around 95%. To the best of the knowledge of authors, it is the first time that such comprehensive study is performed that compares the performance of machine learning algorithms for the two popular approaches of tracing users, user location coordinates and cells of association with reference to a global indexing system.

Contribution and Organization

The main contributions of the work presented in this chapter are:

- Identification of contemporary and futuristic applications of mobility pattern analysis and prediction in mobile networks.
- Classification of methods of locating users with their pros and cons
- Classification of machine learning algorithms and implementation schemes commonly used for mobility prediction in mobile networks
- Comparative study of the performance of regression and classification based algorithms for the prediction of next multiple consecutive locations, using real GPS data and global grid based indexing system called Geohash.
- Development of prediction model for the prediction of user's location with high accuracy based on Global indexing system. Such that the model uses very simple algorithm and fewer features.

The rest of the chapter is organised as follows: Section 4.2 presents some key applications of machine learning based mobility prediction in the field of 5G and beyond networks. Key ingredients of the machine learning approaches for the prediction of mobility patterns are discussed in Section 4.3. Comparative study based on the results from regression and classification machine learning algorithms applied on coordinates and grid indexed data is presented in Section 4.4. Section 4.5 concludes how Geohash based global referencing system can help to predict users location at different resolution levels of Geohash grid system.

4.2 Enabling Network Optimization via Mobility Prediction

Traditionally, data-driven models for mobility prediction have been used to optimize network performance metrics like latency, call drop rates, and signaling overhead. However, predictive mobility models can unlock several advanced use cases in 5G networks:

4.2.1 Proactive Caching and Pre-buffering

Mobility prediction enables efficient caching and buffering by estimating user trajectories. This allows small cells (SCs) to cache user-specific content or pre-buffer data in areas with predicted connectivity challenges, thereby reducing latency and backhaul overhead [89, 90]. For instance, [89] leveraged user trajectory estimation to optimize SC cache usage, while [90] proposed algorithms to schedule data fetches proactively based on mobility patterns.

4.2.2 Energy-Efficient Cell Switching

Predicting future traffic loads helps optimize energy savings by scheduling small cell sleeping cycles or deciding which base station (BS) a user should connect to along their route [75,91]. For example, [91] used past handover(HO) traces to forecast traffic loads and develop SC sleeping patterns in ultra-dense networks.

4.2.3 Base Station Skipping

Frequent HOs in ultra-dense networks increase signaling overhead and reduce QoS. Mobility prediction aids in HO skipping techniques, such as location-aware or cell-size-aware skipping [92], which minimize unnecessary HOs by extending user connections with selected BSs.

4.2.4 Pop-up Networking

Mobility prediction can optimize drone-assisted temporary networks for scenarios like events or disasters [93]. By forecasting user density, drones equipped with SCs can dynamically position themselves to provide additional resources and coverage.

4.2.5 Secondary User Mobility Predictions in Cognitive Radio Networks

In 5G's IoT-driven ecosystem, cognitive radio networks (CRNs) improve spectral efficiency by enabling secondary users to access available spectrum opportunistically [94]. Mobility prediction can proactively identify spectrum availability and reduce latency for secondary users, as highlighted in [95].

Nonetheless, given the intense ML implementations to make the decision, the proposed approaches are more suitable for stationary users, since all the processes should be repeated while the SUs are moving. Therefore, mobility prediction could be a good candidate in solving this problem by predicting the future locations of the SUs in order to proactively carry out the necessary ML implementations for deciding the most appropriate spectrum portion to sense.

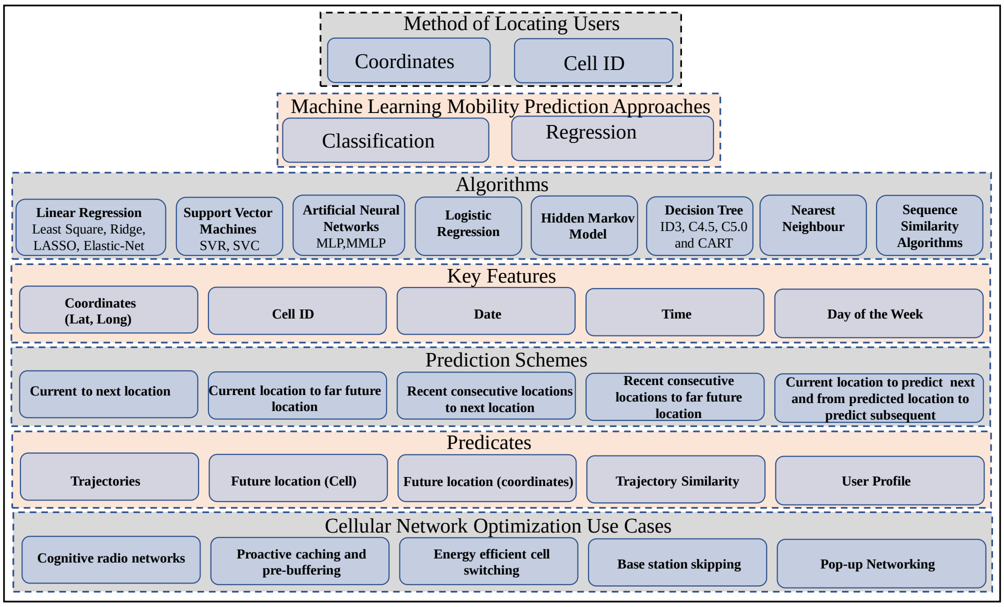

Figure 4.1: Machine learning driven mobility prediction components and cellular network optimization use cases.

4.3 Machine Learning Driven Mobility Prediction

Mobility pattern prediction in wireless mobile networks like 5G is driven by both predefined requirements and untapped opportunities derived from big data generated by networks and users. This data includes user interaction times, geographical locations, and types of services used. By leveraging data-driven ML algorithms, this contextual information can be transformed into insights about user mobility behavior. Key components of mobility pattern prediction for mobile networks are outlined in Fig. 4.1. These components represent a flexible framework that can vary based on specific research or business needs.

4.3.1 Methods of Locating Users

In mobile networks, user locations are determined using two primary methods: coordinate-based and grid-based systems.

Coordinates

Coordinates, comprising latitude, longitude, and altitude, provide exact user geolocation. GPS systems are widely used to obtain this data. Latitude and longitude are commonly used for tracing mobility patterns, while altitude assists in analyzing vertical mobility, such as movements between floors in buildings. However, due to privacy concerns, network operators are often reluctant to share such precise data, limiting its research applications.

Grid

A grid represents a set of adjacent cells covering a geographical area. In mobile networks, each cell can refer to the coverage area of a base station (BS) or a user-defined grid. The grid-based approach, which associates users with specific cells, is popular for studying mobility behavior, particularly in scenarios where cells act as autonomous entities.

Key parameters for defining grid systems include:

- Resolution: Resolution basically defines the size or area of each cell in a grid. In mobile networks, cell resolution is important because different cell sizes can assist in the study of mobility patterns that are important at different network levels. For example at high resolution (grid of cells with small size) can be useful for studying mobility patterns important for operations related to micro and pico cells whereas the low resolution analysis can be useful for macro cell operations. In an actual mobile network, BS cells are commonly categorized as macro (radius 3-30 km) and SCs e.g. micro (radius 500-2,500 meters), pico (radius 100-250 meters) and femto (10-50 meters).

- Shape: In mobile network, different shapes of cells can be used to draw a grid for the study of mobility behaviour of users. It can be hexagonal, rectangular, diamond or triangular shape depending on the requirements e.g. triangular grid can be used to study BS antennas beamforming operations. Most commonly used grids are rectangular in shape as coordinates can easily be translated into a Cartesian System which makes different computational functions very easy. Cells in rectangular grid shape can be of fixed size or different sizes depending on the scheme used to define it.

- Hierarchical Structure: Another criterion to differentiate various grid systems is whether they are hierarchical or non-hierarchical. Hierarchical grid system combines multiple grids, each grid with different cell size. At the first level of the hierarchy, the grid has broad cells and at subsequent levels cell sizes decreases and smaller cells fall within the broad cells at a level above. One such hierarchical grid system is Geohash. It is described in section 4.3.1 and its sample cells at resolution level V to VIII are shown in Fig. 4.2. If the grid is non-hierarchical, each cell has a unique identifier, independent of the labels of

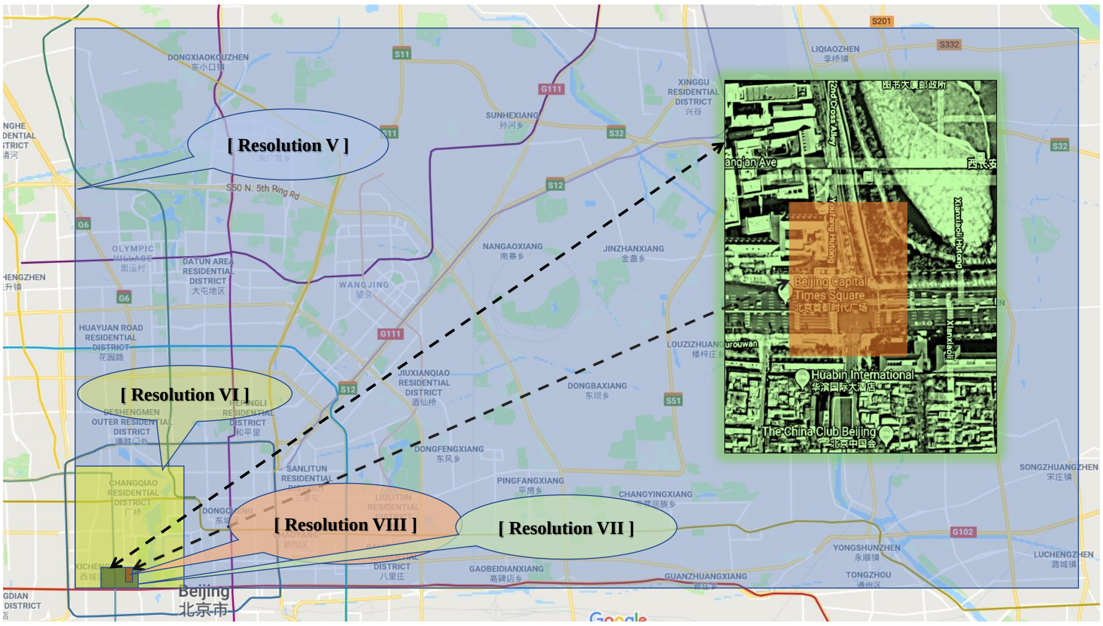

Figure 4.2: Geohash cells at resolution levels V to VIII.

the cells on the top levels. Concept of a hierarchical grid is important to replicate the concept of SCs deployed inside macro cells in real mobile networks.

Global Grid Systems

Global grid systems like Geohash provide a standardized framework for mobility studies, supporting cross-disciplinary applications like urban planning and marketing. Geohash translates user coordinates into alphanumeric tags based on grid cell locations, with longer tags representing finer resolutions. Table 4.1 lists examples of global geo-coding systems, while Table 4.2 summarizes Geohash specifications. These systems can convert the coordinates of a user's location into string labels as cell identifier using some geo-coding scheme and can also do the reverse.

Geohash

Geohash is a global geocoding scheme comprising hierarchical grid system of rectangular cells of different sizes at different resolution levels [96]. It was introduced by Gustavo Niemeyer in 2008 and it is public domain indexing system. It translates the coordinates (latitude and longitude) of a user's location into an alphanumeric string label called 'Geotag' or 'Geohash'. Length of the Geohash tags represents the resolution level of the grid. Longer the Geohash tag, higher the resolution and smaller the cell size are. This resolution also determines the precision, up-to how close a user can be located to its actual position using Geohash tag. Cell dimensions and

Table 4.1: Different global geo-coding systems

Name	Translates	Presents	Resolution	Labels
Mapcode (2001)	latitude, longitude	Points	Arbitrarily precise	Two alphanumeric strings joined by dot of length between 4-7
What3-words (2013)	latitude, longitude	Grid of cells	One, 57 million squares of 3m*3m area	Three words
Geohash (2008)	latitude, longitude, precision	Grid of cells	twelve levels, hierarchical	Alphanumeric string, 32 characters with the base 32 encoding
Plus Code (2014)	latitude, longitude, precision	Grid of cells	Various hierarchical levels up to 3m*3m area	Up to 10 characters with + sign before last two, first four character for area code and last 6 for local code

precision levels up to resolution IX of Geohash grid system are presented in Table 4.2 with reference to the length of the Geohash tag [97]. Fig. 4.2 shows the cells in the Geohash grid system at resolution level V to VIII. In Geohash tag first character presents the broad cell at level one and first two characters together present a smaller cell at resolution II. Similarity between the prefixes of Geohash of two cells represents how close they are to each other, but it is not always true for the vice versa case. For example, Geohash tag 'wx4g00w' contains seven characters which represents a cell at resolution level VII. If a cell has Geohash tag "wx4g00f' it means both cells are falling into same cells up-to resolution level VI.

4.3.2 ML Approaches

Based on these methods of locating users different ML algorithms can be used to predict mobility patterns. These algorithms can be broadly grouped into two categories

Table 4.2: Geohash grid specifications

Geohash length	Cell width (km)	Cell height (km)	Precision (Error km)
1	5,000	5,000	$\pm$ 2500
2	1,250	625	$\pm$ 630
3	156	156	$\pm$ 78
4	39.1	19.5	$\pm$ 20
5	4.89	4.89	$\pm$2.4
6	1.22	0.61	$\pm$0.61
7	0.153	153	$\pm$0.076
8	0.038	19.1	$\pm$0.019
9	0.0048	4.77	$\pm$0.002

namely regression and classification.

Regression

Regression algorithms predict continuous numeric values, such as user coordinates (latitude and longitude). Examples include Least Square, Ridge, LASSO, Elastic-Net, SVR, and Decision Tree Regression (DTR). Multitask regression models can predict multiple outputs simultaneously, e.g., both latitude and longitude. Non-numeric categorical variables, such as dates, must be converted into numeric values for use in regression models.

4.3.3 Predicates

Classification

Classification algorithms, such as SVM, Decision Trees, and KNN, predict discrete outcomes like grid cell IDs or user trajectories. These methods are particularly suitable for tasks like predicting the next cell in a user's path. Sequence similarity algorithms can identify recurring mobility patterns, aiding user profiling and personalized service delivery. In Section 4.4 we compare the performance of both approaches at the different resolution.

4.3.4 Feature Space

Mobile network data is rich in features like user location, time-stamps, and service details. Features can be directly used (e.g., coordinates) or derived (e.g., time-stamps converted into day or time-slot). These features are critical for training ML models to predict mobility patterns.

4.3.5 Prediction Schemes

Using any combination of the above features, different approaches can be adopted for prediction according to the network requirements.

- Predicting immediate or distant future locations using the current location.
- Using sequences of recent locations to forecast trajectories.
- Iteratively predicting far-off destinations by chaining predictions. These schemes vary in computational cost and suitability for specific use cases, such as real-time vs. long-term predictions.

4.3.6 Predicates

Predicates in mobility prediction include predicting user coordinates, grid cell IDs, or paths. Mobility pattern analysis can also be used for user profiling, trajectory similarity analysis, and identifying popular routes. Such insights enable enhanced marketing, resource management, and service delivery in future networks.

4.4 Comparative Study of Machine Learning Approaches for the Prediction of next Location

In this section, we present and compare results for different ML approaches, regression, and classification, for the prediction of future locations, based on the current location, at the different levels of resolution of 'Geohash' spatial indexing system.

4.4.1 Data Set Description

In this study, we have used mobility data of a user in Beijing. Data is extracted from a bigger GPS data-set collected for Geolife project by Microsoft Research Asia [98]. Data is collected from 182 participants using different GPS loggers and GPS-phones at different sampling rates with around 91% of the data logged in dense representation i.e. at every 1-5 seconds or 5-10 meters. The data used here covers mobility patterns of a user in a dense urban area of Beijing for VI trajectories. Data comprises coordinates (latitude and longitude) of location points of the user with time-stamp. These trajectories cover an area of approximately 2250 km^2. It has 27,627 sample points. It crosses 21 unique cells at resolution V of the Geohash grid system and 6,188 unique cells at resolution VIII. For the data selection following aspects have been in consideration:

- Data covers a dense urban area, as study of 5G and beyond networks in the urban context is crucial.
- Data of such participant is used who has more trajectories than any other participant in that area. So we have data of as many trajectories as possible for a single participant.
- For the participant, mentioned above, top VI trajectories with the highest number of foot prints (data samples) are selected. So we have more number of data samples for training models.

4.4.2 Methodology

This study is conducted in two stages. At first stage, simply two features (latitude and longitude) are used to compare the performance of multiple ML algorithms from the two main ML approaches, regression, and classification. We evaluate the impact of resolution (size of cells to identify user location) on predicting the next location of the user under the both approaches. In a classification approach, first the coordinates are translated into Geohash tag at certain resolution then the label (cell id), identifier for current cell, is used as feature and 'cell id' of the next location is used as predicate to train and evaluate classification algorithms namely Support Vector Machine (SVM), Logistic Regression, Decision Trees (DT), K-nearest neighbours (KNN) and Gaussian Naive Bayes.

In Regression approach, we have used linear regression and regression-based implementation of ML algorithm decision tree (DTR). In this approach, we have trained and tested separate models for the prediction of latitude and longitude with each algorithm. In any single model, the latitude of the current location is used as feature and the latitude of the next is used as a predicate, same is repeated for the prediction of longitudes. To use accuracy as an evaluation metric, later the predicted latitude and longitudes are translated into 'cell id' using Geohash. Geohash of the predicted and actual destination points are then compared to calculate the accuracy of regression-based models. Accuracy is used as a single evaluation metric so that the results of regression and classification can be compared. Separate classification and regression models are trained and tested for the resolution level I to VIII of Geohash grid system. For simplification, only better performing algorithms (DTR, LR, DT, KNN and NB, SVM) are presented for meaningful resolution levels V to VIII for the prediction of the next location. In the second stage, we have developed models using best performing classification algorithm and features from the set of features listed in Table 4.3 at resolution VIII. Performance of the models is compared to identify the

Table 4.3: Set of features exploited in enriched data with the help of KNN

Sr#	**Sets of Feature**
1	Set I: {Latitude, Longitude}
2	Set II: {Latitude, Longitude, Time}
3	Set III: {Latitude, Longitude, Time, Date}
4	Set IV: {Latitude, Longitude, Time, Date, Speed}
5	Set V: {Geohash }
6	Set VI: {Geohash, Time }
7	Set VII: {Geohash, Time , Date}
8	Set VIII: {Geohash, Speed}
9	Set IX: {Geohash, Time , Date, Speed}

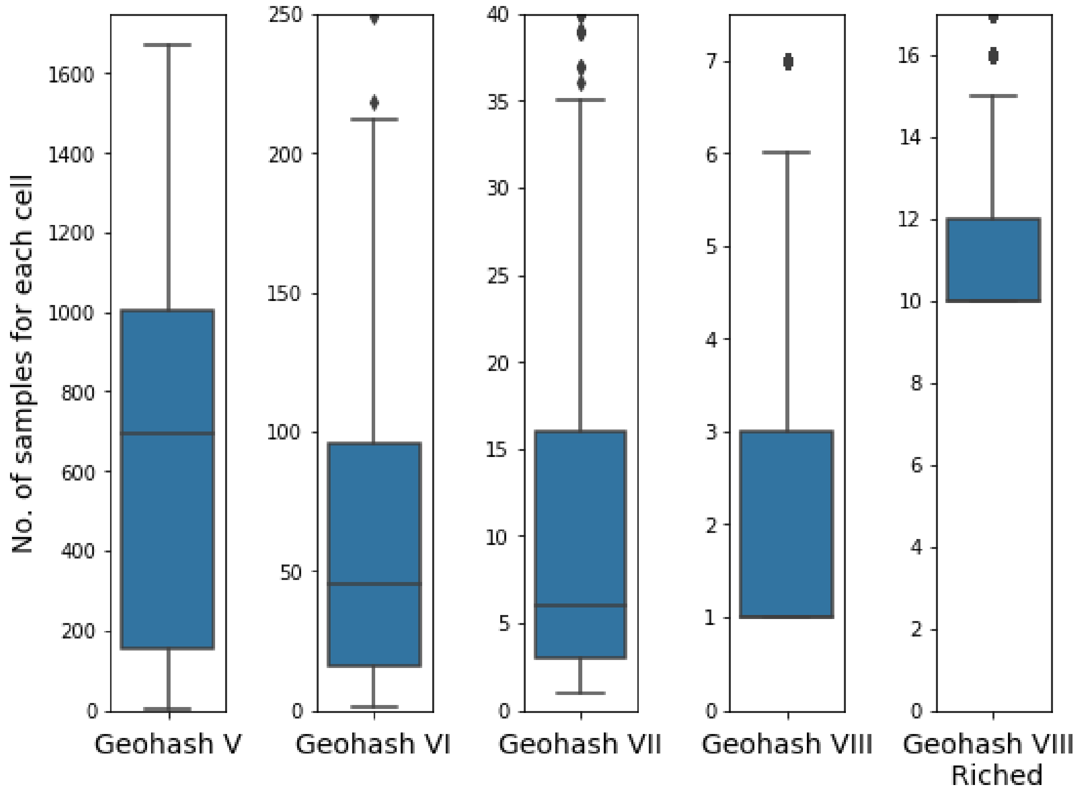

Figure 4.3: Number of samples for each cell at different resolution levels of Geohash

combination of features and tuning parameters that lead to best performing model. But first data is enriched by increasing the number of samples for the cells with fewer sample to over come the challenge of data scarcity. No. of samples for each cell are presented in Fig. 4.3 for each resolution level and enriched data at resolution level VIII.

4.4.3 Data Processing

Since the data used here is real data so it require loads of pre-processing and cleaning before its use in the model. The original data has individual 17,621 trajectory plots of 182 users representing sequence of time-stamped geographical coordinates latitude, longitude and altitude. First through the reverse geo-coding each point in the trajectory is labeled with corresponding city and administrative area. Trajectories in the most dense area Beijing are screened, and in Beijing such participant is selected that has the maximum number of trajectories. For that selected person, top six trajectories with the highest number of for prints are selected. Using location and time information first distance and then speed of travel for the participant is calculated among consecutive points. Using 'pygeohash' library in python Geohash labels for each point are extracted. These extracted information, Geohash ID, speed etc., are used as features in

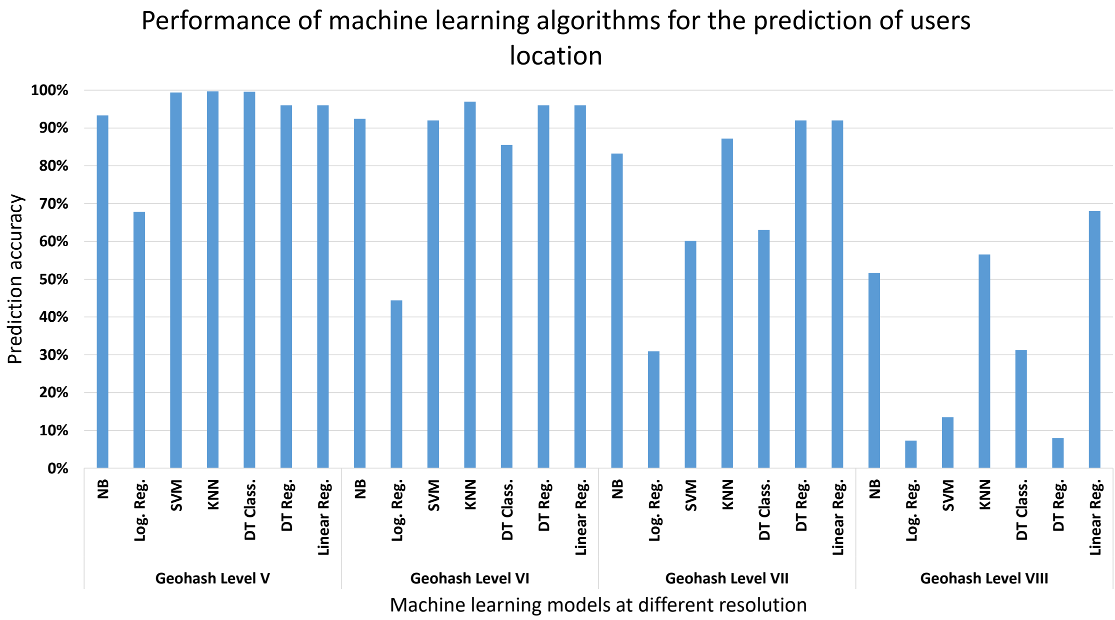

Figure 4.4: Performance comparison of regression and classification approaches at different resolution level of Geohash spatial indexing system for the prediction of next location(step). DT Reg.: Decision Tree Regression, Linear Reg.: Linear Regression, DT Class: Decision Tree Classifier, KNN: K Nearest Neighbours, NB: Naive Bayse, Log. Reg.: logistic Regression

the study later. Categorical features like date, time, Geohash ID are also encode into numeric values so tha they can be used as features in machine learning algorithms. At stage II, data is enriched by creating duplicate samples for the cells where they had very few instances of occurrences. Simple duplication is performed as, at a sampling rate of 1-5 second and 5-10 meter, creating more sample points by interpolation of latitude and longitude will also lead to coordinate points falling in the same Geohash cells.

4.4.4 Results and Discussion

Stage I

Comparison of the performance of regression and classification models at resolution level V to VIII is presented in Fig. 4.4. Resolution levels I to IV are excluded because for those resolutions models can predict multiple future locations with an accuracy almost 100%. It is so because the cell size of the Geohash grid system is very large and in the data used here users' locations is logged after very short intervals of time or distance i.e., at every 1-5 seconds or 5-10 meters. It is highly unlikely for the user to change its cell in that short interval of time and predicting next location becomes

Table 4.4: Results of classification algorithms at different resolutions

Model	Geohash Level	Parameters selected based on cross validation	Mean Val. Accuracy	Test Accuracy
Gaussian Naive Bayes	V	NA	92.82%	93.35%
	VI	NA	92.20%	92.43%
	VII	NA	83.61%	83.23%
	VIII	NA	53.96%	51.64%
Logistic Regression	V	{C: 100, penalty: L1}	67.42%	67.82%
	VI	{C: 100, penalty: L1}	44.05%	44.39%
	VII	{C: 100, penalty: L1}	31.31%	30.91%
	VIII	{C: 100, penalty: L1}	7.84%	7.30%
Support Vector Machine	V	{C: 1000, kernel: linear}	99.07%	99.38%
	VI	{C: 1000, kernel: linear}	90.61%	91.99%
	VII	{C: 1000, kernel: linear}	59.07%	60.16%
	VIII	{C: 1000, kernel: linear}	18.24%	13.46%
KNN	V	{Metric: Minkowski, No. of neighbors: 5, Weights: distance}	99.49%	99.70%
	VI	{Metric: Manhattan, No. of neighbors: 5, Weights: distance}	96.99%	96.96%
	VII	{Metric: Manhattan, No. of neighbors: 5, Weights: distance}	88.06%	87.20%
	VIII	{Metric: Minkowski, No. of neighbors: 10, Weights: distance}	58.75%	56.55%
Decision Tree	V	{Criterion: Gini, Maximum depth: 19}	99.39%	99.59%
	VI	{Criterion: Entropy, Maximum depth: 11}	96.63%	85.51%
	VII	{Criterion: Entropy, Maximum depth: 15}	86.68%	63.04%
	VIII	{Criterion: Entropy, Maximum depth: 13}	55.15%	31.32%

meaningless. Another reason for ignoring these resolutions is that the femto to macro cells are expected to fall within same cells where as mobility prediction is important for knowing the movement of a user from one cell to another. Resolution IX and higher are also excluded from the results here, because it identifies each location almost with a unique identifier which will be same as using coordinates to locate the users.

From Fig. 4.4 it can be seen that all models, whether it be regression or classification models, perform equally well at low resolution as cell size is large. It is observed that accuracy decreases as the resolution increases, i.e, with the decrease in the cell size of the grid system. All models except Logistic Regression, perform well, with an accuracy above 90%, for the prediction of the next immediate location at resolution V.

As it can be seen from Table 4.4, at resolution V where cell size is 4.89Km by 4.89km, SVM, KNN, and Decion tree based models can predict the next location of the participant even with an accuracy more than 99%. at this resolution these classification models perform better than regression models. Reason being the same, the cell size is big so classification models can classify non changing cells easily while regression model predict continuous values and have comparatively low accuracy. For the resolution VI with a cell size 1.22km by 610m, it can be seen that the prediction accuracy is slightly decreased for all models as compared to their accuracy at level V. It can be observed that the best performing regression model Linear Regression and decision tree and classification model like KNN predict next location with almost same accuracy. KNN with 96.96% accuracy overall performs the best and slightly better than the other two models.

At resolution VII with cell size 153*m* by 153*m*, performance degrades further for the all models. KNN classfier, Linear and DT regression remains pretty consistent with slight degradation in performance. Here Linear and DT regression models perform slightly better tahn KNN contrary to their behviur at resolution V and VI. Whereas at Resolution VIII, all models have drastic decrease in accuracy as compared to other resolutions for predicting next location of the participant. Among all these models, here, linear regression performs better than others with an accuracy around 68% for the prediction of next location. From Fig. 4.4 a trend of gradual decrease in performance of all models can e seen from Resolution V to VIII. At low resolution i.e., large cell size, Classification algorithms seem to perform better whereas at high resolution i.e small cell size, regression models perform better. From the comparison of the results, it can be said convincingly that Geohash indexing system can be used to identify and predict mobility patterns of users. It can also be concluded that at low resolution (i.e. in a grid with larger cells) both regression and classification models can achieve similar high accuracy to predict immediate next location, but as the resolution increases and cell size decreases the regression-based prediction approach can provide better results both for immediate and far locations as compared to classification schemes.

Stage II

Looking at the trend from Fig. 4.4 the decrease in the performance from level VII to VIII should have been gradual rather than drastic. The gradual further increase is expected as the cell size decrease further, but this strange behaviour needs to be investigated. Beside that, considering future cellular networks with small cell deployments, study of mobility at high resolution is very important. Therefore the study is extended at high resolution level VIII of Geohash grid system with the best performing classification algorithm. Study is extending with the best performing classification algorithm rather than the best performing regression algorithm. It is so because the natural option for developing a prediction model with a Grid based indexing system is classification. To exploit the Geohash indexing system with its cell id based labeling, it is important to develop better performing classification algorithm.

To understand the performance lag at resolution VIII and to develop a better performing classifier, therefore we further investigated the data, extracted and exploited new features. We extended study with KNN as it was consistently exhibiting the best performance among other classification algorithms. Data exploration lead to insight that key reason for the drastic decrease in performance at resolution VIII is due to data scarcity. As it can be seen from Fig. 4.3 that there are sufficient number of samples for each cell at low resolution V, i.e., around 75% of the cells have 200 or more number of footprints or sample data points at resolution V. But 75% of the cells

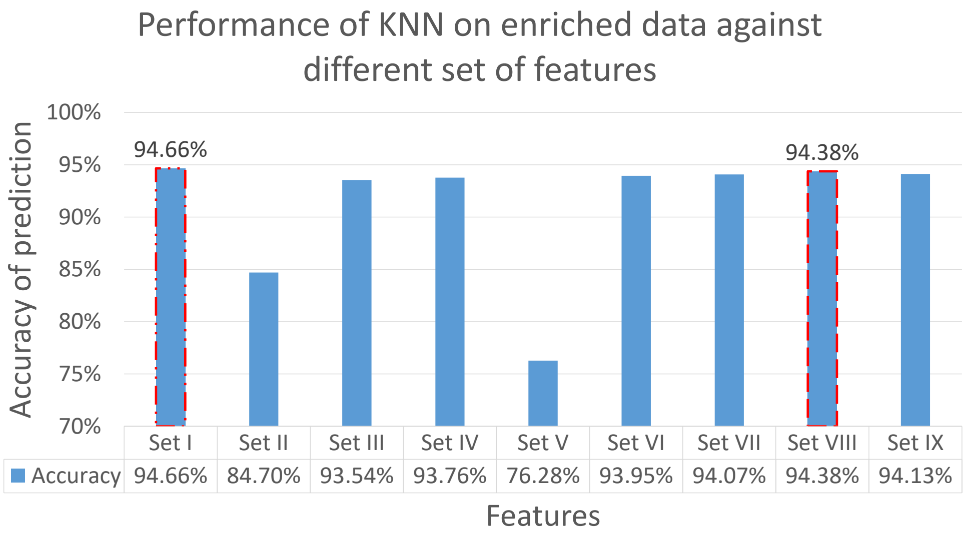

Figure 4.5: Performance of KNN on enriched data for the prediction of next location with the set of features from Table 4.3

at resolution VIII have less than three data points. With the increase in resolution level the number of cells on each resolution level increases whereas the data points in each cell decreases leading to data scarcity in cells at high resolution. It means 27,627 data points fall in only 21, 175, 1100, and 6188 unique cells at resolution V, VI, VII, and VIII respectively To address this challenge more data is generated by duplication of sample points with less than ten instances. As a result of data enrichment the number of total samples increased to 83,131 from 27,627. As it can be seen from the result for feature set I in Fig. 4.5, the performance of the KNN model significantly increases after data enrichment. But goal here to develop a highly accurate model that exploits the Geohash ID. For this purpose new features are extracted and set of different features are exploited for the model development. Set of the features evaluated are listed in Table 4.3 and correlation between the feature set and the predicate Geohash id of the destination location is shown in Fig. 4.6. From Fig. 4.6 it can be seen that Geohash id for the current location and future location has the strongest correlation, Cartesian coordinates (Latitude and Longitude) are next in the list for strong correlation with the predicate 'Geohash VIII Destination'.

From Fig. 4.5 it can be seen that feature set I produces the best results. Whereas use of set II, only Geohash ID does not generate some good results. But when Geohash ID is used along the speed as features in the KNN model it produce very good accuracy of around 94% for the prediction of the next location. So finally we have a model that rely on Geohash ID and can predict the future location of an individual with very high accuracy based on the current location.

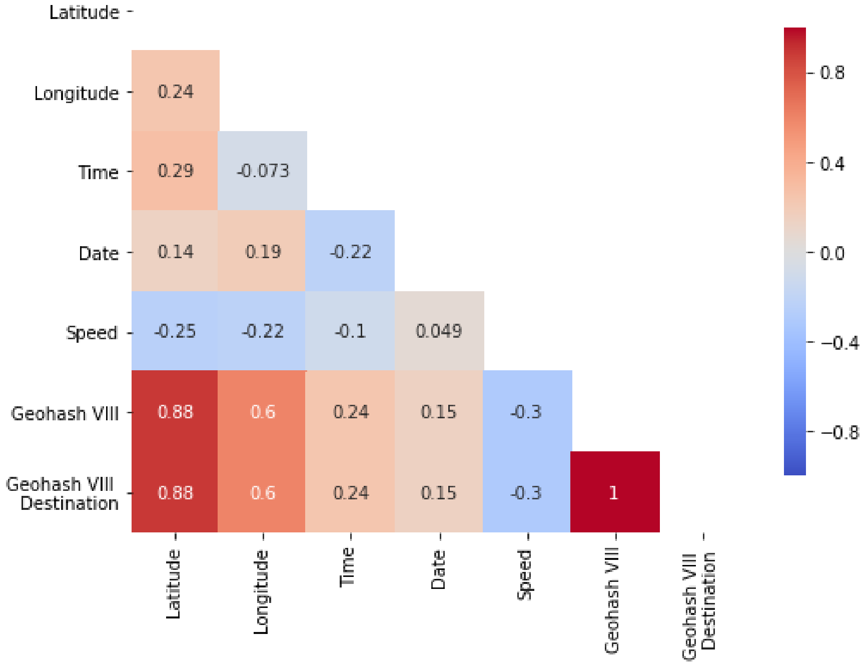

Figure 4.6: Correlation of features and predicate

4.5 Conclusion

This chapter demonstrated that a global hierarchical spatial indexing system like Geohash is a powerful tool for mobility pattern prediction in mobile networks. By leveraging Geohash-based spatial indexing, both regression and classification ML approaches were evaluated, revealing that larger grid cells (e.g., resolution V) allow models like DTR, LR, DT, KNN, and SVM to achieve high prediction accuracy above 96% for the next location of users. However, prediction accuracy declines progressively as the resolution increases (smaller cell sizes), with a noticeable performance drop at resolution VIII, primarily due to data scarcity and smaller cell dimensions.

Data enrichment through duplication and feature augmentation using Geohash Cell ID and speed significantly enhanced model performance. The KNN model achieved an accuracy of over 94% for next-location prediction at resolution VIII, demonstrating the effectiveness of enriched data and feature selection. Overall, the results indicate that classification models generally outperform regression models at lower resolutions (larger cells), while regression models perform better at higher resolutions (smaller cells). The integration of additional features, such as speed, further improves classification accuracy, making it suitable for high-resolution mobility prediction.

These findings align with the thesis's broader goal of enabling intelligent, data-driven mobility management in 5G and beyond networks. By optimizing prediction accuracy at varying resolutions, this work supports enhanced mobility management, network optimization, and service delivery in ultra-dense and heterogeneous networks.

Chapter 5

Performance Based Cells Classification in Cellular Network Using CDR Data

The fifth generation of mobile networks (5G) is now a reality as Verizon turns on the world's first 5G network[1] and EE launches the first 5G trial in Canary Wharf, London[2]. The launch of 5G networks promises great gains in capacity, speed, resilience, and latency, but at the cost of unseen complexity. In order to unlock the 5G potential, a network requires ultra dense small cell deployments, multi-RAT coexistence (radio access technology), catering for an unprecedented spectrum of services, and advanced features (e.g., network slicing), just to name a few. Such complexity requires permanent monitoring, diagnosis, rectification, and actuation for three main reasons. First, it would require an inhibitive capital expenditure to be able to offer the promised 5G infinite capacity perception to all users by following the traditional over-engineering approach. To this end, network agility and flexibility are key to reshuffling the exactly dimensioned resources in ”near” real-time in a user-centric method. Authors in [99] refer to this feature as ”Resource Elasticity” and propose mechanisms for the exploitation of this elasticity by softwarising network functions pertinently and via cross-slice resource provisioning. Second, the increased number of network nodes coupled with the key role of each node in creating the infinite capacity perception renders the failure (or sub-par performance) of any node a significant impediment on the overall network performance.

Thus, in order to offer the best sustainable quality of experience, an augmented smart root cause analysis with ”near” real-time classification of network nodes is paramount. With that in mind, the authors in [100] propose a modified local outlier

[1] https://www.techradar.com/uk/news/verizon-turns-on-the-worlds-first-5g-network
[2] https://www.mirror.co.uk/tech/ee-launches-first-5g-network-13368492

factor approach to identify cells in outage and apply rectification immediately in the context of heterogeneous cloud radio access networks. Third, the amount of data generated by the network to report on its performance and experience of its users is massive and very diverse in nature, content, and frequency. As such, it is essential to synthesize the knowledge from multiple sources in ”near to real-time” with permanent inspection and analysis to match the pace of changes in the network such as capacity demand, users expectations, interference conditions, mobility patterns, etc. Authors in [101] analyze the role of machine learning in the implementation of self-organized network management, with an end-to-end perspective of the network, taking into account the entirety of data generated by the network.

The methodologies employed in current networks for performance management and optimization is by far insufficient to avail real-time optimization and requires a disruptive approach to scale it to 5G timing stipulations. Today's networks are firstly dissected to subsystems (e.g., radio access, transport, core, etc.), then to regions, and features (e.g, 3G, LTE, 4G, etc.) and sometimes to vendors (e.g., Nokia equipment, ZTE node B, Ericsson core, etc.) and more. Such a silo-like approach to examining the network performance is no longer valid in the world of 5G where the distinction between radio access and transport is blurred (e.g. C-RAN: centralized radio access networks with pooled base band units) and the inter-RAT operation is no longer optional (e.g., ATSSS: Access Traffic Steering, Switch and Splitting). Moreover, the abundance of data sources and data types has further increased with features such as minimization drive tests (MDT) [102] and advanced self organization mechanisms. There is an equal rise in specialized solutions to examine these new data sources, alas, these are often vendor specific and result in narrower silos.

In this chapter, we propose a novel semi-supervised machine-learning scheme for automating the detection of under-par network performance and classification of the behaviour of network nodes which allows for a ”near” real time detection of the nodes causing performance degradation. The described method is particularly designed to operate on various and multiple types of data streams that are continuously generated by the network. For instance, the performance indicators continuously generated by the radio access network may be jointly analyzed with minimization drive test data streams for the performance degradation detection. We posit that such a feature is crucial for overcoming the traditional silo-approach of network performance management.

The proposed solution offers a user-centric perception of the network's performance as opposed to the traditional network-centric perception. This aspect is critical in the prioritization of identified problems and in the distribution of resources/efforts for the optimization and maintenance of network nodes. We have tested our solution on CDR data from an African GSM operator and it was successful in identifying under par performing nodes and categorize the “type” of degradation recorded within an hour of the incident.

We thus summarize the contributions of this work in the following points:

- Development of unique method for “near” real-time identification of under-par performing cells that may be applied to any network generated stream of data. Moreover, the identification is further categorized to highlight the “type” of degradation recorded in preparation for the automated smart root cause analysis that would follow.

- A unique method for unsupervised learning based on K-means clustering that auto tunes the pertinent number of clusters based on the streaming data.

- A supervised learning technique based on Support Vector Machines (SVM) that builds on the knowledge extracted from the previous step to classify similar new data streams. Aim of the classifier to enhance accuracy and scalability for real-world applications.

- Validation of the framework on CDR data from a real GSM network, achieving over 98% classification accuracy.

The rest of the chapter is organized as follows. Section 5.1 offers a survey on state of the art work that employs machine learning for solving network problems. Section 5.2 details the proposed method comprising two phases of clustering and classification. Sections 5.3 and 5.4 describe the data set used for the validation of this method and the corresponding results and analysis.

5.1 Machine Learning for Mobile Networks

Machine learning is a proven technique for solving complex problems and has been successfully applied in many fields such as computer vision, medical diagnosis, recommendation systems, speech recognition, and more. Driven by the success of machine learning in various verticals, both industry and the research communities are exploring its potentials in solving mobile network problems, as in [103].

There are three key features in machine learning that advocate for its application on mobile networks. First, machine learning learns from the data and the acquired knowledge improves when the data volume increases. In the advent of fast and massively parallel graphical processing units and the abundance of network-generated data, this key feature of machine learning is well exploited. Second, machine learning and reinforcement learning in particular, circumvents the requirement of highly complex closed form mathematical formulation since it is model-free and relies mostly on the reward system. Closed-form formulations have been a impediment in classical programming for mobile problems as they create a catch-22; many aspects of the system need to be omitted to secure a closed form, hence, compromising the model's fidelity, alternatively, no closed form can be reached which limits the solution to numerical analysis. The third features of machine learning is the knowledge transfer which, in the field of mobile networks, can exploit the temporal and spatial differences and relevance of different regions. As such, the knowledge acquired in one node can be transferred to another (or new) node to accelerate the learning curve. This is particularly important in the 5G era which is characterized by dynamic changes and diversity. In this case, the knowledge acquired in classifying macro-cell performance can readily be transferred to small cells or indoor cells. Similarly, knowledge acquired in diagnosing quality deterioration for VoIP services can be used to accelerate the diagnose of IoT service (e.g., e-health, smart meters, driver-less vehicles etc.).

Applying machine learning to solving network problems has resulted in a prolific research output in the last few years. Many efforts have been invested in identifying potential applications of machine learning such as [1, 104, 105] in the domain of wireless networks and in particular in the implementation of self-optimization mechanisms. Two recent works apply machine learning techniques into the analysis of Call Data Records (CDRs): [106] and [107]. Authors in [106] employ big data analysis for over ten million CDR records to extract the spatial-temporal predictability of network traffic. These CDR-driven predictions are then applied to a novel mechanism for joint optimization of energy consumption and inter-cell-interference in ultra-dense 5G networks. Motivated by the crippling cost of churn faced by telecom operators, authors in [107] apply deep learning to CDR and customer relationship management (CRM) records to predict which customer is likely to churn to allow for user-centric retention efforts. On the other hand, authors in [108] propose a novel method to automatically diagnose the radio condition in a mobile network cell based on the user's performance as captured by MDT records. To this end, the proposed method applies an unsupervised machine learning technique (Self optimizing maps) to cluster and classify the performance of each cell in the network. Moreover, in the domain of fixed networks, authors in [109] propose a machine learning method for performance monitoring employing SVM and double exponential smoothing (DES) for the prediction of equipment failure. Authors in [110] offer a solution that uses deep learning to predict customer churn. By exploiting the intrinsic property of deep neural networks, the proposed solution can be applied to any type of network and any subscription based events.

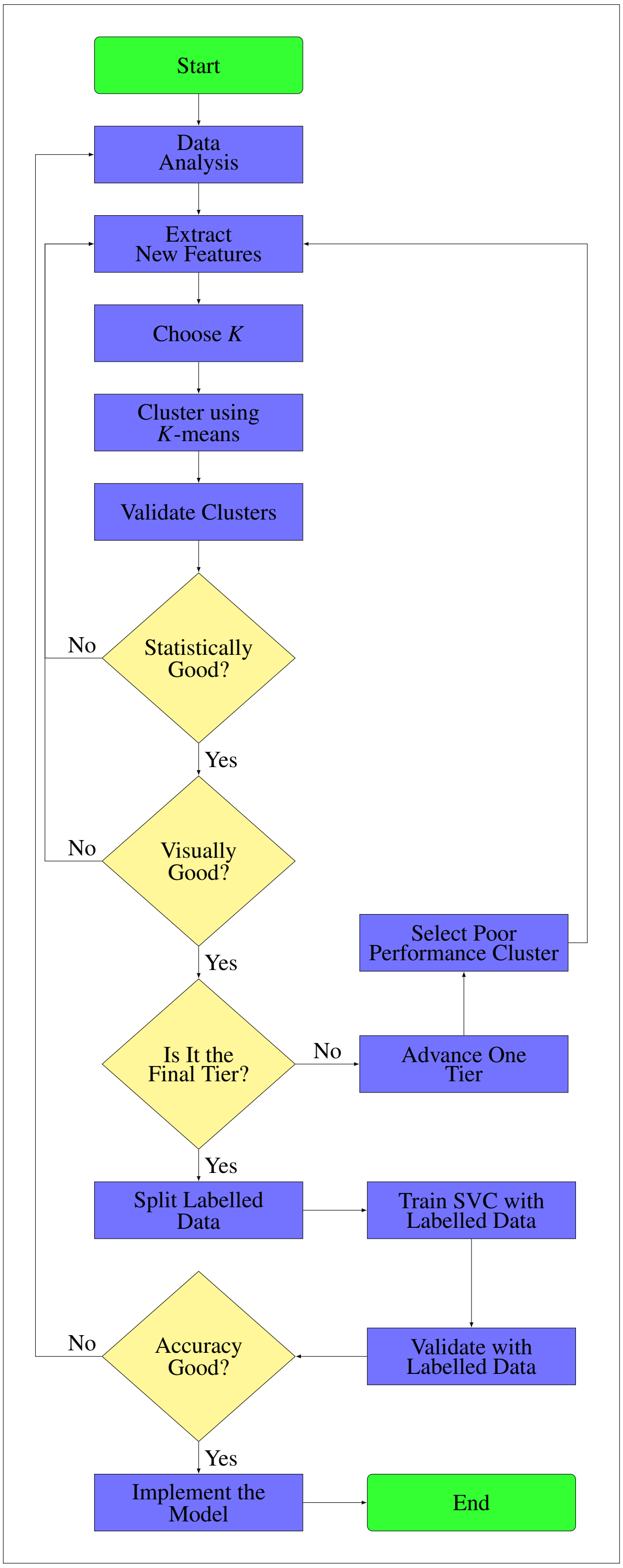

Figure 5.1: Classification process flow chart

5.2 Methodology

In this section, we discuss the proposed solution, first detailing each step of the clustering stage then describing the SVM classifier. Steps taken to devise the solution are outlined in the flowchart presented in Fig. 5.1. The solution is devised with an aim to classify cells according to their performance in certain intervals of time over the day. By performance here we mean the volume of dropped calls (*i.e* duration, quantity) against that of normally terminated calls.

To this end, we start by analyzing the data to find relevant features that are descriptive of the desired performance aspect. Moreover, we apply the concept of sliding window to capture both, the instantaneous streaming values as well as the trend of variation of these values; the length of segments can be user defined, e.g., three hours. As there exists no gold standard for cluster validation in our case, therefore, in the clustering phase, the goodness of clusters is preliminarily determined by the internal measure of clusters compactness and separation. More importantly, we mostly rely on domain expertise via visual validation to determine the quality of the clusters and to ensure that domain specific characteristics are represented correctly. Tier-wise implementation scheme using K-means, as explined in in section 5.2.3, is designed. A certain combination of features is selected for each tier after evaluating possible combinations from the available feature space.

Once the clustering is yielding satisfactory results data is labeled with the help of that clustering scheme. Then the labeled data is used to devise a classifier with supervised learning approach as it can be more simple and generic in terms of implementation on real data while being more deterministic in terms of performance evaluation at the same time. In the development of a classifier, the labeled data is split into training and testing data set. The first is employed to train the SVM classifier which is later evaluated on the testing data set. The entire process of clustering and classification is reiterated until a satisfactory level of accuracy is achieved resulting in a final clustering and classification scheme.

5.2.1 Clustering Model

The goal here is to identify categories of issues present in network and group together segments with similar performance behavior. As, here, different possible categories of performance behavior are not known, so, the use of a clustering algorithm is an obvious choice. But the fact is that there is no best clustering algorithm [111]. Some of the major methods of clustering are based on density estimation, probabilistic estimation, partition, and graph-theoretic.

There are numerous clustering algorithms which basically differ in their objective function computation. Each algorithm has its own pros, cons and implementation methods. Density-based clustering algorithms [112] e.g., DBSCAN that looks for regions of high density is not very efficient particularly for high dimensional sparse data cases because it compares the distance of all pairs of points.
Whereas, though the spectral clustering [113], an example of graph theory, does not require complicated parameters like BDSCAN, but it is not a good algorithm for big data and a higher number of clusters. A Gaussian mixture model [114] from the family of probabilistic methods consider data as a mixture of Gaussian distribution with unknown factors. This model has limitations like it cannot be scaled and it requires too many parameters for implementation.
K-means clustering is the most popular and simplest partition model [115]. It does not require any parameters except the number of clusters. There also exist the methods to compute an optimal number of clusters. It works well even for high-dimensional sparse data [111]. Multiple variants and implementation schemes of K-means exist due to the rich history of research [111]. In addition, it is scale-able for huge data and very efficient for real-time implementation. Considering all these factors K-means clustering is selected here.

5.2.2 Clustering Validation Metrics

Another critical task in clustering is the validation of results, particularly in absence of a ground truth, as it is the case here. We have no predefined labels for any classes in our data. There exist multiple metrics for clustering validation, for data, where exists ground truth or golden standard, also known as external metrics [116]. But the metrics for evaluating the goodness of clustering without ground truth are rare. Such metrics are called internal metrics and they evaluate clusters generally on the basis of two attributes of clustering: compactness within clusters and separation between clusters [117]. Some of the internal metrics take compactness into consideration or separations only, other evaluate clustering taking both parameters into consideration at the same time. We have taken four commonly used and well-acknowledged internal metrics [117]: Root-mean-square standard deviation (RMSSTD) for compactness, R-squared (RS) for separation, Calinski-Harabasz index (CH) and Silhouette index (S) for compactness and separation.
Internal metrics basically help to determine the optimal number of clusters and their numeric values just indicate how compact or separate the clusters are on average. For example, Silhouette index varies between [-1,1] where negative values indicate that clusters are mixed with data points assigned to one cluster from other clusters. When it is zero or close to zero, it reflects clusters are not far from each other and similarity

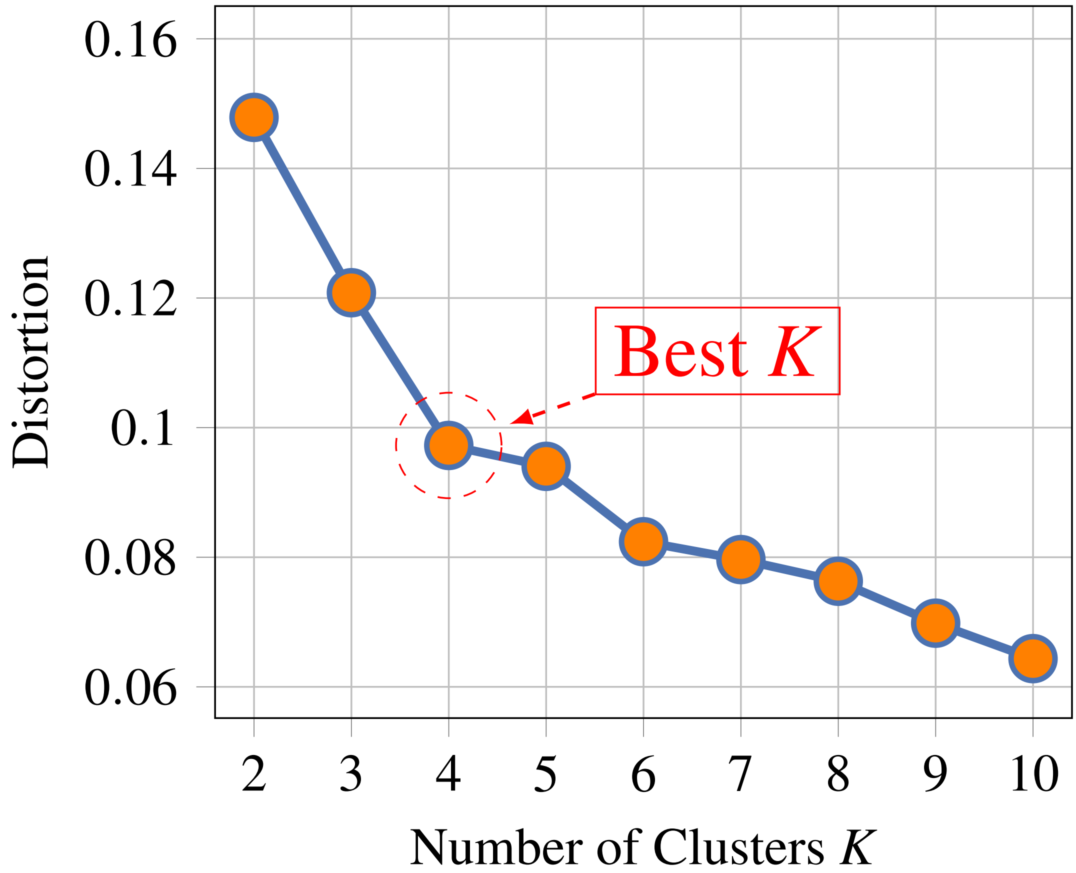

Figure 5.2: Selection of optimal K, at Tier I using elbow methods.

among data points within clusters is low. Moving away from zero on the positive side of the index indicates that the clusters are well separated and well compact. Besides that, visual validation with the help of domain knowledge is a requirement in the absence of a gold standard. These metrics can be a good statistical indicator of compactness or separation of the cluster but they are no guarantee that the obtained clusters are suitable for the desired application [117].

5.2.3 Clustering Implementation Scheme

K-means clustering results in groups of data such that a distance metric between the empirical mean of a sub-cluster and the points in that cluster is minimized and it happens for all the clusters [111]. In this work, we have used the Euclidean distance as the distance metric. The input parameter required by K-means is the number of clusters beside decision function also known as kernel.

In this research, an heuristic approach is applied for the selection of features. For this different combinations of features are used in K-means and the features with optimal results are shortlisted. Clusters here not only need to be statistically

Table 5.1: Performance of K-means against clsuter validation metrics for different values of K on both tiers of clustering

K	RMSS		RS		Silhoutte		CH	
	T-1	T-2	T-1	T-2	T-1	T-2	T-1	T-2
2	0.14	0.18	0.44	0.33	0.48	0.36	9546	312
3	0.11	0.15	0.69	0.54	0.54	0.39	13479	368
4	0.09	0.14	0.78	0.58	0.45	0.37	13963	296
5	0.08	0.14	0.81	0.62	0.46	0.38	13206	262
6	0.07	0.13	0.85	0.66	0.41	0.36	13383	246
7	0.07	0.13	0.87	0.69	0.41	0.31	13126	232
8	0.06	0.12	0.88	0.71	0.42	0.31	13278	223
9	0.06	0.12	0.90	0.73	0.41	0.32	13910	217
10	0.06	0.11	0.91	0.76	0.39	0.29	14052	220

sound but require to group together segments with similar performance in terms of telecommunications characteristics, that is variations of cell load alongside variations in the number of interrupted calls. Therefore cluster validation is done by applying domain knowledge on the visual presentation of segments in clusters alongside the use of internal validation metrics. Moreover, the selection of clustering scheme is another key factor that affects the results. The data has multiple features and co-relations among those features and variation within the values of individual features determine the network behavior. We have applied a two-tier clustering scheme, wherein the first tier separates the segments with poor performance from those with good performance while the second tier further segregates the poor performing segments into the final clusters reflecting different type of network behaviors. Another advantage of using this two-tier approach is that visual validation is easier as there are fewer segments to analyze in the second tier.

In both tiers, we majorly rely on the elbow method beside the metrics of compactness and separation for selecting the optimal K. Bend on the elbow plot as shown in Fig. 5.2 is the optimal number as after that increase in number of clusters does not changes the clustering at that rate as it change it before that, similar behavior is observed when RMSS and RS values in Table 5.1 are plotted against the number of clusters. Calinski-Harabasz index (CH) and Silhouette index (S) are used to evaluate the overall quality of the clusters.

The inputs of each tier are the hourly aggregated CDRs, split into l long vectors for each segment i and cell j. $\vec{D}_{i,j}$ and $\vec{D}'_{i,j}$ are the duration of normally terminated and dropped calls, while $\vec{C}_{i,j}$ and $\vec{C}'_{i,j}$ express the quantity of normally terminated and dropped calls, respectively. Moreover, the vectors $\vec{D}_j$, $\vec{D}'_j$, $\vec{C}_j$ and $\vec{C}'_j$, with length 24, contain data relative to cell j for the entire day.

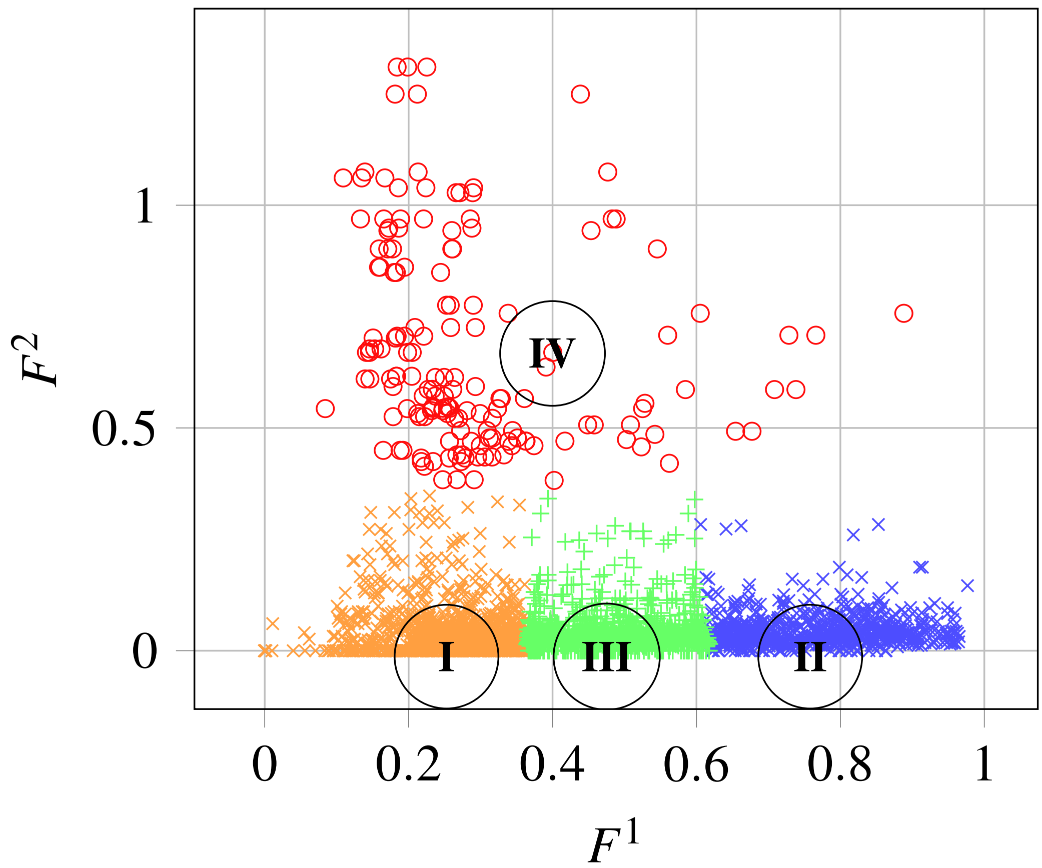

Figure 5.3: Four clusters generated at Tier I

First Tier

Tier I is used to separate segments with poor performance from the segments with good performance in terms of number and duration of interrupted calls. It also segregates segments with good performance according to the cell load. For that we have used two features F^1 and F^2 computed as follows:

$$F^1_{i,j} = \text{mean}\left(\frac{\vec{D}_{i,j}}{\max\left(\vec{D}_j\right)}\right) \tag{5.1}$$

and

$$F^2_{i,j} = 10 \cdot \log\left(\max\left(\vec{C}'_{i,j} \div \left(\vec{C}_{i,j} + \vec{C}'_{i,j}\right)\right)\right), \tag{5.2}$$

where $\div$ indicates an element wise division and log denotes the natural logarithm. To emphasize differences in the number of dropped calls, we have considered a logarithmic scale for F^2. The idea is to make the variation of the bad data more pronounced, such that clustering can find the bad quality segments easily based on Euclidean distance. The output of the first tier is depicted in Fig. 5.3. Note that the bad quality segments are grouped into Cluster IV while the other three clusters contain good quality segments segregated according to their volume of traffic.

Second Tier

In the second tier, we consider only the segments from Cluster IV, the one with poor performing segments from the first tier. It is further sub-clustered into three groups of segments with K-means clustering. There are two sets of features used in the second tier. The first set $F_{i,j}^3$ is obtained via

$$F_{i,j}^3 = \nabla\left(\frac{\vec{C}_{i,j}}{\max\left(\vec{C}_j\right)}\right), \tag{5.3}$$

where ∇ indicates the gradient operation and thus $F_{i,j}^3$ is a vector of $l-1$ features.

The second set $F_{i,j}^4$ is obtained using

$$F_{i,j}^4 = \nabla\left(\vec{D}'_{i,j} \div \left(\vec{D}_{i,j} + \vec{D}'_{i,j}\right)\right), \tag{5.4}$$

and thus also results in a vector of $l-1$ features. Therefore the second tier contains $2l-2$ features in total. $F_{i,j}^3$ captures the variation in the number of bad calls as a gradient whereas $F_{i,j}^4$ encompass variations in bad call duration.

5.2.4 Classification

The labeled data from K-Means clustering is used further to train and test a classifier. Here an SVM based classifier (SVC) is used to classify the labeled data. SVM is very effective for high dimensional sparse data. It is memory efficient for training as it uses a subset of points for the training of the decision function. SVM based models already have proved to perform exceptionally on cellular network data like the call traffic prediction at high granularity [106]. Another huge advantage of SVC is the availability of a diverse range of Kernels to compute the decision function [118], such as linear, polynomial of higher degrees, Gaussian, sigmoid etc. In this research, we have evaluated the three most popular kernels, linear, polynomial (cubic) and radial basis function (RBF) for developing the classifier. SVC decision function constructs hyper-plane(s) in high dimensional space which separates the data points into different classes.

Theoretically, good separation is when the hyper-plane has maximum distance from the nearest training data points of any class. But there is a trade-off between the separation level also known as functional margin and model generalization. The

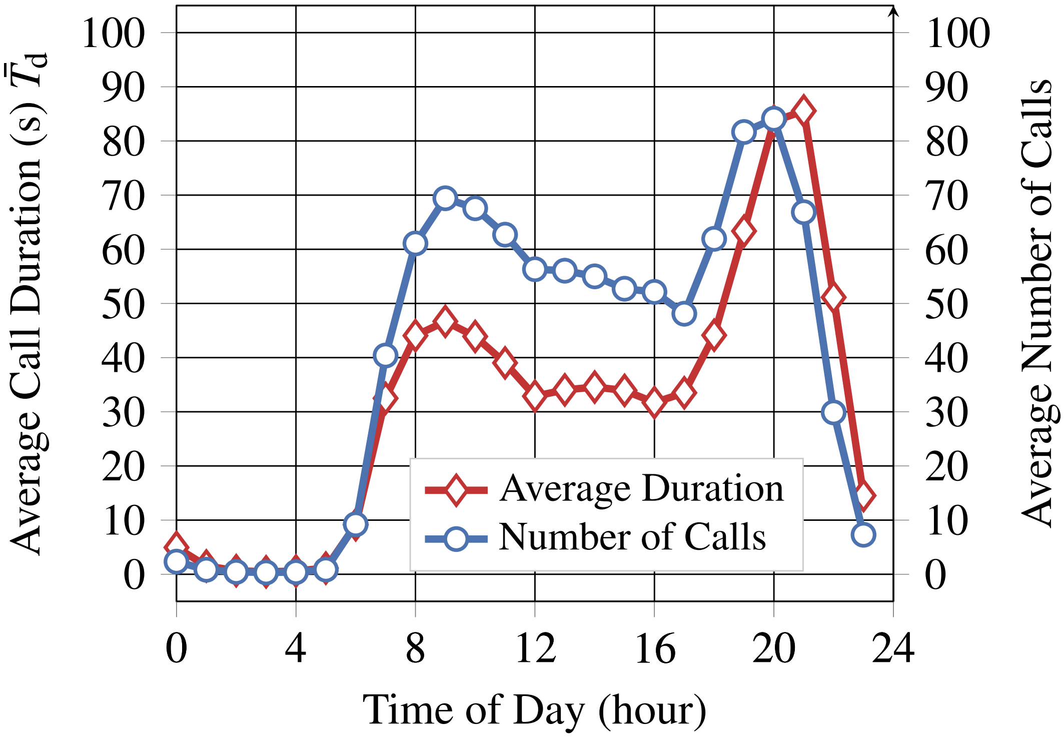

Figure 5.4: Average normalized traffic

higher the margin, the lower the generalization error of the classifier leading to over-fitting. Model generalization and over-fitting can be regulated with the help of kernel parameters like C also known as penalty parameter for the error. A larger C yields a more accurate classification at the cost of a lower generalization. A smaller C means a more smooth decision plane. We have used a set of exponentially increasing values for C for all three kernels. For model training and evaluation data is randomly split into training and testing data sets, such that 75% is allocated for training and validation while the remainder 25% of the data is reserved for testing. A classifier is trained and validated, with three-fold cross-validation, using 75% of the training data with different combinations of kernels and C. The best estimated model is then applied on the test data in the end for the final evaluation of classifier.

5.3 Data-set Description

The data used in this research is extracted from CDRs gathered from a real GSM network in Africa over a period of three weeks for 759 geographical cells. Data used here is numeric in nature and presents duration in seconds and number of calls terminated normally and dropped, both separately. Normal termination means the calls were properly cut off by either user, conversely, a dropped call is a call that has ended due to some network error. We have used hourly aggregated data of 759 cells for one day to train and evaluate our clustering and classification scheme.

From our preliminary analysis, we found that total call traffic in terms of duration and quantity is very low from midnight until morning as shown in the Fig. 5.4, which contains the average[3] total traffic across all the cells for one day. Thus, we have taken the data from 6 a.m. to midnight into consideration for this research. Selecting the data this way also helps to overcome the model biases towards greater number of segments with low traffic. Moreover, network operators are more concerned in dealing with issues which affect larger volumes of calls, which occur during the day.

5.4 Results

The results of the proposed clustering and classification methodology are presented in this section. The analysis demonstrates the framework's effectiveness in categorizing network cells based on performance and identifying specific patterns of degradation using CDR data. Detailed insights into the clustering and classification processes, along with their implications for network optimization, are discussed below.

5.4.1 Clustering

Tier I

From the results of elbow method as shown in Fig. 5.2 and those of RMSS and RS shown in Table 5.1 it is found that $K = 4$ is the optimal number of cluster at Tier I. Using $K = 4$ for the feature sets F^1, F^2 produces optimal clusters presented in Fig. 5.3. The use of F^2, which incorporates a logarithmic scale for dropped calls, was particularly effective in capturing smaller variations in dropped call volumes. This ensured a meaningful separation of segments, where even minor degradations were distinctly identified. As shown in Fig. 5.3, the four resulting clusters represent different performance profiles:

- Cluster I: Segments with good quality and low traffic, indicative of low activity periods or less congested areas.
- Cluster II: Segments with medium traffic, reflecting moderate activity levels in the network.
- Cluster III: Segments with high traffic, representing peak usage times or highly congested areas.

[3]The traffic is first scaled between 0 and 100 for the whole day and each cell, then averaged. This provides a realistic information of the cell load at each hour of the day.

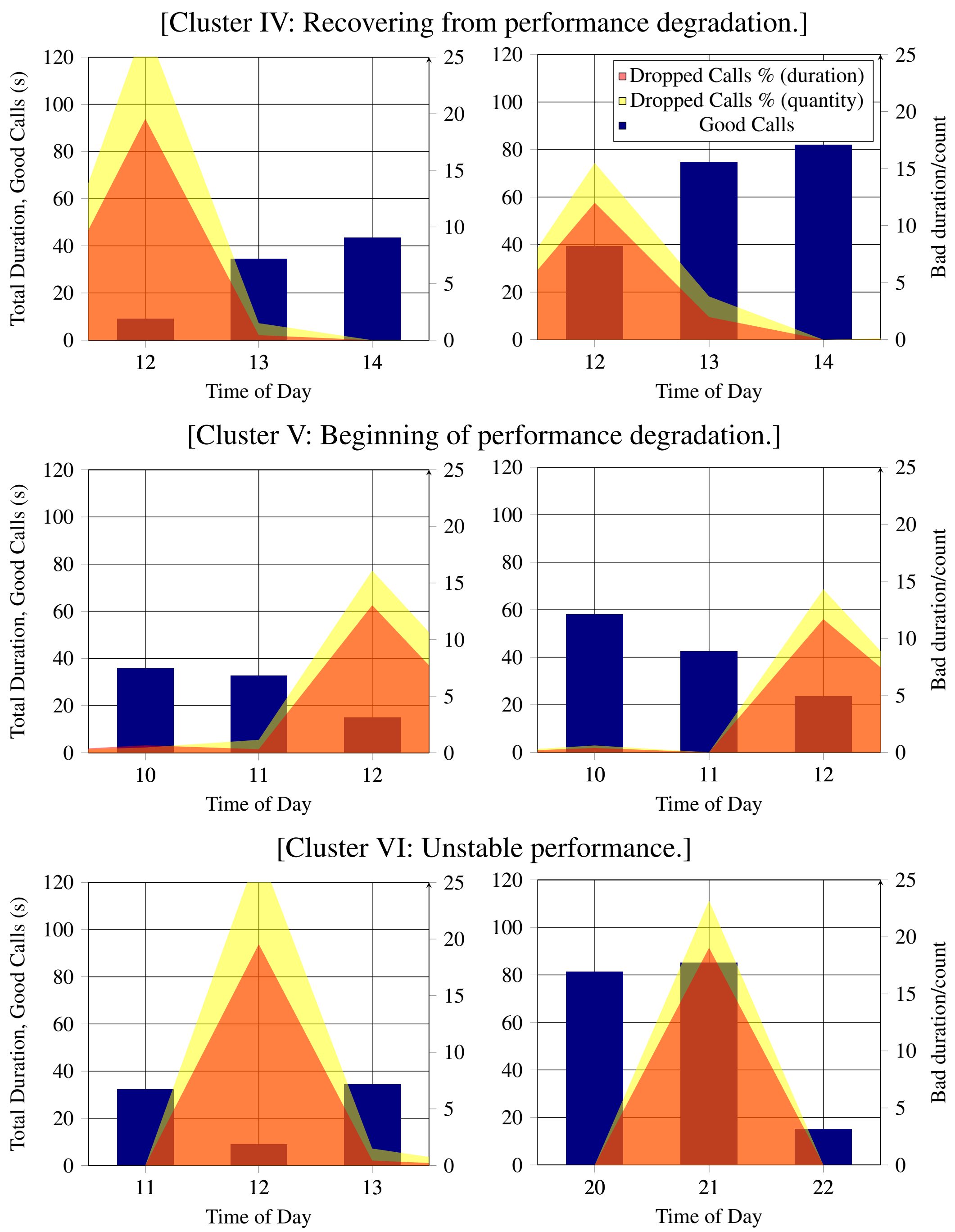

Figure 5.5: Sample segments in their respective clusters.

- Cluster IV: Segments with poor quality, characterized by high volumes of dropped calls and reduced performance.

The ability to isolate Cluster IV with poor performance is a significant achievement, as it enables targeted interventions in problem areas without impacting other network operations. Furthermore, this tier's clustering aligns with the real-world requirements of prioritizing segments with substantial performance issues for further analysis.

Tier II

The second tier of clustering focuses on Cluster IV from Tier I, which contains segments with poor performance. By further analyzing these segments, Tier II aims to categorize specific types of degradation. Using feature sets F^3 and F^4, which capture variations in call duration and dropped call behavior, the optimal number of clusters at this tier was determined to be $K = 3$.

The statistical metrics in Table 5.1 confirm the compactness and separation of clusters at Tier II. These numbers indicate how compact or separated the clusters are at Tier II. Where RS and RMSSD scores respectively give an idea about how separate or compact the clusters are there they also help to choose an optimal number of clusters which is III for Tier II.

Visual validation further supports the quality of these clusters, as shown in Fig. 5.5 , presenting samples from each cluster at Tier II. The three sub-clusters represent distinct degradation patterns:

- Cluster IV: Represents segments recovering from performance degradation, as evidenced by decreasing dropped calls and increasing good call durations.

- Cluster V: Captures segments at the onset of performance degradation(i.e., as soon as the performance starts degrading), characterized by an increase in dropped calls and a decline in good call durations.

- Cluster VI: Identifies segments with unstable performance, where a spike in dropped calls is observed at certain times, disrupting normal traffic patterns.

Red color represents the duration percentage of dropped calls while the yellow color represents the percentage of dropped calls (quantity), both presented as secondary axis on the right y-axis, whereas normalized good calls duration is presented by blue bars on the primary axis on the left y-axis. The top two graphs in Fig. 5.5 represent two sample segments from Cluster IV, it can be seen that segments with decreasing dropped calls and increasing good call duration are grouped in that cluster.

Table 5.2: Classification cross validation results

Rank	C	Kernel	Mean Validation Score	Mean Train Score
1	10^3	Linear	0.989	0.99
2	10^4	Linear	0.986	1.00
3	10^2	Linear	0.986	0.99
4	10^2	RBF	0.985	0.99
5	10^3	RBF	0.984	0.99
6	10^4	RBF	0.984	1.00
7	10^1	Linear	0.983	0.99
8	10^4	Cubic	0.975	0.98
9	10^1	RBF	0.974	0.98
10	10^0	Linear	0.970	0.97
11	10^3	Cubic	0.961	0.97
12	10^2	Cubic	0.931	0.93
13	10^0	RBF	0.923	0.92
14	10^{-1}	Linear	0.914	0.91
15	10^1	Cubic	0.877	0.88
16	10^{-1}	RBF	0.862	0.86
17	10^0	Cubic	0.797	0.80
18	10^{-1}	Cubic	0.581	0.58

Whereas segments with an opposite behaviour, to that observed in Cluster IV, are grouped in Cluster V shown with two graphs in the middle in Fig. 5.5. The bottom two graphs in Fig. 5.5 show segments in Cluster VI which are different from IV and V in the traffic patterns, here we can see a spike in bad calls on the middle hour when the good traffic is low compared to other hours. These sub-clusters provide valuable insights into the nature of performance issues. For example, segments in Cluster IV suggest recovery trends that can be monitored for stabilization, while segments in Cluster V require immediate attention to prevent further degradation. Cluster VI, on the other hand, highlights unpredictable behaviour that may be linked to transient issues such as interference or hardware faults.

The two-tier clustering approach demonstrates its effectiveness in breaking down complex performance issues into manageable categories. This granularity enables network operators to prioritize and tailor their interventions, improving overall efficiency in network management.

5.4.2 Classification

The labeled data from the clustering process serves as input for the classification stage, where an SVM-based classifier is trained, validated, and tested. The classifier's

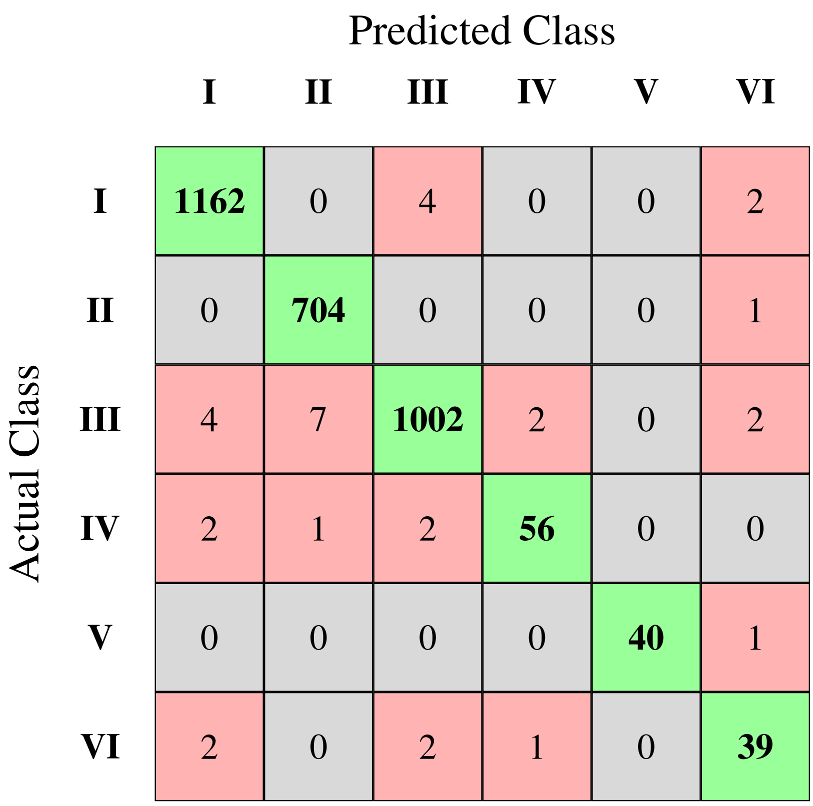

Figure 5.6: Confusion matrix

performance is summarized below.

Model Performance

Results of three-fold cross-validation are shown in Table 5.2. The classifier achieved a mean accuracy of 99.39% during validation using a linear kernel and a penalty parameter (C) of 1000. This high accuracy indicates the robustness of the model and its ability to generalize well across the training data. When applied to unseen test data, the classifier maintained an impressive accuracy of 98.91%. It can be seen from the mean training and validation score for different sets of parameters shown in Table 5.2 that the model performance is consistent.This consistency between training and test performance reduces concerns about overfitting, confirming that the model can reliably classify new data.

Confusion Matrix Analysis

The confusion matrix in Fig. 5.6 provides a detailed view of the classifier's performance across the six clusters. Key observations include:

- High accuracy in identifying large clusters, such as Clusters I, II, and III, which represent segments with good performance.

- Balanced performance across smaller clusters, such as Clusters IV, V, and VI, with accuracies of 92%, 98%, and 88%, respectively. This demonstrates that the model is not biased toward larger clusters, a common issue in classification tasks.

- Minimal misclassification errors, indicating that the feature sets used for training effectively capture the distinguishing characteristics of each cluster.

The ability to accurately classify segments in smaller clusters, such as Cluster VI, is particularly noteworthy. These clusters represent critical performance issues that may otherwise be overlooked in traditional approaches. The classifier's effectiveness in this regard underscores its value as a tool for proactive network management.

Implications for Network Optimization and Maintenance

The classification results have several practical implications for network optimization:

- Proactive Maintenance: By accurately identifying and classifying problem segments, the framework enables network operators to address issues before they escalate, reducing downtime and improving user experience.

- Resource Allocation: The ability to differentiate between stable and unstable performance segments allows for more efficient allocation of resources, ensuring that efforts are focused on areas with the greatest need.

- Scalability: The high accuracy and consistency of the classifier make it suitable for deployment in large-scale networks, where manual monitoring and intervention are impractical.

5.5 Conclusion

As we approach the deployment of 5G and design of 6G networks, the increasing complexity of mobile networks emphasizes the need for automated performance monitoring to ensure optimal Quality of Experience (QoE). This chapter presented a novel machine-learning framework for classifying network node performance using Call Detail Records (CDR) data. The proposed method combines a two-tier clustering approach with a Support Vector Machine (SVM) classifier, achieving an accuracy of 98.91% in near-real-time classification.

The framework effectively identifies performance degradation and categorizes its type, enabling network operators to prioritize and address issues swiftly. By overcoming traditional silo-based approaches, it provides a scalable and user-centric solution, aligned with the broader goal of predictive modeling for cellular network optimization.

Future work can extend this method to include multiple performance indicators and proactive predictive capabilities, paving the way for more comprehensive network management strategies.

Chapter 6

Leveraging Intelligence from Network CDR Data for Energy Consumption Minimization

As emphasized throughout this thesis, network densification is a cornerstone of 5G and future cellular networks, addressing the escalating demand for mobile traffic. Co-channel small cells (SCs), which reuse spectrum with macro cells (MCs), are pivotal in this densification. However, spectrum reuse introduces inter-cell interference (ICI) that can degrade network performance if not managed effectively. Furthermore, while SCs have lower power consumption than MCs, their load-independent power component constitutes a significant portion of their overall energy usage. This ”always-on” operation exacerbates energy inefficiency in ultra-dense networks. Consequently, achieving the vision of 5G and beyond requires addressing two persistent challenges: high ICI and aggregated energy consumption.

This study proposes a proactive approach leveraging real-world Call Data Records (CDRs) to predict spatio-temporal user activity, optimizing SC operational states dynamically. Unlike reactive designs prevalent in the literature, the proposed framework integrates ICI mitigation with energy optimization, minimizing interference and resource wastage while meeting 5G's ambitious quality-of-service (QoS) requirements. To the best of the author's knowledge, this is the first study to exploit CDR data for proactive scheduling of SC sleep cycles while concurrently addressing ICI and energy efficiency (EE) challenges.

Several studies have addressed ICI and EE challenges in heterogeneous networks (HetNets). SC-to-SC interference mitigation has been explored in [119], [120]. However, the interference from SCs to MC users often has a greater impact on network performance due to the higher number of MC users. To tackle this issue, authors in [121–123] propose techniques to reduce SC-induced interference for MC users, balancing user experience and network efficiency.

An alternative line of research involves spectrum partitioning between cell-center and cell-edge users, as seen in [124–126]. While effective for ICI reduction, these methods sacrifice overall capacity, making them less scalable in ultra-dense deployments. Joint EE and ICI management strategies have been explored in [127–129], focusing on interference reduction to improve energy throughput efficiency. However, these approaches fall short of optimizing underutilized network resources during off-peak hours.

Traffic-aware transmission strategies, proposed in [130–133], address this gap by recommending underutilized base stations (BSs) transition to sleep mode during off-peak hours. While promising for improving EE, these strategies are often decoupled from comprehensive ICI mitigation frameworks. Other studies, such as [134–136], explore EE and ICI jointly but focus primarily on uplink transmissions and user equipment (UE) energy savings, leaving downlink energy optimization largely unaddressed despite its significant contribution to operational expenditures (OPEX).

Existing approaches are predominantly reactive, responding to changes in traffic or interference after they occur. These methods may fall short in meeting the real-time demands of 5G and beyond. Building on the Big Data-empowered Self-Organizing Network (BSON) vision [13], this study adopts a fundamentally proactive approach, leveraging CDR data to dynamically optimize SC operations. This framework integrates traffic-aware scheduling with interference mitigation, demonstrating the potential for predictive analytics to enhance network performance and resource efficiency.

The utility of CDR data in network planning and optimization has been highlighted in [137] and [138], where real-world CDR analysis revealed insights often overlooked by analytical models or synthetic datasets. However, prior studies have not utilized CDR data for scheduling SC sleep cycles while simultaneously addressing ICI and EE challenges. This research fills that gap by combining real-world data analysis with predictive modeling to propose a unified framework for energy-efficient, interference-optimized HetNets. This chapter builds upon the data-driven and machine-learning-based methodologies introduced in Chapters 3–5, extending their application to practical network optimization challenges in dense cellular environments. Specifically, it demonstrates how insights from clustering and classification models can inform resource allocation and energy-saving decisions in a predictive framework.

The main contributions of the revised Chapter 6 are:

- Energy Efficiency and Interference Optimization: A joint optimization framework is developed to minimize energy consumption while ensuring

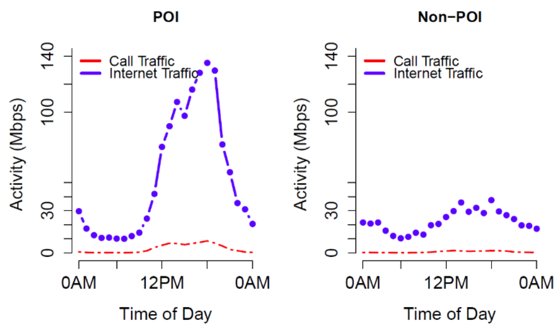

Figure 6.1: Call activity level for POI and non-POI

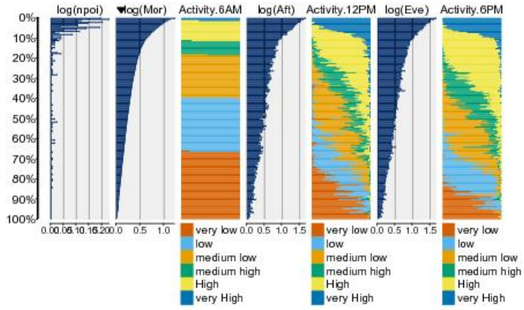

Figure 6.2: Internet activity level for POI and non-POI

minimum guaranteed throughput. The novel contribution is the integration of base station sleep-mode mechanisms with resource allocation strategies, supported by a heuristic low-complexity solution to this NP-hard problem.

- CDR-Driven Small Cell Deployment and Evaluation: Real-world CDR data from Milan City is analyzed to predict spatio-temporal traffic patterns. These insights are applied to design a realistic small cell deployment scenario and evaluate the proposed ECA scheme, achieving up to 8x energy savings under low traffic conditions.

These contributions integrate predictive modeling (Chapter 3), mobility analysis (Chapter 4), and performance classification (Chapter 5) into a cohesive framework for energy-efficient, interference-aware HetNet management.

6.1 Leveraging CDR Data for Proactive ICI–EE Solutions

6.1.1 CDR Analysis and Traffic Insights

The analysis of Call Data Records (CDR) data is a critical foundation for developing proactive scheduling strategies to optimize small cell (SC) operations. Using real-world data from the Telecom Italia Big Data Challenge, the study identifies spatio-temporal patterns in network activity. These findings complement and extend the methodologies presented in Chapter 3, particularly regarding mobility and traffic prediction.

Traffic Periodicity: The analysis reveals consistent periodic patterns in network activity for both Point-of-Interest (POI) and non-POI cells. Fig. 6.1 and 6.2 illustrates the aggregate traffic load over 24 hours, showing:

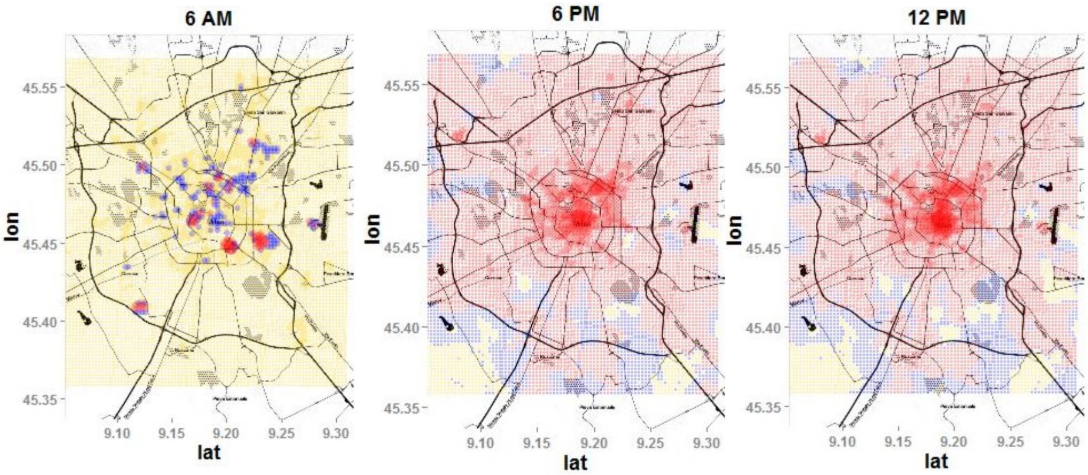

Figure 6.3: Milan activity level at 6 am, 6 pm, 12 pm

Daytime Peaks: POI cells exhibit significantly higher traffic levels during business hours (8 AM–11 PM) compared to non-POI cells.

Off-Peak Hours: Both POI and non-POI cells experience minimal activity during off-peak hours (midnight–6 AM), highlighting opportunities for SC sleep mode scheduling. In Figures 6.3 morning, midday, and evening activity levels are shown, reinforcing the dominance of low activity in the early morning hours. These insights support proactive SC transitions to sleep mode, reducing energy consumption without compromising user experience.

The periodicity observed here aligns closely with temporal mobility patterns discussed in Chapter 3. The grid-based approaches from Chapter 3, particularly hierarchical grids like Geohash, can complement the CDR-based periodicity analysis by offering a spatial indexing framework to enhance resource optimization during high and low traffic periods.

Spatial Variation: Geographic differences in traffic intensity are pronounced, as demonstrated in Fig. 6.4 and Fig. 6.5, POI cells consistently demonstrate higher activity due to their significance (e.g., commercial hubs and landmarks). Conversely, non-POI cells often remain underutilized, suggesting the need for location-aware resource allocation and SC scheduling strategies.

Chapter 3's hierarchical grid systems (e.g., Geohash) are directly applicable for mapping spatial traffic variations. By linking the traffic variations highlighted here with spatially indexed data structures from Chapter 3, more efficient resource distribution can be achieved, particularly in balancing SC usage between POI and non-POI areas.

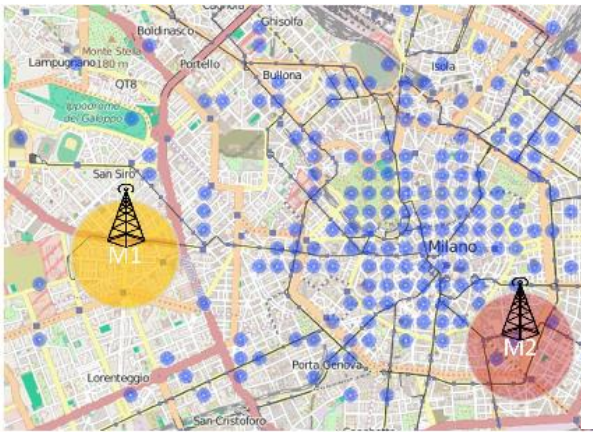

Figure 6.4: POI on Milan map

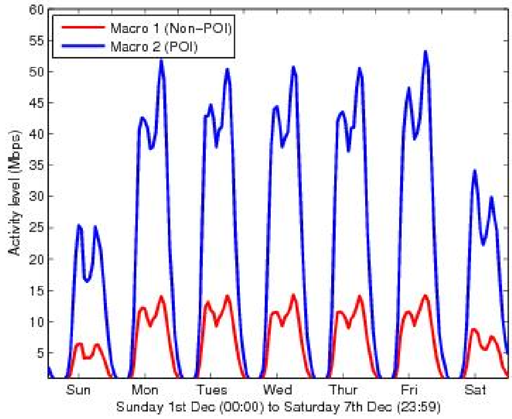

Figure 6.5: Activity level for non-POI and POI macro cells

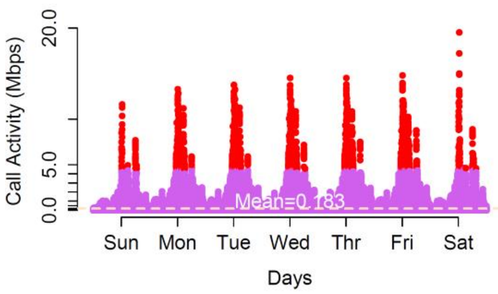

Figure 6.6: Weekly call activity

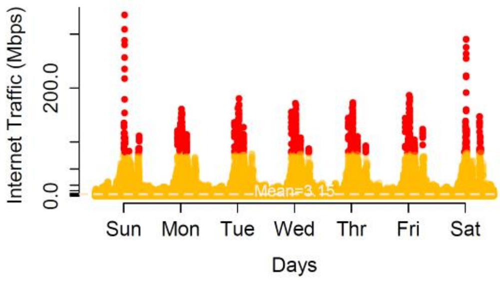

Figure 6.7: Weekly internet activity

Predictability: Weekly trends (as shown in Fig. 6.6 and 6.7) underscore the predictability of user behaviour. Internet activity consistently surpasses calling activity throughout the week, maintaining stable and periodic patterns. This predictability provides a robust basis for forecasting future network conditions and enabling dynamic SC optimization.

Predictive models such as regression and classification techniques discussed in Chapter 3 can be extended to utilize the predictable patterns identified here. By integrating these models, the predictive mobility patterns in Chapter 3 can be enriched with traffic insights from CDR analysis, enhancing the ability to proactively manage SC operations.

6.1.2 System Model and Problem Formulation

This section outlines the comprehensive system model for a heterogeneous network (HetNet) comprising a macrocell (MC) and multiple small cells (SCs) within its coverage area. This model underpins the proactive resource allocation and energy efficiency strategy central to this study. The integration of SCs into the MC's frequency

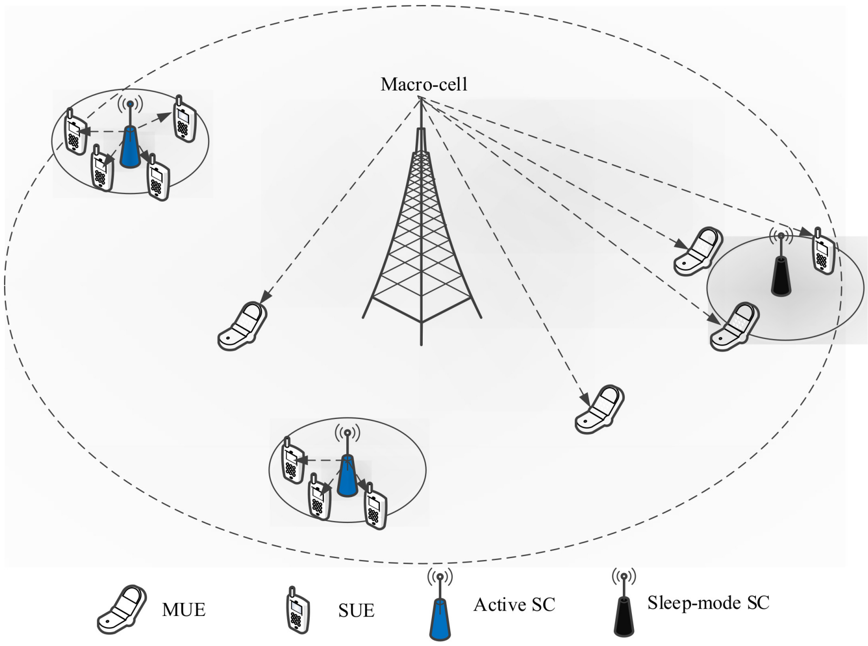

Figure 6.8: System Model with Heterogeneous Cells

spectrum enhances network capacity but also introduces significant interference challenges, necessitating coordinated resource allocation and energy optimization strategies.

HetNet Architecture

The HetNet under consideration includes one macrocell (MC) and M small cells (SCs), which reuse the same spectrum as the MC. This reuse enhances spectral efficiency, a critical requirement for dense 5G deployments. However, this setup also introduces inter-cell interference (ICI) due to overlapping frequency usage. To address this, a coordinated resource allocation strategy is essential to manage interference effectively while maintaining quality of service (QoS) for all users.

The user distribution within this HetNet is dynamic, with K active users connected to either the MC or SCs. The system model assumes that each user is served by one cell at a time, while each cell can support multiple users. This setup reflects the realistic complexity of dense HetNets and serves as the foundation for our optimization strategy. The dynamic user distribution within this HetNet aligns with the spatial traffic variation patterns modeled in Chapter 3, where grid-based approaches were used to predict user mobility and traffic trends.

Key Model Components

The HetNet model includes several critical components, described below:

Resource Blocks (RBs): The total system bandwidth is divided into N resource blocks (RBs). Each RB can be allocated to a single user within a cell to avoid intra-cell interference. Fig. 6.8 illustrates the HetNet layout, showing the spatial arrangement of the MC, SCs, and users, as well as the overlapping resource allocation zones. This figure is crucial for understanding how resource allocation decisions affect interference patterns and overall network performance.

Binary Indicators for Operational States and Allocation: The operational state of each SC is represented by the binary indicator c_m, where:

$$c_m = \begin{cases} 1, & \text{if SC is active,} \\ 0, & \text{if SC is in sleep mode.} \end{cases}$$

Similarly, the RB allocation is represented by $f_{k,m,n}$, where:

$$f_{k,m,n} = \begin{cases} 1, & \text{if user } k \text{ in cell } m \text{ is allocated RB } n, \\ 0, & \text{otherwise.} \end{cases}$$

These binary variables allow for flexible modeling of resource usage and energy-saving strategies.

Performance Metrics: User performance is modeled using the signal-to-interference-and-noise ratio (SINR) and the Shannon capacity formula. The SINR for user k in cell m on RB n is given by:

$$\gamma_{k,m,n} = \frac{p_{m,n} G_{m,k,n}}{\sum_{i \neq m} p_{i,n} G_{i,k,n} + N_0},$$

where:

- $p_{m,n}$: Transmit power of cell m on RB n,
- $G_{m,k,n}$: Channel gain between cell m and user k,
- N_0: Noise power.

The user data rate is then estimated using the Shannon-Hartley theorem:

$$R_{k,m,n} = B \log_2 \left(1 + \gamma_{k,m,n}\right),$$

where B represents the bandwidth of each RB.

Energy Efficiency Metrics

Given the critical importance of energy efficiency (EE) in dense networks, the study evaluates EE using the following metrics:

Energy Consumption Ratio (ECR): The energy required to transmit one bit is calculated as:

$$ECR = \frac{P}{D},$$

where

- P is the total power consumed, and
- D is the data rate.
- Lower ECR values indicate higher energy efficiency.

Area Power Consumption: The power consumption normalized by the coverage area is measured in Watts per square kilometer. This metric is particularly useful for comparing energy efficiency across different network deployments.

Problem Formulation

The objective of this study is to develop a resource allocation strategy that minimizes energy consumption while mitigating interference and maintaining network efficiency. This problem is formulated as a mixed-integer non-linear programming (MINLP) problem:

$$\text{Objective:} \quad \min_{f,c,p} \sum_{m=1}^{M} \sum_{n=1}^{N} p_{m,n} \cdot c_m,$$

subject to:

- SINR constraints to ensure user QoS,
- Resource allocation constraints to avoid intra-cell interference,
- Power and operational state constraints for SCs.

This optimization problem is inherently complex due to its combinatorial nature, requiring heuristic approaches for practical implementation. To address this complexity, we propose an Energy Consumption Aware Resource Allocation Scheme (ECA). The flowchart in Fig. 6.9 provides a step-by-step overview of the ECA algorithm presented by Algorithm 1, illustrating the processes of RB allocation, SC state transitions (ON/OFF), and interference management. The diagram helps to clarify how the heuristic algorithm works iteratively to achieve energy efficiency while maintaining QoS for both MUEs and SUEs.

6.1.3 Proposed Energy Consumption Aware Resource Allocation (ECA) Scheme

To overcome the challenges of energy efficiency (EE) and interference mitigation in dense heterogeneous networks (HetNets), we propose the Energy Consumption Aware (ECA) Resource Allocation Scheme. This heuristic approach addresses the computational complexity of real-time resource allocation while ensuring efficient operation in scenarios with dynamic user demands and dense SC deployments. By leveraging predictive modeling based on historical CDR data, the ECA scheme proactively optimizes Resource Block (RB) allocation and adjusts Small Cell (SC) operational states.

The ECA scheme's unique focus on proactive adjustments, underutilized SC transitions, and computational efficiency positions it as a practical solution for real-time implementation in Self-Organizing Networks (SONs). This is especially critical in the ultra-dense 5G and 6G environments discussed in Chapters 3 and 5,

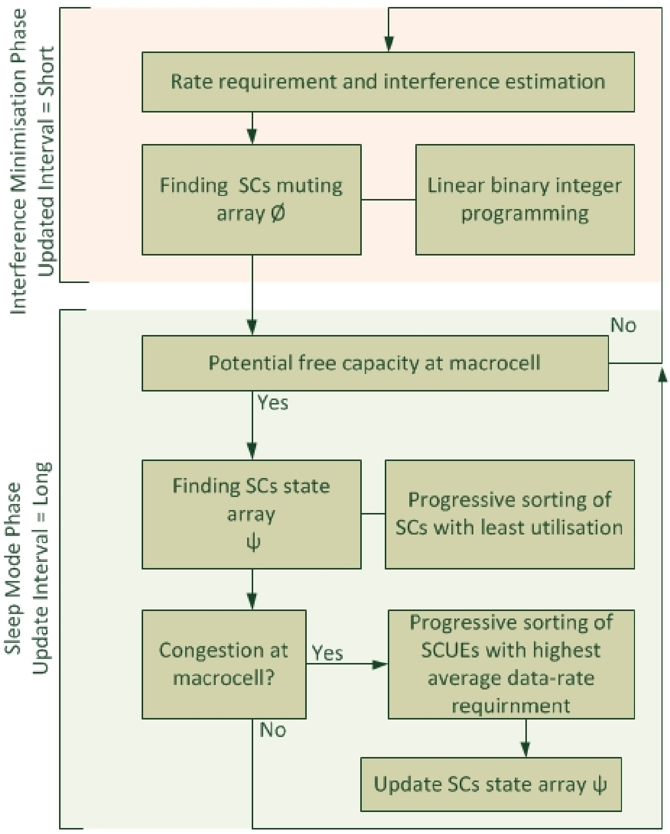

Figure 6.9: Flow diagram for ECA scheme

where user mobility, traffic variations, and network optimization are intricately linked.

Algorithm 1 : Energy Consumption Aware Resource Allocation (ECA)

for $n = 1 \rightarrow N$

  Initialise : $\vec{\phi}_n = \vec{0}$

  Calculate : $\Omega_n^{\max}$,$\omega_{u,0,n}^m$,$R_{k,m,n}$ as in eq. 3, 5 and 6 [75]

  $\vec{\phi}_n$= *bint prog* ($\vec{R}_n$, $\vec{\omega}_{u,0,n}$, $\Omega_n^{\max}$)

end

Notify SCs with their respective $\phi_{m,n}$.

Small Cell Sleep Mode Phase

Analyse Available n RBs at Macrocell

if $RB_0^{\text{Avail}} > RB_0^{\text{Thres}}$

  Sort small cell utilisation $\vec{U}$ in descending order

  while $RB_0^{\text{Avail}} < RB_0^{\text{Thres}}$

    Send sleep mode activation message to SC on top of $\vec{U}$.

    Update RB_0^{Avail}, Remove top element from $\vec{U}$.

  end

Small Cell Wake-up Phase

For every SC in sleep mode **do**:

if $\sum_k RB_{k,m}^{\text{Req}} > RB_0^{\text{Avail}}$ **OR** $C_0 = 1$

  Sort $\sum_k RB_{k,m}^{\text{Req}}$ in ascending order for all m

  while $\sum_k RB_{k,m}^{\text{Req}} > RB_0^{\text{Avail}}$ **OR** $C_0 = 1$

    Send wakeup message to SC on top of the list

    Update RB_0^{Avail}

  end

end

Key Components of the ECA Scheme

Resource Block Allocation The ECA scheme begins with RB allocation. The Macro Cell (MC) takes a central role by:

Allocating RBs to its Macro User Equipment (MUEs) based on their minimum data rate requirements and SINR constraints. Guiding SCs on their muting parameters to ensure interference from SCs to MUEs remains within acceptable thresholds. This process is critical for minimizing inter-cell interference (ICI) while maintaining the Quality of Service (QoS) for MUEs. This step incorporates Fig. 6.8, which visually illustrates the RB allocation dynamics in a HetNet with multiple SCs and an MC. The diagram shows how RBs are distributed among users and how interference management is achieved through coordinated allocations.

SC Sleep Mode Decision Underutilized SCs are identified based on traffic conditions, which are predicted using historical CDR data. SCs with minimal load or high interference factors are transitioned to sleep mode if:

The MUEs and Small Cell User Equipment (SUEs) associated with the SC can be seamlessly accommodated by the MC. The MC has sufficient resources to handle the additional load without compromising its performance. This step emphasizes energy efficiency by reducing the idle power consumption of SCs during off-peak hours. Fig. 6.9 (Flowchart of the ECA Algorithm) provides a clear overview of this process, showing how RB allocations and SC operational states are iteratively optimized.

Proactive Adjustments with Predictive Insights The ECA algorithm presented as Algorithm 1 leverages predictive modeling of user behavior, as explored in Chapter 3, to anticipate traffic trends and make preemptive adjustments to resource allocations. Historical CDR data, analyzed for spatio-temporal traffic patterns (e.g., traffic periodicity and spatial variation from Chapter 3), provides the basis for these forecasts.

For example:

During off-peak hours, SCs in residential areas (non-POI cells) can be transitioned to sleep mode based on low expected traffic volumes. In contrast, SCs in POI areas with consistently high activity are allocated additional resources to ensure seamless operation. These proactive adjustments reduce the reliance on reactive mechanisms, which are often slower and less efficient in ultra-dense network scenarios.

Advantages of the ECA Scheme

Low Computational Complexity: The heuristic nature of the ECA scheme ensures that it operates efficiently, making it suitable for real-time applications in SONs. Unlike exhaustive search methods, the ECA algorithm achieves sub-optimal solutions with significantly reduced computational overhead.

Energy Efficiency Gains: By transitioning underutilized SCs to sleep mode and optimizing RB allocations, the ECA scheme achieves notable reductions in network energy consumption, aligning with the energy-saving strategies emphasized in Chapter 5. Interference Mitigation: Coordinated RB allocations and SC muting ensure that interference remains manageable, preserving QoS for both MUEs and SUEs.

6.2 Results and Analysis

This section evaluates the effectiveness of the proposed Energy Consumption Aware (ECA) algorithm in leveraging spatio-temporal traffic predictability for energy savings and interference management within dense HetNet deployments. By integrating real-world CDR data and deterministic simulation models, the analysis highlights the algorithm's role in achieving the thesis's objectives of predictive modeling and proactive network optimization.

6.2.1 Traffic Predictability and Analysis

The analysis of real CDR data from Telecom Italia's Milan network reveals clear periodicity and predictability in spatio-temporal traffic patterns. Figures 6.1 and 6.2 illustrate weekly variations in call and internet activity, distinguishing between Point-of-Interest (POI) and non-POI cells. POI cells consistently experience higher traffic volumes due to their location-specific significance, such as proximity to commercial hubs, entertainment centers, and transport nodes. Conversely, non-POI cells exhibit lower activity, particularly during off-peak hours.

This predictable behavior, forms the foundation for proactive scheduling strategies. The higher variance in traffic levels for POI cells underscores their dynamic nature, requiring adaptive and granular resource allocation approaches. These findings align with insights discussed in Chapter 3, where predictive modeling techniques are employed to forecast user mobility and network demand. This synergy across chapters reinforces the thesis's focus on predictive analytics as a critical tool for cellular network optimization.

The predictable traffic patterns observed in this analysis validate the premise of the ECA algorithm, which utilizes historical CDR data to anticipate traffic demands and optimize network operations. By exploiting this predictability, the algorithm minimizes energy consumption and interference, achieving tangible improvements in network performance.

6.2.2 Simulation-Based Energy Efficiency Gains

The simulation results, presented in Figures 6.10, provide a detailed evaluation of the ECA algorithm's energy efficiency (EE) performance across POI and non-POI cells. The results demonstrate the algorithm's ability to address the circuit power-dominant nature of small cells and optimize their energy usage:

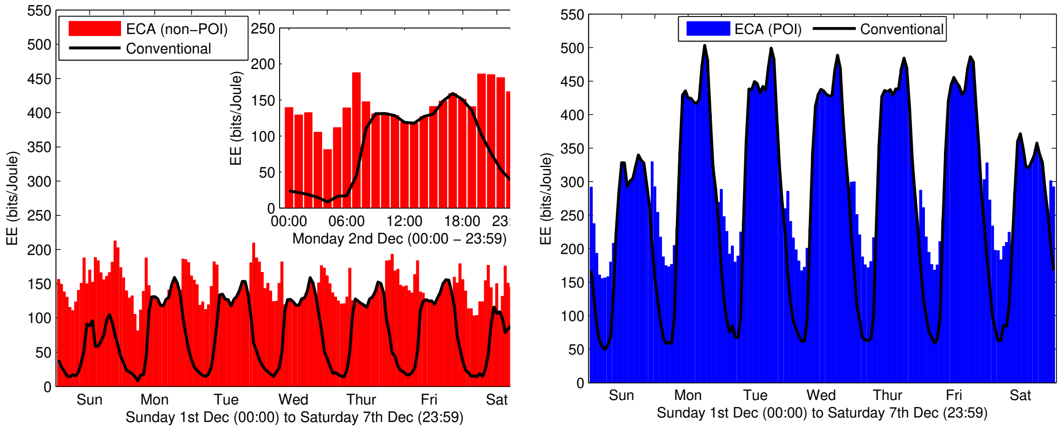

Figure 6.10: EE (bits/Joule) performance for non-POI cells(left), and POI cells (right)

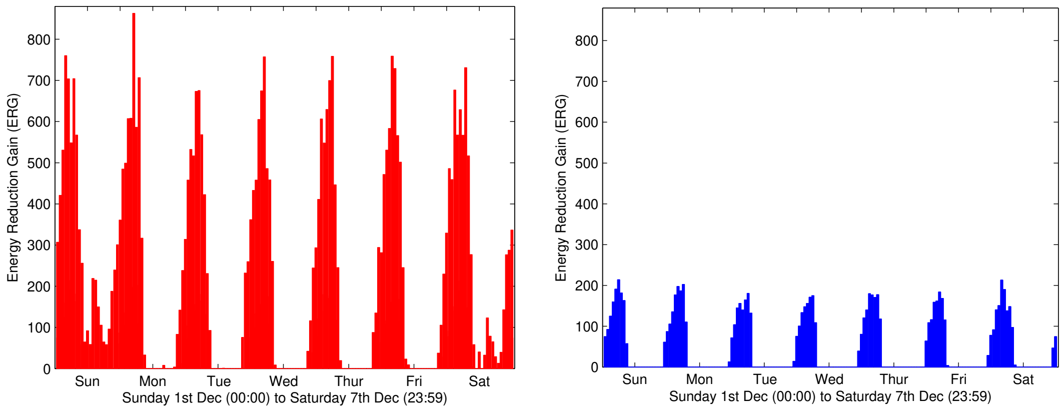

Figure 6.11: Energy reduction gain (ERG) performance for non-POI (left) and POI case (right)

Energy Efficiency Improvements: POI cells exhibit higher EE due to their higher utilization rates, while non-POI cells, often underutilized, achieve significant EE gains with the ECA algorithm. The transition from the conventional always-on operation (solid black line) to the ECA-enhanced operation (red line for non-POI and blue line for POI) underscores the algorithm's impact.

Energy Reduction Gain (ERG): Fig. 6.11 highlights the algorithm's effectiveness in reducing energy consumption. For non-POI cells, the ECA scheme achieves up to 8 times ERG during off-peak traffic conditions, while POI cells exhibit up to 2 times ERG improvement under similar conditions. These results demonstrate the algorithm's ability to optimize energy usage, particularly in underutilized scenarios. The observed improvements align with the predictive modeling insights from Chapter 3, interference-aware scheduling strategies in Chapter 4, and machine learning-based optimizations in Chapter 5, highlighting the integration of these methodologies within the thesis framework.

6.2.3 Service Quality and Scalability

A critical challenge in dense HetNets is maintaining quality of service (QoS) while optimizing energy efficiency. The ECA algorithm addresses this challenge through intelligent resource allocation and dynamic scheduling mechanisms:

SINR Improvements for MUEs: The algorithm mitigates interference from muted SCs, significantly enhancing SINR for MUEs. This is consistent with interference management strategies detailed in Chapter 4, where grid-based mobility prediction is leveraged for resource allocation optimization.

QoS Consistency for SUEs: Dynamic RB allocation ensures that small cell users (SUEs) maintain satisfactory data rates, even during peak traffic conditions. This adaptability highlights the algorithm's ability to balance energy savings with service quality.

The results further demonstrate the scalability of the ECA algorithm across diverse traffic scenarios, from POI to non-POI cells, and its practical applicability for future 5G and beyond networks. By integrating predictive modeling, interference-aware scheduling, and machine learning techniques, the ECA scheme exemplifies the thesis's overarching goal of leveraging data analytics for cellular network optimization.

6.3 Conclusion

This chapter introduced the Energy Consumption Aware (ECA) scheme, aimed at addressing the pressing challenges of energy inefficiency and interference management in dense HetNets by leveraging spatio-temporal traffic patterns derived from real-world CDR data. The ECA algorithm demonstrated its effectiveness in dynamically optimizing small cell operations and resource allocation. Notably, the algorithm achieved energy savings of up to 8x Energy Reduction Gain (ERG) in non-POI cells during off-peak hours and up to 2x ERG in POI cells under similar conditions. These results underscore the significant potential for energy efficiency improvements, particularly in underutilized network segments. Importantly, these gains were realized without compromising the quality of service (QoS) for either macrocell users (MUEs) or small cell users (SUEs), highlighting the robustness of the proactive resource management enabled by traffic predictability.

The methodologies developed in this chapter build upon and integrate insights from earlier chapters. Specifically, the traffic load estimation and predictive modeling techniques explored in Chapter 3 laid the foundation for understanding the temporal and spatial dynamics of network activity. Similarly, the mobility pattern predictions and grid-based indexing strategies detailed in Chapter 4 were instrumental in enhancing the granularity of traffic data analysis, which the ECA scheme leveraged to schedule small cell operations effectively. Furthermore, the performance-based classification and clustering methods from Chapter 5 provided a framework for identifying underutilization and adapting small cell operations to dynamic network conditions.

By integrating predictive analytics from Chapters 3–5, this chapter exemplifies the cohesive framework of this thesis, demonstrating how data-driven approaches can address the multifaceted challenges of energy efficiency, interference mitigation, and QoS assurance in next-generation cellular networks. This alignment reinforces the thesis's overarching goal of leveraging predictive modeling and data analytics for optimized network performance in 5G and beyond.

Chapter 7

Conclusion and Future Directions

7.1 Conclusion

5G and beyond network technologies are becoming part and parcel of our socio-economic system. They are envisioned to connect everything with everything, i.e., massive machines to machines communication is expected along with human to human and human to machine communication. Where it makes the wireless cellular network systems too complex to be managed manually there it leads to the generation of an enormous amount of data carrying valuable information about network and user behaviour. This data can be exploited by machine learning to gather intelligence required to perform SON functions considered as an auto management solution for the complex wireless cellular networks. The capacity of machine learning to predict can help SON enabled networks to be proactive rather than reactive. It can help SONs to better perform two of their key function namely self-optimization and self organization. A major share of the OPEX of networks goes towards the maintenance of network facilities. In this research, classical machine learning is used to develop a generic high-resolution internet traffic prediction model, global mobility pattern prediction models, both helpful towards SON use cases of self-optimization function. Besides that, a classification scheme is developed for the classification of network cells based on their performance that also identifies and categorise sub-performing cells. In addition to that, the traffic load prediction model is applied in a use case of wireless cellular network to address common issues of inter-cells interference and energy consumption. Though deep learning is getting very popular, but one of the objectives of this study had been to exploit classical machine learning to avoid deep learning because deep learning is very complex, computationally costly and data-hungry. Secondly, 5G and beyond networks are expected to rely on the application of edge and fog computing to make fast decisions for the specific tasks in hand. In such cases classical machine learning algorithms become crucial. An SVR based prediction models are proposed in Chapter 3 that predict minimum, maximum, and mean hourly internet traffic density with a prediction accuracy of around 95% much higher than the state of the art statistical and deep learning models. This prediction model can help to enable SON self-optimization function for uses case like radio resource management, load optimization, energy optimization etc. Towards the enabling of the same self-optimization function, in Chapter 4, for the first time such a comprehensive study is presented on mobility pattern prediction schemes with an empirical comparison of the performance of classical machine learning algorithms with the help of a global indexing system for locating users. Besides the radio resource optimization Users' location prediction can also help in self-optimization function in use cases like handover management, caching etc. Towards another key function of SONs, self-healing, a comprehensive hybrid scheme is developed with the help of classical clustering algorithm K-means and classification algorithm SVC, presented in 5. This scheme groups together cells based on their performance andidentifies sub-performing cells exploiting CDRs. Use of machine learning for the identification and classification of cells based on their performance can help in enabling self-healing function in SON. Cellular network traffic prediction study is further extended to apply in a use case for the joint optimization of inter-cell-interference and energy consumption presented in Chapter 6. Except for the Chapter 4, throughout this study real network CDR data is used for the development of the models. It is concluded that the real network data like CDR can be exploited with the help of classical machine learning algorithms for gathering intelligence needed to enable SON functions, e.g., self-optimization and self-healing in a proactive manner. It is also established that classical machine learning algorithms can outperform deep learning algorithms for the tasks like traffic density estimation which make them a potential candidate for their use in edge or fog computing.

7.2 Future Direction

Research presented in this thesis aimed at developing broad comprehensive prediction and diagnosis machine learning models that can work as base models for providing support to enable SON functions of self-optimization and self-healing. As a case study, the prediction model is also applied and evaluated in the cellular network use case for the optimization of interference aware energy consumption. Further research can be conducted to extend or build on the work done so far and presented in this thesis. Some areas are highlighted here with potential for further research as follows:

7.2.1 Prognostics in SON

Prognostics, the capability to predict failures or performance degradation before they occur, offers a significant opportunity to enhance the self-healing capabilities of SONs. By leveraging historical data and machine learning techniques, prognostics can provide timely insights into potential issues, allowing preemptive actions that reduce downtime and improve network reliability.

Prognostics models use data such as Call Detail Record (CDR) trends, KPIs (e.g., signal quality, call drops, and throughput), and hardware performance metrics to anticipate faults. For instance, machine learning algorithms like Long Short-Term Memory (LSTM) networks and Support Vector Regression (SVR) can model these trends to detect patterns that signal impending failures. By predicting such events, operators can proactively perform maintenance, reallocating resources and ensuring service continuity. For example, cells prone to traffic congestion during peak hours can be identified in advance, enabling network adjustments that balance loads across the system.

Incorporating prognostics into SON functions not only improves fault management but also aligns with the thesis's broader objective of leveraging predictive analytics for proactive decision-making. Chapters 5 and 6 lay the foundation for such advancements by exploring methods to detect underperforming cells and optimize resource usage dynamically. Building on this, future research can focus on integrating heterogeneous data sources—such as hardware logs, user behavior, and environmental conditions—to improve the accuracy of prognostics models. Additionally, ensemble machine learning models can be developed to combine multiple algorithms for enhanced fault prediction reliability, and uncertainty quantification techniques can be explored to provide confidence levels for predictions, enabling more informed decision-making.

7.2.2 Reinforcement Learning in SON

Reinforcement Learning (RL) has emerged as a powerful tool for enabling dynamic decision-making in SONs. Unlike traditional approaches, RL models learn from the environment by iteratively adjusting their policies to maximize long-term rewards, making them highly suitable for complex and adaptive tasks in cellular networks.

In the context of SONs, RL can address several critical challenges. For example, RL models can dynamically allocate Resource Blocks (RBs) to Mobile User Equipment (MUEs) and Small User Equipment (SUEs), balancing resource efficiency and interference mitigation. By analyzing traffic patterns and resource utilization, RL can optimize energy efficiency by deactivating underutilized small cells during off-peak hours, thereby reducing unnecessary energy consumption. Additionally, RL can significantly enhance fault recovery processes. By learning optimal strategies for rerouting traffic or reconfiguring network parameters, RL ensures minimal service disruption during cell outages or performance degradation. Mobility management is another area where RL can play a pivotal role, as highlighted in Chapter 4. Predicting user mobility patterns with RL facilitates seamless handovers, reducing latency and maintaining Quality of Service (QoS).

Future advancements in RL for SONs can further enhance these capabilities. Developing lightweight RL models that operate efficiently in edge computing environments will be crucial for real-time decision-making. Multi-agent RL, where multiple agents collaborate to optimize the entire network, can also address the challenges posed by large-scale heterogeneous networks. Furthermore, integrating RL with supervised learning techniques can overcome the ”cold start” problem, allowing RL models to perform effectively even in unfamiliar network conditions. Finally, RL’s potential can be expanded to focus on improving user Quality of Experience (QoE). By incorporating real-time feedback from users, RL models can dynamically adjust network policies to ensure an optimal experience for end-users.

Incorporating RL into SON functions, as demonstrated in Chapters 4, 5, and 6, reinforces the thesis’s overarching goal of integrating data-driven methods for cellular network optimization. By leveraging RL’s dynamic adaptability and ability to learn from complex environments, SONs can achieve significant advancements in resource allocation, energy efficiency, and fault management.

7.2.3 Machine Learning on Big Data

Through the machine learning applied on real data in this research, it is established that real network data has valuable information and machine learning can exploit that data to gather intelligence needed to make futuristic cellular network more smart and independent. Except for the Chapter 4, throughout this research only CDR data is used for the tasks like internet traffic estimation in Chapter 3 fault detection in Chapter 5 and overall traffic load prediction in Chapter 6. The volume of the data used in this research has been significantly huge which helps in identifying and predicting important patterns about network behaviour. But the fact is, including CDR data gigantic amount of heterogeneous data is generated from different ends of the network as highlighted in Chapter 2. This data can be rightly characterized as big data and has the same potential as attributed to big data. This cellular big data is a mine of valuable information and can be exploited using proper big data analytic and machine learning approach. So machine learning applied on the big data with proper big data analytic steps as few listed below can open a new world of meanings for the future cellular networks.

Data Fusion

Heterogeneous data generated from different ends of the network can be exploited independently with the help of machine learning to meet requirements for different KPIs. But data fused from multiple sources can be more meaningful, reliable and informative but it becomes more complicated and huge in volume that needs to be treated as big data for storage, processing and machine learning. For example, as mentioned earlier, in this study the main source of information has been CDR data which is itself enormous in volume and independently sufficient to gather intelligence to perform certain tasks like the prediction of change in cell association through machine learning applied onCDR data which can be helpful in better management of handover for mobile users. But a mobile user may be using a mobile app e.g. watching a video on some popular online video platform if a machine learning model is trained not only on spatiotemporal data of location but also on the mobile app usage behaviour of the users it can not only predict when to perform a handover but it can also inform the network which data stream need to be cached on which base station at what time. So in such cases a fused data from data of user's location with time line and data of history of mobile apps usage can be very helpful. So the research work already done and presented in this thesis can be extended by applying machine learning on fused data to facilitate single SON function or multiple SON functions at the same time.

Distributed Storage and Parallel Processing

Though in this study as well parallel processing is used for the efficient processing wherever it could be possible like in SVR based model development in Chapter 3 and big data ecosystem is used in Chapter 6. But here, main purpose has been to expedite the model development process. But considering the big data aspect, it becomes inevitable to use big data processing tools like distributed storage and parallel processing for the data processing, analysis and machine learning. So the existing research presented can be extended further to develop big data focused models. Research work can be performed to develop machine learning algorithms that can crunch enormous heterogeneous data streams in parallel to train single model. For example, in fault diagnosis, when there are hundreds of thousands of sites and over a single day, performance of each site can be evaluated multiple time for a shorter interval of times. If the model evaluate each site one after the other, it will take too much time to evaluate the performance of all sites and by the time it evaluates them, the results may not be useful any more. On the contrary, with a model that can run in parallel and evaluate performance of multiple sites at the same time, it can help in timely or near real-time decisions.

7.2.4 Traffic Load Estimation and Applications

The focus of this research has been to develop prediction and diagnosis models using classical machine learning algorithm for being computationally low cost and less data hungry during training as compared to deep learning algorithms. In Chapter 3, prediction model is developed using classical machine learning algorithm Support Vector Machine and its performance is compared against some state of the art statistical and deep learning algorithms for the prediction of three levels of internet traffic. This research can be possibly extended in any of the following three dimensions:

1. Though, in this research traffic density is already estimated at a significantly high resolution of one hour important for decision making in a shorter interval. But this research can be extended for even higher resolutions like few minutes important for future cellular networks with small cells in place where keeping all cells active with the maximum resources is not optimal. Key challenges for extending research in high resolution will be to address the data sparsity as in smaller areas (i.e., small cells) and the shorter intervals very few samples can be expected. Use of big data and mitigation techniques for data sparsity can help to address this challenge.

2. Research presented in Chapter 3 covers the development of machine learning based prediction model and comparison of its performance with the other state of the art algorithms. This research can be extended to evaluate the application of the models developed in the relevant use cases of cellular networks. For example, to evaluate, how much it can help in saving energy or optimize radio resources when the predicted values are used in some optimization model as compared to 'always on' or maximum resources allocated scenarios.

3. It is well-known fact that deep learning algorithms are computationally complex, expensive and data hungry but they can lead to better accuracy. But with our prediction model developed and presented in Chapter 3, it is established that conventional machine learning models can outperform complex deep learning models in terms of accuracy. But the potential of deep learning models can not be undermined as they can work on raw heterogeneous data. Hence, the research can be extended to develop such lighter versions of deep learning models for the traffic load estimation which do not only have good accuracy but they are also efficient in terms of computational cost.

7.2.5 Mobility Pattern Predictions and Applications

To the best of my knowledge it was the first time that such study is performed that comprehensively discusses the machine learning based mobility pattern prediction schemes and present comparative study of some popular machine learning algorithms for the prediction of mobility patterns with the help of a global indexing system in the context of cellular networks, for details please see Chapter 4. The study presented is very generic and can be used in multiple use cases of SON functions in wireless cellular networks. Further research can be conducted in the mobility pattern prediction in the following potential areas:

1. Further research can be conducted by reproducing results for some other global indexing system. User location is very important information in wireless cellular networks for the provision of services. Commonly in research grid systems are defined by researchers for the study of users association with cells and their mobility patterns. But there is need of global indexing system which can help in the transferability and scalability of the machine learning models. Hence further research using different global indexing systems can be a very helpful trend for the future cellular networks.

2. As mentioned earlier the main goal of this research has been the development of prediction and diagnosis models using conventional machine learning algorithms rather than its applications. So the research presented on mobility pattern prediction in Chapter 4 can be extended toward the implementation of these models in SON use case. For example to evaluate, how it can improve the handover, load balancing or content caching.

3. Third dimension to extend this work can be to develop deep learning models for the prediction of mobility patterns and compare their performance with conventional machine learning models. It can be good contribution if such model is developed which are not only more accurate but also computationally efficient.

7.2.6 Fault Diagnosis

A comprehensive scheme for the identification and classification of sub-performing cells is presented in Chapter 5. A clustering model is developed for the identification of sub-performing cells as the data is not labelled. Cells are grouped into clusters, based on similarity and labelled with the help of domain knowledge. This study can be extended further for the auto diagnosis on top of detection of sub-performing cells. So the model not only identifies that there is some issue with cell but it also diagnoses the reason of fault like temporary traffic congestion, some equipment failure or high traffic loads at some cells with not sufficient resources. Some of the reasons can be identified by applying data analytic on a single source data like CDR. But others may require big data from multiple sources as an alternative for the domain knowledge and human expertise. For example, if a cell sub performs on specific time intervals with high traffic it can be due to temporary traffic congestion. But if it always sub performs and it has high traffic in general then the cell is overloaded or less equipped. If a cell sub performs due to other technical reasons then the data of other technical reports may help in the identification of root cause. So the research presented here in Chapter 5 can be extended for the further root cause analysis by comparative study of multiple time slots of the same cells or exploiting fused big data from multiple sources.

7.2.7 Energy Efficiency by Proactive Instead of Cell Switching

Energy efficiency is one of the major use cases being targeted in the emerging cellular network. In Chapter 6 research is presented to address not only this issue but also to minimize inter-cells interference another use case for the future wireless cellular networks. In the envisioned ultra-dense 5G networks, there will be hundreds of small BSs deployed in a heterogeneous environment. These BSs will be switched ON/OFF based upon user's requirement for the sake of efficient resource utilization and energy saving. To choose which BSs need to be turned ON, it often requires to know the channel quality which is fed from the UE. However, it is not possible for a turned OFF BS to listen to UE so that it can know the radio channel information. Here this BS selection procedure can be enabled through data driven machine learning. The idea is to collect the history of the channel quality indicator (CQI) of different locations in the coverage. By analyzing the history of the channel information with respect to those different locations, the radio signal quality nearby any BS can be predicted and thus efficient BS switching can be enabled. In a user-centred multi-tier heterogeneous network, the macro cell and the small cells have coordinated resource allocation for uplink soundings reference signals. Based on the uplink random access signals of new UEs or UEs with handover requirement or traffic variation of UEs in the network, the central controller or macro BS will decide whether and which dormant smalls cells need to be activated and which active cell should go to dormant mode. This decreases UE access latency and achieves maximum possible energy saving according to dynamic traffic variations. Thus machine learning can enable optimal implementation of small cell activation and UE access to the network avoiding the waste of energy in a heterogeneous environment. Here again, study presented in **??** can be extended with the fused big data from multiple sources instead of just once source like CDR data although it was also huge in volume. But data from multiple sources can help to make more robust and timely decisions.

7.2.8 Integrated Model

Though in this study a conscious effort is made to develop generic models that can be helpful to address multiple use cases of SON. For example, internet traffic prediction model in Chapter 3 and model of mobility prediction patterns in Chapter 4 can be used for multiple use cases of self-optimization in SON like for energy efficiency, load balancing, caching, handover management etc. But still, the development of a single integrated comprehensive model to address maximum use cases of SON functions is an ideal scenario. Research can be further extended towards that end by integrating multiple models as modules into one algorithm or by developing one comprehensive algorithm that takes heterogeneous big data but it is also robust to missing data and make accurate predictions and decisions related to several use cases of SON functions. For example, a model can be envisioned with multiple modules and based on the end KPIs or PIs it automatically decides the data set and model to be used and make decisions accordingly or a comprehensive single big data based algorithm is developed that ignores the missing streams and exploits the available data streams to make decisions help for performing SON functions in different use cases.

7.2.9 Light Efficient Models

As highlighted already that focus in this study had been on the use of classical machine learning algorithms instead of deep learning algorithms because deep learning algorithms are complex, computationally costly and data hungry. But deep learning algorithms have a huge potential of developing comprehensive single algorithm for all or multiple use cases. But such comprehensive algorithm is expected to be even more complex whereas it is envisioned that in the future wireless cellular networks some decisions need to be made on centre but some decision needs to be delegated as referred to the concepts of edge or fog computing. Running such complex models on the edge or not practical not only because of the limitations of computational resources but also because the tasks to be performed on the edge can be very specific. Light and efficient machine learning models are needed there to address such challenges. Classical machine learning algorithms developed here can be crucial in such scenarios for specific tasks or lighter versions of deep learning algorithms need to be developed, deep learning algorithms developed from detailed deep learning algorithms are getting popular in this regard.

7.2.10 Control and Data Plane Separated Architecture

Due to the network densification, heterogeneity of RANs and diverse use cases forecasted in 5G networks, the conventional architectures are not deemed suitable from energy efficiency, efficient planning, and interference mitigation perspectives. Thus a new architecture has been proposed in which the control and data plane are logically separated. A complete survey on this architecture can be found in [11]. In control plane of the architecture, a few macro BS (MBS) provide control signalling to all the users and data transmission to only low data rate users, whereas there are many data BS (DBS) in the data plane that provides high rate data transmission within the macro base station (MBS) footprint. However, interference from the DBS to MBS users is still a challenging problem. In such a scenario, since the MBS has an MDT report from all the users, effective mitigation technique can be developed through coordination and cooperation to overcome the interference between the two planes. The massive MDT report (big data) containing radio signal quality and location information of the users, collected at the control base station (CBS) can be analyzed to come up with comprehensive framework to overcome interference and provide better signal quality. Apart from interference mitigation, if there occurs network fault e.g. cell outage, those DBS can be cooperated to compensate the outage until the problem is rectified. These solutions for proposed architectures are only possible through data driven machine learning.

7.2.11 Bridging Gap between QoS and QoE

Although data driven machine learning promises automatic optimization of QoE of individual subscriber sessions under dynamically changing conditions, the current machine learning solutions do not meet the performance and real-time needs to telecom networks. This is because of the vast size of big data generated in the network. For example, a wireless network with 10 million subscribers runs about 1,000,000 transactions per second during peak hours. To enhance the individual customer's QoE, the insights from the big data needs to be combined with intelligent decisions based on machine learning to automate immediate actions enabling dynamic QoE management. This automatic, prompt and iterative loop between analytics, decision and optimization is still missing.

Bibliography

[1] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, “A survey of machine learning techniques applied to self-organizing cellular networks,” *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2392–2431, Fourthquarter 2017.

[2] H. Fourati, R. Maaloul, and L. Chaari, “Self-organizing cellular network approaches applied to 5g networks,” in *2019 Global Information Infrastructure and Networking Symposium (GIIS)*. IEEE, 2019, pp. 1–4.

[3] 3rd Generation Partnership Project (3GPP), “General packet radio service (gprs) enhancements for evolved universal terrestrial radio access network (E-UTRAN), Technical Specification TS 123.401,” 2018.

[4] Ericsson, “Ericsson mobility report: November 2023,” Ericsson, Tech. Rep., 2023, accessed: 2024-12-02. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report

[5] E. Hossain and M. Hasan, “5g cellular: key enabling technologies and research challenges,” *IEEE Instrumentation & Measurement Magazine*, vol. 18, no. 3, pp. 11–21, 2015.

[6] O. Idowu-Bismark, K. O. Okokpujie, H. Ryan, and M. O. Adedokun, “5g wireless communication network architecture and its key enabling technologies,” *International Review of Aerospace Engineering (I. RE. AS. E)*, vol. 12, no. 2, pp. 70–82, 2019.

[7] S. El Hassani, A. Haidine, and H. Jebbar, “Road to 5g: Key enabling technologies,” *Journal of Communications*, vol. 14, no. 11, pp. 1034–1048, 2019.

[8] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, “5g roadmap: 10 key enabling technologies,” *Computer Networks*, vol. 106, pp. 17–48, 2016.

[9] L. Qian, J. Zhu, and S. Zhang, “Survey of wireless big data,” *Journal of Communications and Information Networks*, vol. 2, no. 1, pp. 1–18, Mar 2017.

[10] X. Cheng, L. Fang, L. Yang, and S. Cui, “Mobile big data: The fuel for data-driven wireless,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1489–1516, Oct 2017.

[11] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, “Control-data separation architecture for cellular radio access networks: A survey and outlook,” *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 446–465, Firstquarter 2016.

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

[13] A. Imran, A. Zoha, and A. Abu-Dayya, “Challenges in 5g: how to empower son with big data for enabling 5g,” *IEEE Network*, vol. 28, no. 6, pp. 27–33, Nov 2014.

[14] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, “Big data analytics in mobile cellular networks,” *IEEE Access*, vol. 4, pp. 1985–1996, 2016.

[15] O. Onireti, A. Zoha, J. Moysen, A. Imran, L. Giupponi, M. A. Imran, and A. Abu-Dayya, “A cell outage management framework for dense heterogeneous networks,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2097–2113, April 2016.

[16] O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, “Cell outage detection in heterogeneous networks with separated control and data plane,” in *European Wireless 2014; 20th European Wireless Conference*, May 2014, pp. 1–6.

[17] 3rd Generation Partnership Project (3GPP), “Universal terrestrial radio access (UTRA) and evolved universal terrestrial radio access (E-UTRA); radio measurement collection for minimization of drive tests (MDT); 3GPP TS 37.320 v11.1.0 ,” 2012.

[18] W. A. Hapsari, A. Umesh, M. Iwamura, M. Tomala, B. Gyula, and B. Sebire, “Minimization of drive tests solution in 3gpp,” *IEEE Communications Magazine*, vol. 50, no. 6, pp. 28–36, June 2012.

[19] B. Sayrac, J. Riihijärvi, P. Mähönen, S. Ben Jemaa, E. Moulines, and S. Grimoud, “Improving coverage estimation for cellular networks with spatial

bayesian prediction based on measurements," in *Proceedings of the 2012 ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design*. New York, NY, USA: ACM, 2012, pp. 43–48.

[20] J. V. D. Beek, T. Cai, S. Grimoud, B. Sayrac, P. Mahonen, J. Nasreddine, and J. Riihijarvi, "How a layered rem architecture brings cognition to today's mobile networks," *IEEE Wireless Communications*, vol. 19, no. 4, pp. 17–24, August 2012.

[21] O. F. Celebi, E. Zeydan, O. F. Kurt, O. Dedeoglu, O. Ileri, B. A. Sungur, A. Akan, and S. Ergut, "On use of big data for enhancing network coverage analysis," in *ICT 2013*, May 2013, pp. 1–5.

[22] J. Puttonen, J. Turkka, O. Alanen, and J. Kurjenniemi, "Coverage optimization for minimization of drive tests in lte with extended rlf reporting," in *21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Sept 2010, pp. 1764–1768.

[23] D. Soldani and I. Ore, "Self-optimizing neighbor cell list for utra fdd networks using detected set reporting," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, April 2007, pp. 694–698.

[24] J. Kurjenniemi, T. Henttonen, and J. Kaikkonen, "Suitability of rsrq measurement for quality based inter-frequency handover in lte," in *2008 IEEE International Symposium on Wireless Communication Systems*, Oct 2008, pp. 703–707.

[25] Z. Su, Q. Xu, and Q. Qi, "Big data in mobile social networks: a qoe-oriented framework," *IEEE Network*, vol. 30, no. 1, pp. 52–57, January 2016.

[26] S. Rao, "Operational fault detection in cellular wireless base-stations," *IEEE Transactions on Network and Service Management*, vol. 3, no. 2, pp. 1–11, April 2006.

[27] F. Xu, Y. Li, M. Chen, and S. Chen, "Mobile cellular big data: linking cyberspace and the physical world with social ecology," *IEEE Network*, vol. 30, no. 3, pp. 6–12, May 2016.

[28] R. Barco, V. Wille, and L. Díez, "System for automated diagnosis in cellular networks based on performance indicators," *European Transactions on Telecommunications*, vol. 16, no. 5, pp. 399–409, September 2005.

[29] G. N. K. B. Cheung, S. G. Fishkin and S. A. Rao, "Method of monitoring wireless network performance," Patent App.10/946,255, 2004.

[30] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "Data-driven analytics for automated cell outage detection in self-organizing networks," in *2015 11th International Conference on the Design of Reliable Communication Networks (DRCN)*, March 2015, pp. 203–210.

[31] Y. Ma, M. Peng, W. Xue, and X. Ji, "A dynamic affinity propagation clustering algorithm for cell outage detection in self-healing networks," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013, pp. 2266–2270.

[32] W. Xue, M. Peng, Y. Ma, and H. Zhang, "Classification-based approach for cell outage detection in self-healing heterogeneous networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2014, pp. 2822–2826.

[33] I. A. Karatepe and E. Zeydan, "Anomaly detection in cellular network data using big data analytics," in *European Wireless 2014; 20th European Wireless Conference*, May 2014, pp. 1–5.

[34] Chernogorov, Turkka, Ristaniemi, and Averbuch, "Detection of sleeping cells in lte networks using diffusion maps," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–5.

[35] F. Chernogorov, S. Chernov, K. Brigatti, and T. Ristaniemi, "Sequence-based detection of sleeping cell failures in mobile networks," *Wireless Networks*, vol. 22, no. 6, pp. 2029–2048, August 2016.

[36] J. Turkka, F. Chernogorov, K. Brigatti, T. Ristaniemi, and J. Lempiäinen, "An approach for network outage detection from drive-testing databases," *Journal of Computer Networks and Communications*, vol. 2012, 2012.

[37] M. Amirijoo, L. Jorguseski, R. Litjens, and L. C. Schmelz, "Cell outage compensation in lte networks: Algorithms and performance assessment," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–5.

[38] W. Wang, J. Zhang, and Q. Zhang, "Cooperative cell outage detection in self-organizing femtocell networks," in *2013 Proceedings IEEE INFOCOM*, April 2013, pp. 782–790.

[39] Q. Liao, M. Wiczanowski, and S. Stańczak, "Toward cell outage detection with composite hypothesis testing," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 4883–4887.

[40] L. Wenjing, Y. Peng, J. Zhengxin, and L. Zifan, "Centralized management mechanism for cell outage compensation in lte networks," *International Journal of Distributed Sensor Networks*, vol. 8, no. 11, November 2012.

[41] M. Amirijoo, L. Jorguseski, R. Litjens, and R. Nascimento, "Effectiveness of cell outage compensation in lte networks," in *2011 IEEE Consumer Communications and Networking Conference (CCNC)*, Jan 2011, pp. 642–647.

[42] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, "Human mobility characterization from cellular network data," *Communications of the ACM*, vol. 56, no. 1, pp. 74–82, January 2013.

[43] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 18–26, April 2011.

[44] R. M. Pulselli, P. Romano, C. Ratti, and E. Tiezzi, "Computing urban mobile landscapes through monitoring population density based on cell-phone chatting," vol. 3, pp. 121–134, 08 2008.

[45] C. L. Lee, W. S. Su, K. A. Tang, and W. I. Chao, "Design of handover self-optimization using big data analytics," in *The 16th Asia-Pacific Network Operations and Management Symposium*, Sept 2014, pp. 1–5.

[46] W. Lv, K. Liu, and T. Zhu, "Analyzing city blocks' properties based on mobile data." IEEE, November 2012, pp. 260–263.

[47] B. Fan, S. Leng, and K. Yang, "A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks," *IEEE Network*, vol. 30, no. 1, pp. 6–10, January 2016.

[48] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Pervasive Computing*, K. Lyons, J. Hightower, and E. M. Huang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 133–151.

[49] K. Laasonen, "Clustering and prediction of mobile user routes from cellular data," in *Knowledge Discovery in Databases: PKDD 2005*, A. M. Jorge, L. Torgo, P. Brazdil, R. Camacho, and J. Gama, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 569–576.

[50] B. Fan, S. Leng, and K. Yang, "A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks," *IEEE Network*, vol. 30, no. 1, pp. 6–10, January 2016.

[51] W. Xue, H. Zhang, Y. Li, D. Liang, and M. Peng, "Cell outage detection and compensation in two-tier heterogeneous networks," *International Journal of Antennas and Propagation*, vol. 2014, no. 4, 2014.

[52] P. S. Prasad and P. Agrawal, "Movement prediction in wireless networks using mobility traces," in *2010 7th IEEE Consumer Communications and Networking Conference*, Jan 2010, pp. 1–5.

[53] W. Su, S. J. Lee, and M. Gerla, "Mobility prediction in wireless networks," in *MILCOM 2000 Proceedings. 21st Century Military Communications. Architectures and Technologies for Information Superiority (Cat. No.00CH37155)*, vol. 1, 2000, pp. 491–495 vol.1.

[54] 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)*, June 2015, pp. 3939–3944.

[55] Y. Wang and S. Ram, "Predicting location-based sequential purchasing events by using spatial, temporal, and social patterns," *IEEE Intelligent Systems*, vol. 30, no. 3, pp. 10–17, May 2015.

[56] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo, "Analysis of gsm calls data for understanding user mobility behavior," in *2013 IEEE International Conference on Big Data*, Oct 2013, pp. 550–555.

[57] H. Kanasugi, Y. Sekimoto, M. Kurokawa, T. Watanabe, S. Muramatsu, and R. Shibasaki, "Spatiotemporal route estimation consistent with human mobility using cellular network data," in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, March 2013, pp. 267–272.

[58] F. Girardin, F. Calabrese, F. Dal Fiorre, A. Biderman, C. Ratti, and J. Blat, "Uncovering the presence and movements of tourists from user-generated content." Citeseer, 2008.

[59] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti," *PLOS Medicine*, vol. 8, no. 8, pp. 1–9, 08 2011. [Online]. Available: https://doi.org/10.1371/journal.pmed.1001083

[60] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "Route classification using cellular handoff patterns," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, ser. UbiComp '11. New York, NY, USA: ACM, 2011, pp. 123–132. [Online]. Available: http://doi.acm.org/10.1145/2030112.2030130

[61] Y. Ishikawa, Y. Tsukamoto, and H. Kitagawa, "H.: Extracting mobility statistics from indexed spatio-temporal datasets," *In: STDBM*, vol. 2004, pp. 9–16, 2004.

[62] Y. Qiao, X. Zhao, J. Yang, and J. Liu, "Mobile big-data-driven rating framework: measuring the relationship between human mobility and app usage behavior," *IEEE Network*, vol. 30, no. 3, pp. 14–21, May 2016.

[63] L. Jun, L. Tingting, C. Gang, Y. Hua, and L. Zhenming, "Mining and modelling the dynamic patterns of service providers in cellular data network based on big data analysis," *China Communications*, vol. 10, no. 12, pp. 25–36, Dec 2013.

[64] L. Qian, B. Wu, R. Zhang, W. Zhang, and M. Luo, "Characterization of 3g data-plane traffic and application towards centralized control and management for software defined networking," in *2013 IEEE International Congress on Big Data*, June 2013, pp. 278–285.

[65] M. Ahmed, A. Anwar, A. N. Mahmood, Z. Shah, and M. J. Maher, "An investigation of performance analysis of anomaly detection techniques for big data in scada systems," *EAI Endorsed Trans. Indust. Netw. & Intellig. Syst.*, vol. 2, p. e5, 2015.

[66] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, May 2000. [Online]. Available: http://doi.acm.org/10.1145/335191.335388

[67] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Loop: Local outlier probabilities," in *Proceedings of the 18th ACM Conference on Information and*

Knowledge Management, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 1649–1652. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646195

[68] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 124–161, 2015.

[69] J. Clement. Percentage of mobile device website traffic worldwide from 1st quarter 2015 to 1st quarter 2020. [Online]. Available: https://www.statista.com/statistics/277125/share-of-website-traffic-coming-from-mobile-devices/

[70] D. Gutierrez-Estevez, Ö. Bulakci, M. Ericson, A. Prasad, E. Pateromichelakis, J. Belschner, P. Arnold, and G. Calochira, "Ran enablers for 5g radio resource management," in *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 2017, pp. 1–6.

[71] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5g cellular networks: challenges, solutions, and future directions," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 86–92, 2014.

[72] J. P. Romero, O. Sallent, R. Agusti, and M. A. Diaz-Guerra, *Radio resource management strategies in UMTS*. John Wiley & Sons, 2005.

[73] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5g: how to empower son with big data for enabling 5g," *IEEE network*, vol. 28, no. 6, pp. 27–33, 2014.

[74] M. F. Iqbal and L. K. John, "Power and performance analysis of network traffic prediction techniques," in *2012 IEEE International Symposium on Performance Analysis of Systems & Software*. IEEE, 2012, pp. 112–113.

[75] 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, 2017.

[76] T. Italia, "Telecommunications - SMS, Call, Internet - MI," 2015, accessed: 2024-12-02. [Online]. Available: https://doi.org/10.7910/DVN/EGZHFV

[77] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ data science*, vol. 4, no. 1, p. 10, 2015.

[78] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer, "A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1790–1821, 2017.

[79] Z. Zhang, F. Liu, Z. Zeng, and W. Zhao, "A traffic prediction algorithm based on bayesian spatio-temporal model in cellular network," in *2017 International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 2017, pp. 43–48.

[80] P. Yu, W. Jian-min, and P. Xi-yuan, "Traffic prediction with reservoir computing for mobile networks," in *2009 Fifth International Conference on Natural Computation*, vol. 2. IEEE, 2009, pp. 464–468.

[81] S. Wang, J. Guo, Q. Liu, and X. Peng, "On-line traffic forecasting of mobile communication system," in *2010 First International Conference on Pervasive Computing, Signal Processing and Applications*. IEEE, 2010, pp. 97–100.

[82] A. Y. Nikravesh, S. A. Ajila, C.-H. Lung, and W. Ding, "Mobile network traffic prediction using mlp, mlpwd, and svm," in *2016 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2016, pp. 402–409.

[83] C.-W. Huang, C.-T. Chiang, and Q. Li, "A study of deep learning networks on mobile traffic forecasting," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2017, pp. 1–6.

[84] Gowrishankar and P. Satyanarayana, "Neural network based traffic prediction for wireless data networks," *International Journal of Computational Intelligence Systems*, vol. 1, no. 4, pp. 379–389, 2008.

[85] Y. Zhang, W. Wang, S. Zhang, D. Fan, B. Kou, and W. Zhou, "Svr based voice traffic prediction incorporating impact from neighboring cells," in *2016 19th International Symposium on Wireless Personal Multimedia Communications (WPMC)*. IEEE, 2016, pp. 142–146.

[86] H. Zhang and L. Dai, "Mobility prediction: A survey on state-of-the-art schemes and future applications," *IEEE Access*, vol. 7, pp. 802–822, 2019.

[87] H. Tabassum, M. Salehi, and E. Hossain, "Mobility-aware analysis of 5G and B5G cellular networks: A tutorial," *arXiv preprint arXiv:1805.02719*, 2018.

[88] M. Manalastas, H. Farooq, S. M. A. 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)*. IEEE, 2020, pp. 1–6.

[89] Y. Guan, Y. Xiao, H. Feng, C.-C. Shen, and L. J. Cimini, “MobiCacher: Mobility-aware content caching in small-cell networks,” in *2014 IEEE Global Communications Conference*. IEEE, dec 2014.

[90] M. Borkowski, O. Skarlat, S. Schulte, and S. Dustdar, “Prediction-based prefetch scheduling in mobile service applications,” in *2016 IEEE International Conference on Mobile Services (MS)*. IEEE, jun 2016.

[91] 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.

[92] R. Arshad, H. Elsawy, S. Sorour, T. Y. Al-Naffouri, and M.-S. Alouini, “Handover management in 5G and beyond: A topology aware skipping approach,” *IEEE Access*, vol. 4, pp. 9073–9081, 2016.

[93] P. V. Klaine, J. P. B. Nadas, R. D. Souza, and M. A. Imran, “Distributed drone base station positioning for emergency cellular networks using reinforcement learning,” *Cognitive Computation*, vol. 10, no. 5, pp. 790–804, Oct 2018. [Online]. Available: https://doi.org/10.1007/s12559-018-9559-8

[94] Y. Zhao, Z. Hong, Y. Luo, G. Wang, and L. Pu, “Prediction-based spectrum management in cognitive radio networks,” *IEEE Systems Journal*, vol. 12, no. 4, pp. 3303–3314, Dec 2018.

[95] M. Ozturk, M. Akram, S. Hussain, and M. A. Imran, “Novel QoS-aware proactive spectrum access techniques for cognitive radio using machine learning,” *IEEE Access*, pp. 1–1, 2019.

[96] C. Zhou, H. Lu, Y. Xiang, J. Wu, and F. Wang, “Geohashtile: Vector geographic data display method based on geohash,” *ISPRS International Journal of Geo-Information*, vol. 9, no. 7, p. 418, 2020.

[97] K. Yu and C. Qu, “Prediction and optimization of sharing bikes queuing model in grid of geohash coding,” *Measurement and Control*, p. 0020294019877521, 2020.

[98] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li, "Geolife gps trajectory dataset-user guide," *Microsoft Research. Available online at: https://www. microsoft. com/enus/research/publication/geolife-gps-trajectory-dataset-user-guide*, 2011.

[99] D. M. Gutierrez-Estevez, M. Gramaglia, A. de Domenico, N. di Pietro, S. Khatibi, K. Shah, D. Tsolkas, P. Arnold, and P. Serrano, "The path towards resource elasticity for 5G network architecture," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, April 2018, pp. 214–219.

[100] P. Yu, F. Zhou, T. Zhang, W. L. L. Feng, and X. Qiu, "Self-organized cell outage detection architecture and approach for 5G H-CRAN," *Wireless Communications and Mobile Computing*, vol. 2018, no. Article ID 6201386, p. 11, 2018.

[101] J. Moysen, L. Giupponi, and J. Mangues-Bafalluy, "A mobile network planning tool based on data analytics," *Mobile Information Systems*, vol. 2017, no. Article ID 6740585, p. 16, 2017.

[102] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "A SON solution for sleeping cell detection using low-dimensional embedding of mdt measurements," in *Personal, Indoor, and Mobile Radio Communication (PIMRC), 2014 IEEE 25th Annual International Symposium on*. IEEE, 2014, pp. 1626–1630.

[103] F. Pervez, M. Jaber, J. Qadir, S. Younis, and M. A. Imran, "Memory-based user-centric backhaul-aware user cell association scheme," *IEEE Access*, vol. 6, pp. 39 595–39 605, 2018.

[104] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Machine learning for wireless networks with artificial intelligence: a tutorial on neural networks," *arXiv preprint arXiv:1710.02913v1*, 2017.

[105] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: key techniques and open issues," *arXiv preprint arXiv:1809.08707v1*, 2018.

[106] 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, Nov 2018.

[107] F. Castanedo, G. Valverde, J. Zaratiegui, and A. Vazquez, “Using deep learning to predict customer churn in a mobile telecommunication network,” *http://wiseathena.com*, 2016.

[108] A. Gómez-Andrades, R. Barco, P. Muñoz, and I. Serrano, “Data analytics for diagnosing the RF condition in self-organizing networks,” *IEEE Transactions on Mobile Computing*, vol. 16, no. 6, pp. 1587–1600, June 2017.

[109] P. Yu, F. Zhou, T. Zhang, W. L. L. Feng, and X. Qiu, “Failure prediction using machine learning and time series in optical network,” *OPTICS EXPRESS 18553*, vol. 25, no. 16, August 2017.

[110] P. Spanoudes and T. Nguyen, “Deep learning in customer churn prediction: unsupervised feature learning on abstract company independent feature vectors,” *arXiv preprint arXiv:1703.03869v1*, 2017.

[111] A. K. Jain, “Data clustering: 50 years beyond K-means,” in *Machine learning and knowledge discovery in databases*. Springer, 2008, pp. 3–4.

[112] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, “Density-based clustering,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[113] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: analysis and an algorithm,” in *Advances in neural information processing systems*, 2002, pp. 849–856.

[114] C. Biernacki, G. Celeux, and G. Govaert, “Assessing a mixture model for clustering with the integrated classification likelihood,” Ph.D. dissertation, INRIA, 1998.

[115] C. Ding and X. He, “K-means clustering via principal component analysis,” in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 29.

[116] R. J. Campello, “Generalized external indexes for comparing data partitions with overlapping categories,” *Pattern Recognition Letters*, vol. 31, no. 9, pp. 966–975, 2010.

[117] M. Hassani and T. Seidl, “Using internal evaluation measures to validate the quality of diverse stream clustering algorithms,” *Vietnam Journal of Computer Science*, vol. 4, no. 3, pp. 171–183, 2017.

[118] B. Scholkopf, K.-K. Sung, C. J. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with Gaussian kernels to radial basis function classifiers," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2758–2765, 1997.

[119] J. Ling, D. Chizhik, and R. Valenzuela, "On resource allocation in dense femto-deployments," in *2009 IEEE International Conference on Microwaves, Communications, Antennas and Electronics Systems*, Nov 2009, pp. 1–6.

[120] F. Bernardo, R. Agusti, J. Cordero, and C. Crespo, "Self-optimization of spectrum assignment and transmission power in OFDMA femtocell," in *Proc. Sixth Advanced Int Telecommunications Conf(AICT)*, 2010, pp. 404–409.

[121] A. Saeed, E. Katranaras, M. Dianati, and M. A. Imran, "Control and data channel resource allocation in macro-femto heterogeneous networks," in *2015 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug 2015, pp. 1272–1276.

[122] J. D. Hobby and H. Claussen, "Deployment options for femtocells and their impact on existing macrocellular networks," *Bell Labs Technical Journal*, vol. 13, no. 4, pp. 145–160, 2009. [Online]. Available: http://dx.doi.org/10.1002/bltj.20341

[123] X. Chu, Y. Wu, D. Lopez-Perez, and X. Tao, "On providing downlink services in collocated spectrum-sharing macro and femto networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 4306–4315, December 2011.

[124] X. Chen, H. Xia, Z. Zeng, S. Wu, W. Zuo, and Y. Lu, "Energy-efficient heterogeneous networks for green communications by inter-layer interference coordination," in *2014 International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Sept 2014, pp. 70–74.

[125] K. Davaslioglu, C. C. Coskun, and E. Ayanoglu, "Energy-efficient resource allocation for fractional frequency reuse in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 10, pp. 5484–5497, Oct 2015.

[126] A. Y. Al-Zahrani and F. R. Yu, "An energy-efficient resource allocation and interference management scheme in green heterogeneous networks using game theory," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5384–5396, July 2016.

[127] C. Kosta, B. Hunt, A. U. Quddus, and R. Tafazolli, "Distributed energy-efficient inter-cell interference coordination (icic) in multi-cell hetnets," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5.

[128] C. Yang, J. Li, A. Anpalagan, and M. Guizani, "Joint power coordination for spectral-and-energy efficiency in heterogeneous small cell networks: A bargaining game-theoretic perspective," *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1364–1376, Feb 2016.

[129] R. Ramamonjison and V. K. Bhargava, "Energy allocation and cooperation for energy-efficient wireless two-tier networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6434–6448, Sept 2016.

[130] R. Litjens and L. Jorguseski, "Potential of energy-oriented network optimisation: Switching off over-capacity in off-peak hours," in *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sept. 2010, pp. 1660 –1664.

[131] M. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *IEEE International Conference on Communications Workshops, ICC*, Jun. 2009, pp. 1 –5.

[132] I. Ashraf, F. Boccardi, and L. Ho, "Sleep mode techniques for small cell deployments," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 72–79, August 2011.

[133] Z. Niu, "Tango: traffic-aware network planning and green operation," *IEEE Wireless Communications*, vol. 18, no. 5, pp. 25–29, October 2011.

[134] F. Z. Kaddour, E. Vivier, L. Mroueh, M. Pischella, and P. Martins, "Green opportunistic and efficient resource block allocation algorithm for lte uplink networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4537–4550, Oct 2015.

[135] E. Katranaras, M. A. Imran, M. Dianati, and R. Tafazolli, "Green inter-cluster interference management in uplink of multi-cell processing systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 12, pp. 6580–6592, Dec 2014.

[136] E. Katranaras, M. A. Imran, and R. Hoshyar, "Energy aware transmission in cellular uplink with clustered base station cooperation," in *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–5.

[137] S. Zhang, D. Yin, Y. Zhang, and W. Zhou, "Computing on base station behavior using erlang measurement and call detail record," *IEEE transactions on emerging topics in computing*, vol. 3, no. 3, pp. 444–453, 2015.

[138] P. Di Francesco, F. Malandrino, and L. A. DaSilva, "Assembling and using a cellular dataset for mobile network analysis and planning," *IEEE Transactions on Big Data*, vol. 4, no. 4, pp. 614–620, 2017.