Universidad de Málaga

Escuela Técnica Superior de Ingeniería de Telecomunicación

Programa de Doctorado en Ingeniería de Telecomunicación

UNIVERSIDAD DE MÁLAGA

TESIS DOCTORAL

Next-Generation Self-Organizing Networks through a Machine Learning Approach

Autor:

David Palacios Campos

Directores:

Raquel Barco Moreno

Isabel de la Bandera Cascales

2018

AUTOR: David Palacios Campos

http://orcid.org/0000-0002-4898-3427

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

Por la presente, **Dra. Dª. Raquel Barco Moreno y Dra. Dª. Isabel de la Bandera Cascales**, profesores doctores del *Departamento de Ingeniería de Comunicaciones* de la Universidad de Málaga,

CERTIFICAN:

Que **D. David Palacios Campos**, Ingeniero de Telecomunicación, ha realizado en el Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga bajo su dirección, el trabajo de investigación correspondiente a su TESIS DOCTORAL titulada:

"Next-Generation Self-Organizing Networks through a Machine Learning Approach"

En dicho trabajo se han propuesto aportaciones originales para la gestión de redes móviles. En particular, se han propuesto métodos para la mejora de las tareas de diagnosis y de optimización automáticas de la red, dentro del marco de las redes autoorganizadas de nueva generación. Los resultados expuestos han dado lugar a publicaciones en revistas, patentes y aportaciones a congresos.

Por todo ello, consideran que esta Tesis es apta para su presentación al Tribunal que ha de juzgarla. Y para que conste a efectos de lo establecido, AUTORIZAN la presentación de esta Tesis en la Universidad de Málaga.

En Málaga, a _____ de _____ de _____

Fdo: Raquel Barco Moreno, Isabel de la Bandera Cascales

UNIVERSIDAD DE MÁLAGA
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

Reunido el tribunal examinador en el día de la fecha, constituido por:

Presidente: Dr. D. _____

Secretario: Dr. D. _____

Vocal: Dr. D. _____

para juzgar la Tesis Doctoral titulada *Next-Generation Self-Organizing Networks through a Machine Learning Approach* realizada por D. David Palacios Campos y dirigida por la Dra. D^a. Raquel Barco Moreno y la Dra. D^a. Isabel de la Bandera Cascales, acordó por

_____ otorgar la calificación de

_____ y para que conste,

se extiende firmada por los componentes del tribunal la presente diligencia.

Málaga a _____ de _____ de _____

El Presidente: El Secretario: El Vocal:

Fdo.: _____ Fdo.: _____ Fdo.: _____

A mis padres.

Acknowledgements

To some extent, a PhD slightly resembles the *hero's journey*; the story of a character who leaves his ordinary world in search of adventure, and after undergoing an ordeal, returns home transformed, having gained a great wisdom. However, no *hero's journey* is possible without the figures of the *hero*'s mentor, fellows, friends and family, which I would like to thank here.

First of all, I would like to sincerely thank my supervisors (my mentors) Raquel and Isabel, two of the most hard-working people that I have met and without who this journey would have not been possible. Thank you Raquel for giving me the chance to be part of this group, which has allowed me to learn and become the researcher I currently am. Thank you very much for your support and for trusting me. Isabel, I would like to thank you for the time and effort that you have spent on my guidance, for your encouragement, and for the discussions we have had, which have made me grow as a better researcher.

Together with these, I would like to thank my lab mates, my fellows. The people which filled every day with laughter, good times and enthusiasm for learning. Among them, I would like to specially thank Emil, fellow sufferer in several occasions and the person who passed his eagerness for machine learning on me, and Ana, for always being keen to help in whatever she could.

Beyond the border, I would like to appreciate the kindness, the hard-working attitude and the overwhelming knowledge that I found in the people working at Nokia-Bell Labs in Aalborg. I would like to thank Beatriz and István for guiding Emil and me in our struggle with machine-type communications and the *ns-3* simulator. I would also like to acknowledge Daniela for her always sensible and wise advices. Together with these, I strongly appreciate the effort of Lucas and the rest of the Wireless Communications Networks team, for making my stay in Aalborg not only be a research milestone, but a life experience.

At this point, I acknowledge the people from Ericsson, which, in the first stretch of this thesis, contributed with their experience, being this my first contact point with the management of cellular networks from the point of view of the industry.

I must also acknowledge the financial support given by the projects mentioned below, together with the research group TIC-102 Ingeniería de Comunicaciones and Universidad de

Málaga. They made this work possible and allowed me to present results and exchange knowledge and skills in conferences and journals.

I could not conclude the acknowledgments section without thanking my friends in Málaga for their continuous support and our endless and always enriching talks, which have helped me acquire a more critical thinking. Special thanks to Manuel for his constant and sincere encouragement, and also Pedro, Arturo and Fran, for always being ready to spend an evening of laughter after work. I also want to thank my friends in Jaén for having accompanied me through all these years: Juande, David, Javi, Dani and Alberto.

I would like to save a special place to thank my family for their endless support and love. Thanks to my sister, Paloma, for her affection and trust through all this time, and thanks to my parents, for having been there in every single step and for their motivation and understanding along this hard journey. This would not have been possible without them, and I will always be grateful to them.

This thesis was partially funded by the following projects:

- Gestión integral avanzada de funciones SON (Self-Organizing Networks) para redes móviles futuras, P12-TIC-2905, Proyectos de Execelencia, Junta de Andalucía.
- 8.06/5.59.3722, contract with Optimi-Ericsson, with support from the Junta de Andalucía (Agencia IDEA, Consejería de Ciencia, Innovación y Empresa) and ERFD.
- ONE5G: E2E-aware Optimizations and advancements for the Network Edge of 5G New Radio, funded under H2020-ICT-2016-2, project number: 760809.
- MONROE: Measuring Mobile Broadband Networks in Europe, funded under: H2020-ICT-11-2014, project number: 644399.

Contents

Abstract

At the present time, for most of the population, mobile phones have become the instruments through which to interact with the surrounding world. From its original service of voice transmission, cellular communications have evolved to provide a variety of services that could be hardly imagined just four decades ago. From that starting point, cellular networks were first enhanced to support data transmission, which opened the door to services like videocalls or web surfing. Later on, successive improvements were made so as to reach resource- and energy-efficient networks, while enhancing the users' perceived quality of experience (QoE) by means of an increasingly higher performance. In order to consequently reduce the management costs in such scenario, some further improvements needed to be made. This led to the concept of self-organizing networks (SONs). That is, the automation of the management tasks of a cellular network to reduce the operational and capital expenditure (OPEX and CAPEX, respectively).

SON tasks are divided in three categories: self-configuration, self-optimization and self-healing. Self-configuration aims at automating the actions required when a network is to be deployed, like the initial configuration parameter setting. Self-optimization tasks pursue maximizing the efficiency of the networks in a time-varying environment, which takes shape as a variety of mechanisms addressing mobility, accessibility and integrity issues. Finally, the targets of self-healing are identifying and repairing possible failures that may arise while the network is operated.

Thus, one of the main tasks of self-healing is determining the cause of a failure, which is called root cause analysis (RCA). Tools for RCA are automatic systems which, in the shape of classification systems, aim at determining a class (or network state) regarding a set of features (or key performance indicators, KPIs). Although different mechanisms have been proposed until now as tools for RCA, there is a long way to develop accurate systems that can deal with the large amount of performance information that is normally collected in a cellular network.

Together with self-healing, self-optimization appears as the SON function group that attracts the most attention from industry and academia. This is mainly due to the optimization opportunities that novel functionalities from the upcoming networks bring. In particular, the management of multi-link connections, and within these, multi-connectivity (MC), occupies an eminent place in the next generation of mobile communications: the Fifth Generation New Radio (5G NR). However, given its novelty, multi-link communications currently lack of efficient

management mechanisms, which will be one of the research hot topics in the coming years.

The objective of this thesis is the improvement of SON functions through the development and use of machine learning (ML) tools for the network management. In particular, its target is twofold. On the hand, self-healing is addressed through the proposal of a novel tool for RCA, which takes the shape of a combination of multiple RCA baseline systems to develop an enhanced ensemble-based system. In order to further enhance the RCA accuracy while lowering both the CAPEX and OPEX, ML techniques for dimensionality reduction are proposed and assessed in combination with RCA tools. On the other hand, multi-link functionalities within self-optimization are studied, and techniques for automatic link management are proposed. In the field of enhanced mobile broadband (eMBB) communications, a component carrier manager implementing network operators' policies is proposed, whereas in the field of low-latency vehicular communications, a mechanism for multi-path traffic steering is proposed.

Many of the methods proposed in this thesis have been assessed using data from live cellular networks, which has allowed them to demonstrate both their validity in realistic environments and their ability to be deployed in current and forthcoming cellular networks.

Resumen

A día de hoy, para la mayoría de la población, los teléfonos móviles se han convertido en el principal instrumento a través del cual interactuar con el mundo. Desde su servicio original de transmisión de voz, las comunicaciones móviles han evolucionado hasta proporcionar una variedad de servicios que apenas podían ser imaginados hace cuatro décadas. Desde ese punto inicial, en primer lugar, las redes celulares se mejoraron para soportar la transmisión de datos, lo que abrió la puerta a servicios como las videollamadas o la navegación web. Las siguientes mejoras persiguieron hacer eficientes a las redes en términos de consumo de energía y uso de recursos radio, a la vez que se mejoraba la calidad de experiencia percibida por los usuarios. Con el objetivo de, simultáneamente, reducir los costes de gestión de las redes, surgió el concepto de las redes autoorganizadas, o *self-organizing networks* (SON). Es decir, la automatización de las tareas de gestión de una red celular para disminuir los costes de infraestructura (*capital expenditure*, CAPEX) y de operación (*operational expenditure*, OPEX).

Las tareas de las SON se dividen en tres categorías: autoconfiguración, autooptimización y autocuración. La autoconfiguración tiene como objetivo automatizar las tareas requeridas cuando se despliega una nueva red, como la configuración inicial de sus parámetros de operación. Las tareas de autooptimización, por otro lado, persiguen maximizar la eficiencia de las redes en un entorno variable en el tiempo, lo que se traduce en múltiples mecanismos de optimización, que abordan desde problemas de movilidad y accesibilidad a cuestiones como la integridad de la comunicación. Finalmente, la autocuración se centra en la identificación y reparación de posibles fallos que aparezcan en la red durante su fase de operación.

Una de las tareas fundamentales de la autocuración es determinar la causa de un fallo, tarea también conocida *root cause analysis* (RCA). Las herramientas de RCA son sistemas automáticos que, en forma de sistemas de clasificación, buscan determinar una clase (o estado de la red) atendiendo a un conjunto de características (o indicadores clave de rendimiento, KPIs, por sus siglas en inglés, *key performance indicators*) que describen una observación o muestra. Aunque actualmente son varios los sistemas que se han propuesto como herramientas para RCA, aún hay un largo camino para el desarrollo de sistemas de RCA precisos, que puedan lidiar con la gran cantidad de información de rendimiento que normalmente se recoge en una red celular.

Junto con la autocuración, la autooptimización aparece como el grupo de funciones SON

que más atención atrae desde los ámbitos industrial y académico. Esto se debe principalmente a las oportunidades de optimización que las nuevas funcionalidades de las futuras redes brindan. En particular, en la nueva generación de las comunicaciones móviles (la quinta generación o *Fifth-Generation New Radio*, 5G NR) cabe destacar la gestión de conexiones multienlace, y dentro de éstas, la multiconectividad (*multi-connectivity*, MC). Dada su novedad, sin embargo, las comunicaciones multienlace actualmente carecen de mecanismos de gestión eficientes, siendo éste uno de los temas de investigación más candentes en los próximos años.

El objetivo de esta tesis es la mejora de las funciones SON a través del desarrollo y uso de herramientas de aprendizaje automático (*machine learning*, ML) para la gestión de la red. En concreto, el objetivo de esta tesis es doble. Por un lado, se aborda la autocuración a través de la propuesta de una novedosa herramienta para RCA, consistente en la combinación de múltiples sistemas RCA independientes para el desarrollo de un sistema compuesto de RCA mejorado. A su vez, para aumentar la precisión de las herramientas de RCA mientras se reducen tanto el CAPEX como el OPEX, en esta tesis se proponen y evalúan herramientas de ML de reducción de dimensionalidad en combinación con herramientas de RCA.

Por otro lado, en esta tesis se estudian las funcionalidades multienlace dentro de la autocuración y se proponen técnicas para su gestión automática. En el campo de las comunicaciones mejoradas de banda ancha (*enhanced mobile broadband*, eMBB), se propone una herramienta para la gestión de portadoras, que permite la implementación de políticas del operador, mientras que en el campo de las comunicaciones vehiculares de baja latencia, se propone un mecanismo multi-camino para la redirección del tráfico a través de múltiples interfaces radio.

Muchos de los métodos propuestos en esta tesis se han evaluado usando datos provenientes de redes celulares reales, lo que ha permitido demostrar su validez en entornos realistas, así como su capacidad para ser desplegados en redes móviles actuales y futuras.

Acronyms

3G	3rd generation
3GPP	3rd Generation Partnership Project
5G	5th generation
5GC	5G core network
AMF	Access and mobility management function
AMISE	Asymptotic mean integrated squared error
ANR	Automatic neighbor relation
AP	Access point
ARFCN	Absolute radio frequency channel number
ARQ	Automatic repeat request
AS	Access stratum
BBU	Baseband unit
BLER	Block error rate
BPSK	Binary phase-shift keying
BS	Base station
BSC	Base station controller
CA	Carrier aggregation
CAPEX	Capital expenditure
CBR	Case-based reasoning

CC	Component carrier
CCM	Component carrier manager
CDF	Cumulative distribution function
CDMA	Code division multiple access
CEM	Customer experience management
CMC	Connection mobility control
CN	Core network
CPU	Central processing unit
CQI	Channel quality indicator
C-RAN	Centralized RAN
C-RNTI	Common radio network temporary identifier
CS	Circuit-switched
CSFB	Circuit-switched fallback
CSMA	Carrier sense multiple access
DC	Dual connectivity
DER	Diagnosis error rate
DL	Downlink
DL-SCH	Downlink shared channel
D-RAN	Distributed RAN
DRB	Data radio bearer
DSRC	Dedicated short-range communications
E2E	End-to-end
eCNS	Evolved core network solution
eMBB	Enhanced mobile broadband
eNB	Evolved node B
EPC	Evolved packet core
EPS	Evolved Packet System

E-RAB	E-UTRAN radio access bearer
ETSIT	Escuela Técnica de Ingeniería de Telecomunicación
ETU	Extended typical urban
E-UTRA	Evolved UMTS terrestrial radio access
E-UTRAN	Evolved UMTS terrestrial RAN
FDD	Frequency division duplex
FLC	Fuzzy logic controller
FM	Fault management
FNR	False negative rate
FPR	False positive rate
FR	Proposed framework
FTP	File transfer protocol
GEV	Generalized extreme value
gNB	Next-generation node B
GPRS	General Packet Radio Service
HARQ	Hybrid ARQ
HO	Handover
HOSR	Handover success rate
HPBW	Half-power beam width
HSS	Home subscriber server
ICA	Independent component analysis
ICIC	Inter-cell interference coordination
IEEE	Institute of Electrical and Electronics Engineers
IMEI	International mobile station equipment identity
IMSI	International mobile subscriber identity
IP	Internet protocol
IRAT	Inter-RAT

ISDN	Integrated services digital network
ITS	Intelligent transportation systems
KDE	Kernel density estimation
kNN	k-nearest neighbors
kPCA	Kernel PCA
KPI	Key performance indicator
LAU	Location area update
LDA	Linear discriminant analysis
LLE	Locally-linear embedding
LOOCV	Leave-one-out cross-validation
LS	Laplacian score
LTE	Long-Term Evolution
LTE-A	Long-Term Evolution Advanced
MAC	Medium access control
MC	Multi-cluster feature selection technique
MC	Multi-connectivity
MCCV	Monte Carlo cross-validation
MCS	Modulation and coding scheme
MDT	Minimization of drive tests
MeNB	Master eNB
ML	Machine learning
MME	Mobility management entity
mMTC	Massive machine-type communications
MN	Master node
MNO	Mobile network operator
MOS	Mobile-originated signaling
MOS	Mean opinion score

MR-DC	Multi-RAT dual connectivity
MRO	Mobility robustness optimization
MSC	Mobile switching center
NAS	Non-access stratum
NCFS	Neighborhood component feature selection
ng-eNB	Next-generation evolved node B
NG-RAN	Next-generation RAN
NG-SON	Next-generation self-organizing networks
NN	Nearest neighbors
NR	New radio
OER	Overall error rate
OFDM	Orthogonal frequency division multiplexing
OFDMA	Orthogonal frequency division multiple access
OPEX	Operational expenditure
OSS	Operations support systems
OTT	Over-the-top
OV	Overlap index-based technique for feature selection
OVL	Overlapping area
PAPR	Peak-to-average power ratio
PCA	Principal component analysis
PCell	Primary cell
PCI	Physical cell identity
PCRF	Policy and charging rules function
PDCP	Packet data convergence protocol
PDF	Probability density function
PDN	Packet data network
PDU	Packet data unit

P-GW	PDN gateway
PLMN	Public land mobile network
PRB	Physical resource block
PS	Packet-switched
PSCell	Primary secondary cell
PUCCH	Physical uplink control channel
QAM	Quadrature amplitude modulation
QoS	Quality of service
QPSK	Quadrature phase-shift keying
RAC	Radio admission control
RACH	Random access channel
RAN	Radio access network
RAT	Radio access technology
RAU	Routing area update
RBC	Radio bearer control
RCA	Root cause analysis
RE	Resource element
RL	ReliefF algorithm for feature selection
RNC	Radio network controller
ROHC	Robust header compression
RRC	Radio resource control
RRM	Radio resource management
RSRP	Reference signal received quality
RSRQ	Reference signal received quality
RSSI	Received signal strength indicator
RSU	Roadside unit
RTT	Round-trip time

SCell	Secondary cell
SC-FDMA	Single-carrier frequency division multiple access
SDAP	Service data application protocol
SDU	Service data unit
SE	Spectral embedding
SeNB	Secondary eNB
S-GW	Serving gateway
SINR	Signal-to-interference-plus-noise ratio
SISO	Single input single output
SL	Sidelink
SMF	Session management function
SN	Secondary node
SOM	Self-organizing maps
SQ	Sequential feature selection technique
SRB	Signaling radio bearer
SRVCC	Single-radio voice call continuity
TB	Transport block
TDD	Time division duplex
TE	Troubleshooting expert
TT	Trouble ticket
TTI	Time transmission interval
UDP	User datagram protocol
UE	User equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
UP	Unsupervised technique for feature selection
UPF	User plane function

URLLC	Ultra-reliable low-latency communications
V2I	Vehicle-to-infrastructure
V2N	Vehicle-to-network
V2P	Vehicle-to-pedestrian
V2V	Vehicle-to-vehicle
V2X	Vehicle-to-everything
VoLTE	Voice over LTE

Chapter 1

Introduction

This first chapter introduces the main topics of this thesis, presenting the motivation, the objectives and the document structure.

1.1 Motivation

Mobile communications, which have evolved through five generations in just four decades, have attracted a lot of attention from both the industry and the research community in recent years. The reason for this vertiginous growth lies in the fact that they have become an essential part of current life.

In their first twenty years, mobile communications were devoted to voice communication. Currently, however, they support a huge amount of services, among which web browsing, video broadcasting or gaming can be found. Together with these mobile broadband (MBB) services, the scope of mobile communications has widened in order to cover other scenarios, use cases and requirements. This is the case of massive machine-type communications (mMTC) or ultra-reliable low-latency communications (URLLC) [1]. As a result, currently, up to 95% of the world population live in an area covered by a mobile network [2].

The first generation of mobile networks was a region-specific and purely analog technology, and as such, lacked from privacy, reliability and efficiency. With the second generation of mobile communications came digitalization and standardization over several countries. The former brought reliability and efficiency to mobile networks; the latter brought an internationally seamless wireless network. The most widespread second-generation mobile network is Global System for Mobile communications (GSM), which was standardized by the European Telecommunications Standards Institute (ETSI) and allowed the transmission of voice and data through a circuit-switched cellular network. Later on, the upsurge of wired packet-switched networks (like the Internet) led to the development of its wireless counterpart: a twofold cellular network, capable of holding both circuit-switched (voice) and packet-switched data. That is, the General

Packet Radio Service (GPRS)/GSM network. Few years after, the users' demands for new services and further performance led to the third generation (3G) of mobile communications. 3G offered a higher capacity mostly through several enhancements of the radio interface, like the code division multiple access (CDMA). The main third-generation technology for mobile communications is Universal Mobile Telecommunications System (UMTS), which was standardized by the Third Generation Partnership Project (3GPP). Driven by the predominance of Internet protocol (IP)-based traffic, a full-IP cellular network was developed: the Long-Term Evolution (LTE) cellular network [3]. Its end-to-end (E2E) packet-switched nature allowed the network topology to be simplified, and its performance, enhanced. As a result of the emergence of novel types of communication (mMTC and URLLC) and the increase in the performance demands for traditional services (leading to service categories like the enhanced mobile broadband, eMBB) the Fifth-Generation New Radio (5G NR) standard has been recently developed [4]. The second phase of the standardization process of 5G has just started, in which tools and mechanisms for the management and optimization of the recently defined 5G NR will be developed.

The current mobile system is not made up by an isolated cellular network, but by a big amount of networks of different generations. Specifically, second-, third- and fourth-generation cellular networks coexist nowadays in commercial deployments. The particularities of each of these technologies, together with their mutual influence make their management tasks become cumbersome and expensive, increasing both the operational and the capital expenditure (OPEX and CAPEX, respectively). In order to prevent this, the Next-Generation Mobile Networks (NGMN) Alliance proposed the concept of self-organizing networks (SON) in 2008 [5, 6]. That is, the automation of some of the management tasks in cellular communications. Shortly after, 3GPP included the concept of SON as a key element for the management of LTE networks [7], which is expected to be even more relevant for 5G NR networks [8].

Within SON, three different categories for the automation of network management can be found: self-configuration [9], self-optimization [10] and self-healing [11]. The first stands for the group of functionalities devoted to the automation in the deployment of new networks or network elements. Plug-and-play functionalities or algorithms for self-planning can be found in this group [12]. On the other hand, self-optimization functions seek to maximize the network performance, which may be sub-optimal due to different time-varying issues. Examples of these are network-internal issues, like a load imbalance, or network-external issues, like a high interference level. In all these cases, the network configuration parameters should be readjusted in order to drive the network to an optimal working point. The amount of network functionalities for which a performance metric may be elicited make a high number of use cases for self-optimization appear. Capacity and coverage optimization (CCO) [13] and mobility load balancing (MLB) [14, 15] are just some of them. Finally, the aim of self-healing functions is to prevent the network performance (and, eventually, the users' perceived quality of experience, QoE) to be degraded. To that end, four functions are distinguished within self-healing [16]: network fault detection [17–19], fault diagnosis (also known as root cause analysis, RCA) [20–34], fault compensation [35, 36] and recovery.

The increasing complexity of cellular network management and the attempt of mobile network operators (MNO) to reduce their expenditure while providing a better QoE has made

SON become an interesting research topic. In the last ten years, several international research projects have taken place. Some of these are CELTIC Gandalf [37], E3 [38], SOCRATES [39], SELF-NET [40], UniverSelf [41], SEMAFOUR [42] and COMMUNE [43]. Most of them have been devoted to self-configuration and self-optimization. Despite self-healing has attracted less attention in the shape of international projects, some national projects, in which self-healing plays a major role, have provided a wide variety of research results [19, 24–36].

Contrary to self-configuration, which takes place before a cellular network is operated, self-optimization and self-healing functions are run during its operational phase. This makes these functions occupy a particularly relevant place within SON, as they will be run during most of the living time of the cellular network. This, together with the intricacy of nowadays and forthcoming cellular networks due to novel functionalities, use cases and service categories, makes current self-healing and self-optimization schemes insufficient, thus needing a pressing boost in their development. Regarding self-healing, for example, recent works on RCA (like [24–34]) do not consider the combination of several diagnosis models coming from multiple sources (like different troubleshooting experts), which would lead to a noticeable improvement in the diagnosis accuracy. Concerning self-optimization, novel functionalities like the multi-connectivity (MC) schemes expected in 5G NR make single-connectivity solutions (like [13–15, 44, 45]) need to be updated to take full advantage of the new possibilities offered.

Besides, for self-optimization and self-healing functions to know the current network state (i.e., whether this state is sub-optimal or degraded) they rely on network performance indicators, which in the context of cellular network management are called key performance indicators (KPI). These indicators quantify the performance of network processes and functionalities and are monitored and stored by the operations and support system (OSS) of the cellular network. As a result, and for network management tasks to be optimally performed, MNOs and vendors have put big efforts into arranging a sufficiently detailed and varied amount of KPIs. However, many of these KPIs do not provide relevant information about the network state. Such amount of data does not only pose a storage problem for the databases in the network, but often leads to the over-fitting of the algorithms devised for the automation of SON tasks. Therefore, the KPIs to be taken into account must be chosen carefully to develop an efficient SON system. Traditionally, this selection has been manually performed by troubleshooting experts [16–34]. However, due to the heavy time-consuming nature of this task, troubleshooting experts have tended to use a fix set of KPIs that, to their knowledge, has shown the best results in terms of errors metrics. As a consequence, the number and variety of KPIs that have been selected in this way often differ from those which could lead to an optimum performance and processing time of the SON algorithms.

Alongside this, and pushed by the advancement in the computation power of nowadays central processing units (CPU), in the last years an unprecedented progress in machine learning (ML) has been witnessed, whose main tasks are pattern classification and trend prediction. As of today, ML has spread to a huge number of areas, like computer vision [46, 47], econometric analysis [48] or manufacturing processes [49]. At this point, cellular network management is not an exception, and some steps have been taken towards including high performance ML techniques in this area. This is the case of [24], in which an advanced type of neural network is trained to

cluster cellular network measurements for diagnosis tasks, or [34], in which a decision tree is built over a variety of user-reported metrics to identify faulty cells. Nevertheless, automatic network management still can be deeply benefited from the newest advancements in ML. It is because of this, that this thesis aims at giving a step forward towards the full automation and optimization of SON functionalities by developing and integrating novel ML techniques into SON.

1.2 Research objectives

The main objective of this thesis is the enhancement of the cellular network management through the development and integration of novel ML tools. For this purpose, this thesis aims at improving the two most relevant function groups within SON: self-healing and self-optimization.

Figure 1.1 represents self-healing and self-optimization tasks together with the OSS, which is in charge of steadily monitoring the cellular network and storing network observations by means of KPIs. Such measurements can be labeled, by attaching a label standing for the network state under which the measurement was made, or unlabeled, by only storing the measurement itself. These databases make up the knowledge basis for the automatic management tasks, which may be supervised or unsupervised depending on whether they use labeled or unlabeled data. This thesis deals with the enhancement of the system shown in Figure 1.1 by developing and applying ML tools at three points: self-healing tasks, self-optimization tasks and the way that OSS performance information is monitored, processed and used in those SON tasks.

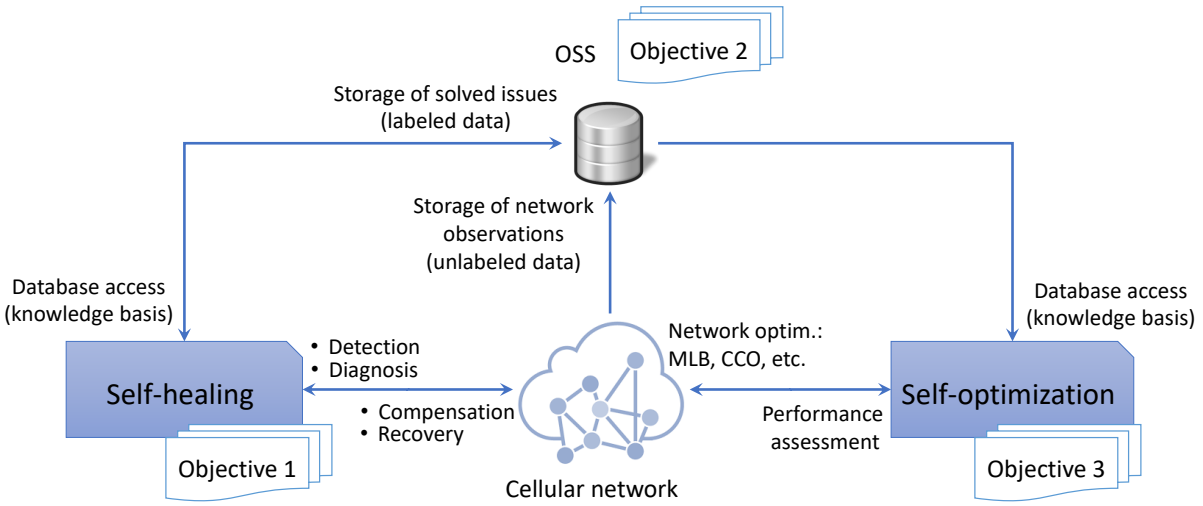


Figure 1.1: Diagram relating self-healing and self-optimization tasks and thesis objectives.

In particular, the research lines of this thesis can be summarized in the following objectives (see Figure 1.1):

- **Objective 1:** *Design of an enhanced tool for automatic diagnosis.* In practice, the selection of the diagnosis technique becomes cumbersome when the aim is to deploy an automatic diagnosis system in a real network. Furthermore, once the technique to be used has been decided, a diagnosis model needs to be built, either from troubleshooters' expertise or

from databases of historical cases. Often, there are different troubleshooting experts or case databases, which lead to building different diagnosis models with different diagnosis accuracies.

Instead of selecting a given technique and then a specific model, this objective pursues developing a framework to combine multiple techniques and/or models in order to overcome the limitations of single techniques/models and, thus, increase the diagnosis accuracy. In the domain of cellular networks, hybrid mixtures (i.e., combination of standalone systems of different kind) of diagnosis systems have not been previously proposed. The adopted approach is to formulate the diagnosis problem from a more general perspective, assuming it is a problem of classifying a set of cases which show some kind of patterns into a not always known number of classes or causes.

- **Objective 2:** *Development and integration of techniques for dimensionality reduction for self-healing.* The second objective of this thesis is to develop a variety of tools to provide a reduced set of KPIs which allows the monitoring and storage needs to be reduced, while improving the performance of SON functions without human intervention. To that end, techniques for dimensionality reduction will be assessed and applied in the field of self-healing tasks.

- **Objective 3:** *Development of algorithms for the performance enhancement of eMBB and low-latency vehicular traffic in 5G through multi-link management.* Nowadays cellular networks face a stage of vertiginous growth, in an attempt to jointly support a set of services, use cases and requirements, that up to now, could only be covered by a variety of wireless technologies [50]. The usage of different and possibly simultaneous connections between the user equipment (UE) and one or several network nodes arises as one of the possible solutions to address this disparity of requirements and services. The third objective of this thesis is to take advantage of this fact and enhance some of the performance metrics associated to eMBB and low-latency traffic in 5G. In particular, this objective is divided in two research lines, one for each of these traffic types:

  - **Objective 3.1** is to develop an algorithm to manage the assignment of component carriers (CC) provided by a number of 5G network nodes relying on the upcoming concept of MC. Under the eMBB scope, the target is to increase the UE throughput by making a proper CC assignment.

  - **Objective 3.2** covers low-latency traffic in a vehicle-to-everything (V2X) communications scenario. Under this scope, this research line aims to develop an algorithm to dynamically select the interface used by the UEs to send low-latency messages based on the evaluation of the performance information obtained from each of them.

1.3 Document structure

This thesis is divided into three main parts. Each part corresponds to the previous objectives, which will be treated independently. Consequently, each of them constitutes one chapter.

Specifically, this thesis consists of six chapters. Chapter 1 corresponds to this introduction, where the motivation and research objectives are presented. In Chapter 2, a technical background is outlined. First, the standards for current LTE networks and the upcoming 5G NR are briefly described to provide context to the rest of the thesis. Then, SONs and their main functions are described. Chapter 3 is devoted to **Objective 1** (Figure 1.1): the design of an enhanced tool for automatic diagnosis under the scope of self-healing. In Chapter 4, **Objective 2** is addressed, integrating dimensionality reduction as one of the main enablers towards the full automation of next-generation SONs. Chapter 5 gathers **Objective 3**: the development of algorithms for multi-link management in 5G networks. Finally, Chapter 6 summarizes the main conclusions of this work and presents future lines of action.

As appendix A, this work includes a brief summary of the thesis in Spanish. As appendix B, an intermediate work, related to self-healing (**Objective 1**) under a multi-radio access technology (RAT) environment (as a precedent to the multi-link management of **Objective 3**) is included. In parallel to the work described in this thesis, David Palacios actively participated in the deployment of the UMA LTE indoor network, which is described in appendix C.

Chapter 2

Technical background

This chapter presents the technical background required to follow the rest of this thesis, and is divided into two sections. First, Section 2.1 summarizes the main aspects of LTE and 5G NR networks, including the network architectures and the main functions at physical, link and network level. Next, Section 2.2 outlines the basic ideas of SON, both with its classic Release-9 conception and with the concept of next-generation self-organizing networks (NG-SONs).

2.1 Overview of current cellular networks

This section aims to provide an overview of the mobile technologies that have generated the most interest in recent years and are the basis on which this thesis lays. First, Section 2.1.1 describes LTE networks, as the currently most advanced commercially deployed networks. Next, Section 2.1.2 describes 5G NR networks, whose first development phase has just finished with the definition of the 5G NR standard.

2.1.1 LTE

The successive integration of functionalities in the UMTS cellular standard and its differentiated circuit-switched and packet-switched subnetworks made this standard become hard and expensive to manage. As a result, an all-IP standard with plain architecture was proposed in Release 8 in 3GPP: LTE [3], which sometimes is referred as Evolved Packet System (EPS). Despite LTE has been marketed as the fourth-generation solution of 3GPP for mobile communications, it has not been until the advent of Long-Term Evolution Advanced (LTE-A) in Release 10 that the standard formally satisfied the requirements of a fourth-generation wireless technology.

The architecture of an LTE network is shown in Figure 2.1 and consists of the evolved UMTS terrestrial radio access network (E-UTRAN) and the evolved packet core (EPC).

The EPC (or core network, CN) is composed of the following elements [51]:

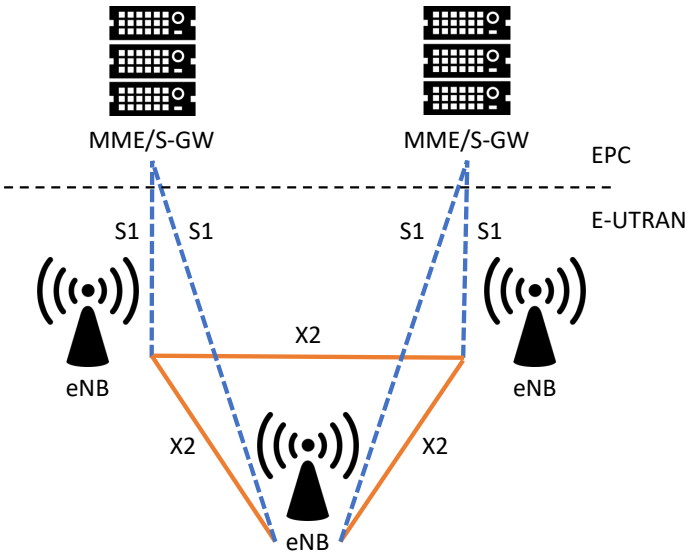


Figure 2.1: LTE overall architecture [3].

- Mobility management entity (MME), which is the main element of the EPC. It is the element in charge of managing the plane control between the CN and the UE. Specifically, the MME is responsible for:
  - Non-access stratum (NAS) signaling and NAS signaling security.
  - Access stratum (AS) security control.
  - RRC_IDLE state mobility handling, where RRC stands for radio resource control.
  - EPS bearer control.
- Serving gateway (S-GW), which is the user plane gateway to the E-UTRAN. Its main functions are:
  - Anchor point for mobility between LTE cells or cells from other 3GPP technologies.
  - Termination of user plane for paging.
  - Packet forwarding, routing and buffering of downlink data for UEs in RRC_IDLE state.
- Packet data network (PDN) gateway (P-GW), which is the user plane gateway to the PDN and the edge router between EPS and external PDNs. It also serves as a global anchor for mobility between 3GPP and non-3GPP access networks. Besides these, its responsibilities are:
  - Policy enforcement.
  - Charging support.
  - UE's IP address allocation.

The E-UTRAN is made up of only one type of network element: the evolved node B (eNB or eNodeB), which interfaces the users' devices. eNBs interconnect each other by means of X2 interfaces, and are connected to the EPC by means of S1 interfaces (Figure 2.1). The main functions of eNBs are the following:

- Radio resource management (RRM).
- IP header compression and encryption.
- Selection of MME at UE attachment.
- Routing of user plane data towards the S-GW.
- Scheduling and transmission of paging messages and information broadcast.
- Radio interface measurement making (uplink direction).
- Configuration reporting for mobility and scheduling purposes.

Figure 2.2 outlines the main functions included in the different elements of the LTE system. Among these functions, one of the most important is RRM. This function is in charge of ensuring an efficient use of the available radio resources to provide a large variety of services to a large number of users, while also ensuring the fulfillment of quality of service (QoS) requirements. Most RRM functions are included in the link and network layer, which are described next.

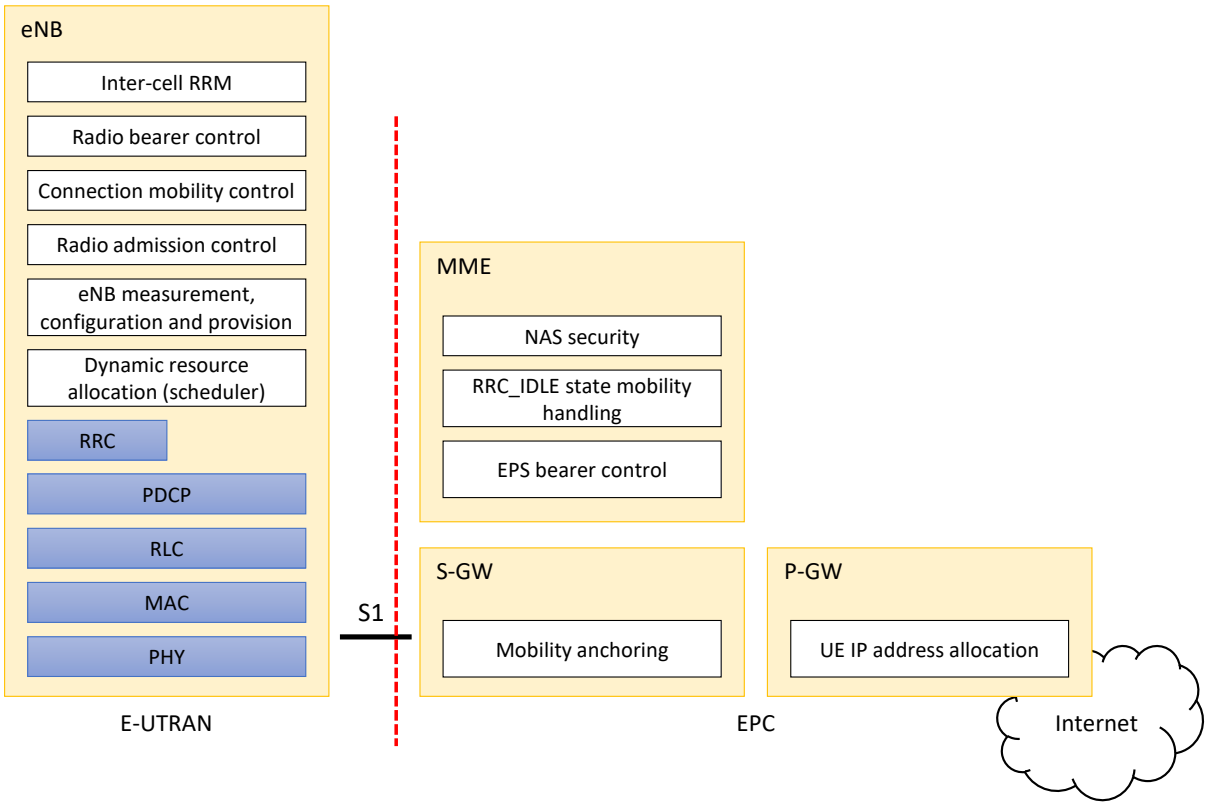


Figure 2.2: Functional split between E-UTRAN and EPC [3].

Physical layer

In LTE, the downlink and uplink multiple access over the air interface is performed by means of the orthogonal frequency division multiple access (OFDMA) and the single-carrier frequency division multiple access (SC-FDMA), respectively.

In the downlink direction, OFDMA makes use of the concept of orthogonal frequency division multiplexing (OFDM), which consists in the parallel transmission of a set of orthogonal subcarriers with the least possible spacing in frequency among them. According to 3GPP Release 15, each LTE subcarrier may be modulated using from a $\pi/2$-binary phase-shift keying (BPSK) to 256 quadrature amplitude modulation (QAM) in uplink and from quadrature phase-shift keying (QPSK) to 256 QAM in downlink [52]. Despite the high spectral efficiency that OFDMA allows, its main drawback is its high peak-to-average power ratio (PAPR), due to combination of a high number of subcarriers. A high PAPR makes radio transceivers have a high power consumption, which makes OFDMA not suitable for the uplink, preferring SC-FDMA instead. This technique combines the low PAPR of single-carrier transmission systems, such as GSM or CDMA, with the multi-path protection and flexible frequency allocation of OFDMA.

The physical layer of LTE is defined in a bandwidth agnostic way based on resource blocks, allowing it to adapt to various spectrum allocations. In LTE, the system (or carrier) bandwidth ranges from 1.4 MHz to 20 MHz. The smallest resource unit in the physical layer of LTE is the resource element (RE), which consists of one OFDMA or SC-FDMA symbol in the time domain and one subcarrier in the frequency domain. REs are grouped in a physical resource block (PRB), that is the smallest unit that can be scheduled for transmission, either in the uplink or in the downlink. A PRB spans over 0.5 ms (a time slot) in the time domain and over 180 kHz in the frequency domain. The number of subcarriers and symbols per PRB depends on the subcarrier spacing and the cyclic prefix length. For the usual case of a 15 kHz-subcarrier spacing and a normal cyclic prefix, a PRB is made up of 12 subcarriers and 7 symbols. Besides user data, a PRB also contains reference signals and other control data.

Finally, the multiple-input multiple-output (MIMO) technique allows the spectral efficiency to be further increased. MIMO is a multi-antenna technique which allows network coverage and capacity to be increased by transmitting multiple data streams simultaneously over the same frequency and time, taking full advantage of the different paths in the radio channel.

Link layer

The link layer is divided into three sublayers: medium access control (MAC), radio link layer (RLC) and packet data convergence protocol (PDCP).

According to [53], the main functions of the MAC sublayer are:

- Mapping between logical channels and transport channels.
- Multiplexing/demultiplexing of MAC service data units (SDUs) belonging to one or different logical channels into/from transport blocks (TBs) delivered to/from the physical layer on transport channels.

- Scheduling information reporting.
- Error correction through hybrid automatic repeat requests (HARQ).
- Priority handling between UEs by means of dynamic scheduling.
- Priority handling between logical channels of one MAC entity.
- Transport format selection.

The RLC sublayer, [54], includes the following functions:

- Transfer of upper layer packet data units (PDUs).
- Error correction through automatic repeat request (ARQ).
- Concatenation, segmentation and reassembly of RLC SDUs.
- Re-segmentation of RLC data PDUs.
- Reordering of RLC data PDUs.
- Duplicate detection.
- Protocol error detection.
- RLC SDU discard.
- RLC re-establishment.

Finally, according to [55], the main functions of PDCP are the following:

- Header compression/decompression of IP data flows using the robust header compression (ROHC) protocol.
- Compression and decompression of uplink PDCP SDU.
- Transfer of data (user plane or control plane).
- In-sequence delivery of upper layer PDUs at re-establishment of lower layers.
- Duplicate elimination of lower layer SDUs at re-establishment of lower layers for radio bearers mapped on RLC acknowledged mode (AM).
- Ciphering/deciphering of user plane and control plane data.
- Integrity protection and integrity verification of control plane data.
- Duplicated transmission and duplicate discarding.

Layer 2 includes three functions regarding RRM: HARQ, adaptive modulation and coding (AMC) and user scheduling, all of which are located at the MAC sublayer [53]. HARQ is a retransmission-based mechanism, which provides a fast and reliable connection between the UE and the eNB. AMC is the mechanism in charge of fighting fading by means of readjusting the modulation and coding scheme (MCS) applied to the TB to be sent in a transmission time interval

(TTI), according to the channel quality estimation. If the channel quality is poor (for example, due to a high interference level), a more robust channel coding and a simpler modulation scheme are applied to limit the TB error rate (also known as block error rate, BLER) to a typical value of 10%. If, on the other hand, a good channel quality is estimated, a higher coding rate and a denser modulation can be used to increase the spectral efficiency. The scheduler determines the assignment of uplink and downlink resources. Its basic task is to decide at each TTI which terminals make use of the available time and frequency resources. To that end, several policies may be followed.

Network layer

The network layer, or layer 3, is made up of the RRC sublayer [56], between the UE and the eNB, and the NAS sublayer, between the UE and the MME. While the latter is in charge of the directly transmitting signaling messages between the UE and the MME (without being interpreted by the eNB), like messages for UE mobility and bearer management, some of the functions of RRC are the following:

- Broadcast of system information related to the AS and NAS.
- Paging.
- Establishment, maintenance and release of an RRC connection between the UE and the E-UTRAN.
- Security functions including key management.
- Mobility functions: cell (re-)selection and handover (HO).
- QoS management functions.
- UE measurement reporting and control of the reporting.
- NAS direct message transfer to/from NAS from/to UE.

Along with the above, the main functions of RRC regarding RRM are the following:

- Radio bearer control (RBC). RRC is in charge of the establishment, maintenance and release of radio bearers, which may either be signaling radio bearers (SRBs), over which the control plane between the UE and the eNB is conveyed, or data radio bearers (DRBs), which are responsible for conveying the user's data plane between the UE and the eNB. When setting up a radio bearer for a service, RBC takes into account the overall resource situation in E-UTRAN, the QoS requirements of in-progress sessions and the QoS requirement for the new service.
- Radio admission control (RAC). Its purpose is accepting or rejecting the establishment of new radio bearers according to the ongoing resource situation in E-UTRAN, the established priority levels and the QoS requirements of the current and requested radio bearers.
- Connection mobility control (CMC). This function is in charge of controlling users' mobility, both the cell (re-)selection in RRC_IDLE state and the HO while the UEs are in

RRC_CONNECTED state. To that end, the E-UTRAN sets and broadcasts several configuration parameters through RRC SRBs, like those related with UEs' measurement and reporting procedures. In particular, the physical layer measurements reported by the UEs to E-UTRAN to be later considered in mobility decisions are:

  - Reference signal received power (RSRP). It is the linear average of the downlink reference signals across the system bandwidth. Since the reference signals exist only for one symbol at a time, the RSRP measurement is made only on those REs that contain cell-specific reference signals.

  - Reference signal received quality (RSRQ). RSRQ quantifies the signal quality and is defined as the ratio of RSRP to the received signal strength indicator (RSSI) of an evolved UMTS terrestrial radio access (E-UTRA) carrier. RSSI accounts for the received power over the whole system bandwidth, including the desired power from the serving cell, co-channel interference and other sources of noise.

Together with this, two novel connectivity schemes related to the RRM function of RBC were added from Release 10 onwards [3]:

- Carrier aggregation (CA). With Release 10, 3GPP introduced the concept of CA, allowing UEs to aggregate the bandwidth of several carriers (referred to as component carriers, CC) from a single eNB in order to support wider transmission bandwidths. According to 3GPP Release 15, a UE may aggregate up to 32 CCs or a total of 640 MHz.

  When CA is configured, the UE only has one RRC connection with the network. At RRC connection establishment/re-establishment/HO, one serving cell provides the NAS mobility information. This cell is referred to as the primary cell (PCell). Depending on UE capabilities, secondary cells (SCells) can be configured to form together with the PCell a set of serving cells. The PCell can only be changed with a HO procedure, whereas the SCells are added/modified/released according to the fulfillment of mobility events without the need to perform a HO.

- Dual connectivity (DC). Later on, with Release 12, 3GPP introduced the concept of DC. That is, the ability of a UE to aggregate CCs from two different eNBs. The eNBs involved in DC for a certain UE may assume two different roles: a master eNB (MeNB), which holds the signaling between the E-UTRAN and the EPC, or a secondary eNB (SeNB), which provides additional radio resources to the UE, but holds no signaling towards the EPC. In DC, a UE is connected to one MeNB and one SeNB.

  In DC, two radio bearer types exist: direct bearers and split bearers. A direct bearer is a bearer which uses radio resources from only one eNB. On the contrary, a split bearer is a bearer which uses radio resources from both the MeNB and the SeNB. Direct bearers may be divided, in turn, in master cell group (MCG) bearers and secondary cell group (SCG) bearers, depending on which node (MeNB or SeNB) they end in E-UTRAN (see Figure 2.3).

  The DC-controlling RRC entity is located in the MeNB, which holds the PCell. The resulting SRBs are always configured as MCG bearer type, only using radio resources of

the MeNB. DRBs, on the other hand, may be configured as MCG, SCG or split bearer type. At least one cell in SCG is configured with physical uplink control channel (PUCCH) resources. This is the primary secondary cell (PSCell).

Regarding the control plane, inter-eNB signaling for DC is performed by means of the X2 interface, whereas the signaling towards the MME is performed through the S1 interface.

For DC, two different user plane architectures are allowed: one in which the S1-U only terminates in the MeNB and the user plane data is transferred from MeNB to SeNB using the X2-U (case *a* in Figure 2.4), and a second architecture where the S1-U can terminate in the SeNB (case *b* in Figure 2.4).

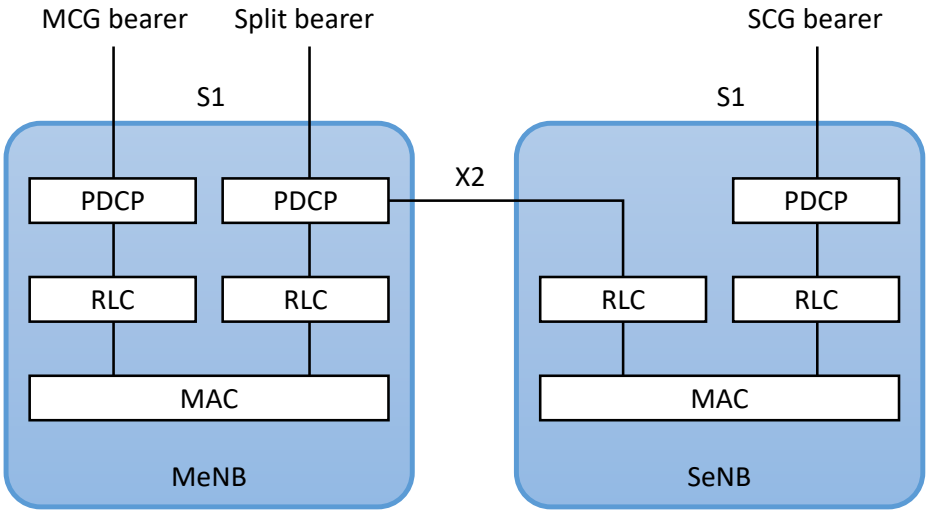


Figure 2.3: Radio protocol architecture for LTE DC [3].

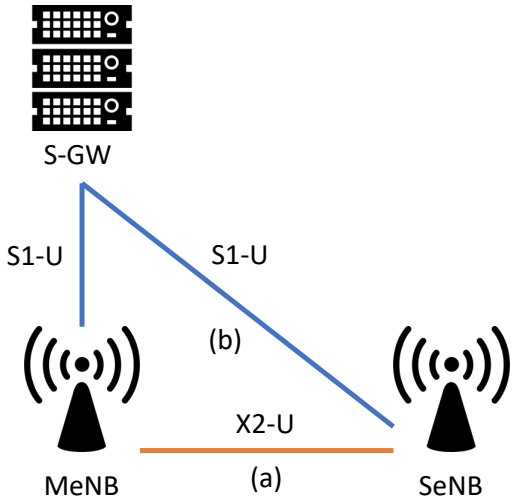


Figure 2.4: User plane connectivity of eNBs involved in LTE DC [3].

2.1.2 5G NR

With Release 15, 3GPP ends the first standardization stage of the 5G NR [4]. This new standard continues the all-IP scheme of LTE, while supporting a wide variety of services, characterized by quite differentiated requirement profiles. This is the case of eMBB, URLLC and mMTC. Whereas eMBB appears as a natural evolution of traditional MBB services, URLLC and mMTC are novel service categories, which aim to optimize other communication aspects rather than throughput, like latency or connection density, respectively. This can be seen in Figure 2.5.

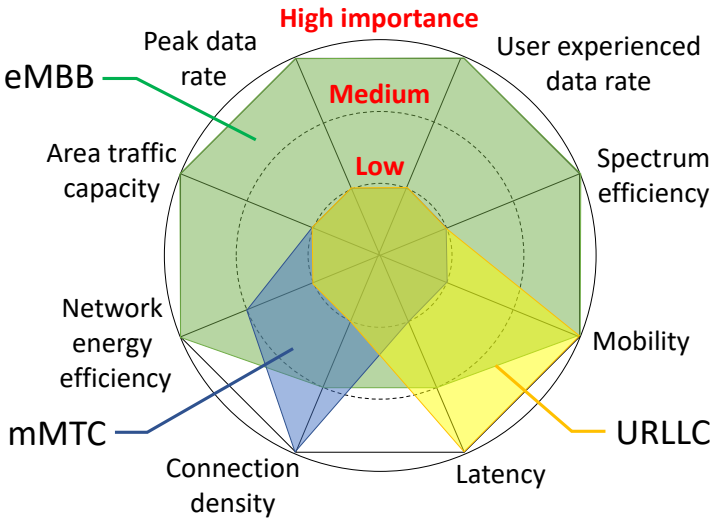


Figure 2.5: Novel service categories to be supported by 5G NR and their main features [1].

Many of the functionalities in 5G NR and the protocols responsible for these are those of LTE, which have been inherited and updated. Nevertheless, some major advancements in both the 5G RAN and the 5G CN (which are referred to as the next-generation RAN, NG-RAN, and the 5G core network, 5GC, respectively) must be highlighted, as they are the enablers for such degree of flexibility. Regarding the NG-RAN, these advancements can be summarized in a scalable numerology in the frame structure and resource shaping (with the definition of mini-slots, bandwidth parts, etc.), the inclusion of millimeter wave frequency bands, leading to an increased capacity and enhanced beam forming capabilities, the use of massive MIMO schemes and an improved QoS management of radio bearers through the service data adaptation protocol (SDAP). Concerning the advancements in the CN, 5GC will be responsible for managing network slicing in the NG-RAN. Network slicing is a concept to allow differentiated treatment depending on each customer requirements. With slicing, it is possible for MNOs to consider customers as belonging to different tenant types with each having different service requirements that govern in terms of what slice types each tenant is eligible to use based on service level agreement and subscriptions.

Figure 2.6 outlines the architecture of 5G NR [4], showing the split between the NG-RAN and the 5GC. The former is made up of NR-RAN nodes, which may be a next-generation node B (gNB), providing NR user plane and control plane protocol terminations towards the UE, or

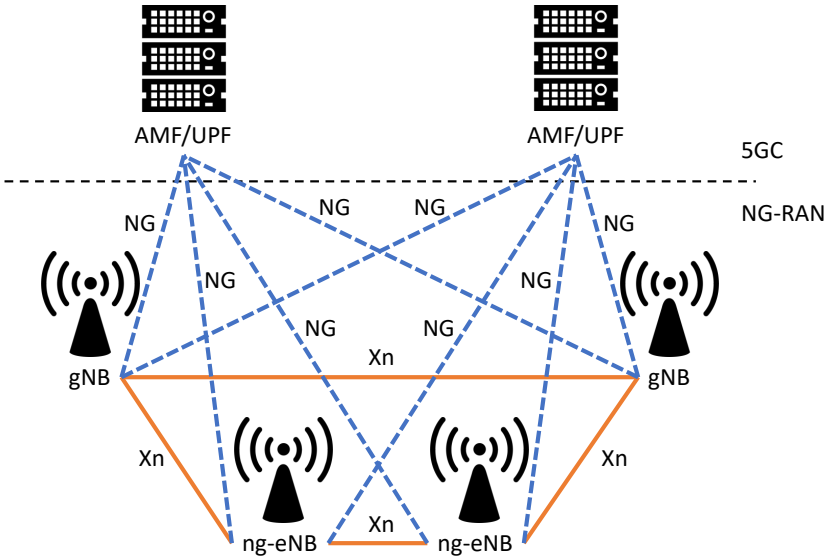


Figure 2.6: 5G NR overall architecture [4].

a next-generation evolved node B (ng-eNB), providing E-UTRA user plane and control plane terminations towards the UE.

The gNBs and ng-eNBs are interconnected with each other by means of the Xn interface. The gNBs and ng-eNBs are also connected to the 5GC by means of the NG interfaces; more specifically, to the access and mobility management function (AMF) by means of the NG-C interface and to the user plane function (UPF) by means of the NG-U interface.

Some of the main functions carried out by gNBs and ng-eNBs are the following:

- Functions for RRM: RBC, RAC, CMC and dynamic allocation of resources to UEs in both uplink and downlink.
- IP header compression, encryption and integrity protection of data.
- Selection of an AMF at UE attachment when no routing to an AMF can be determined from the information provided by the UE.
- Routing of user plane data and control plane information towards UPF(s) and AMF, respectively.
- Scheduling and transmission of paging and system broadcast information.
- Support of network slicing.
- QoS flow management and mapping to DRBs.
- Support of UEs in RRC_INACTIVE state, which is a novel battery-saving RRC state, recently included in 5G NR.
- Measurement and measurement reporting configuration for mobility and scheduling.

- Dual-connectivity.
- Tight interworking between NR and E-UTRA.

On the other hand, the 5GC is composed of an AMF, which roughly inherits the functions of the LTE MME; a UPF, which roughly inherits the functions of the LTE S-GW and P-GW, a session management function (SMF), which covers some of the functions of the LTE MME and P-GW.

The main functions of an AMF are the following:

- NAS signaling termination and security.
- AS security control.
- Inter CN node signaling for mobility between 3GPP access networks and support for intra- and inter-system mobility management control.
- Access authentication and authorization.
- Support of network slicing.
- SMF selection.

Regading the UPF, its main functions are:

- Anchor point for intra- and inter-RAT mobility.
- External PDU session point of interconnect to data network.
- Packet routing and forwarding.
- Packet inspection and user plane part of policy rule enforcement.
- Traffic usage reporting.
- Branching point to support multi-homed PDU session.
- QoS handling for user plane (e.g., packet filtering, gating, uplink/downlink rate enforcement).

Finally, the main tasks of SMF are the following:

- Session management.
- UE IP address allocation and management.
- Configures traffic steering at UPF to route traffic to proper destination.
- Control part of policy enforcement and QoS.

In a similar way to Figure 2.2 regarding LTE, Figure 2.7 summarizes the main functions included in the different elements of the 5G NR system.

The protocols which make up the AS of 5G NR are those of LTE (except for the SDAP protocol, which has been added on top of the user plane protocol stack), thus including all the

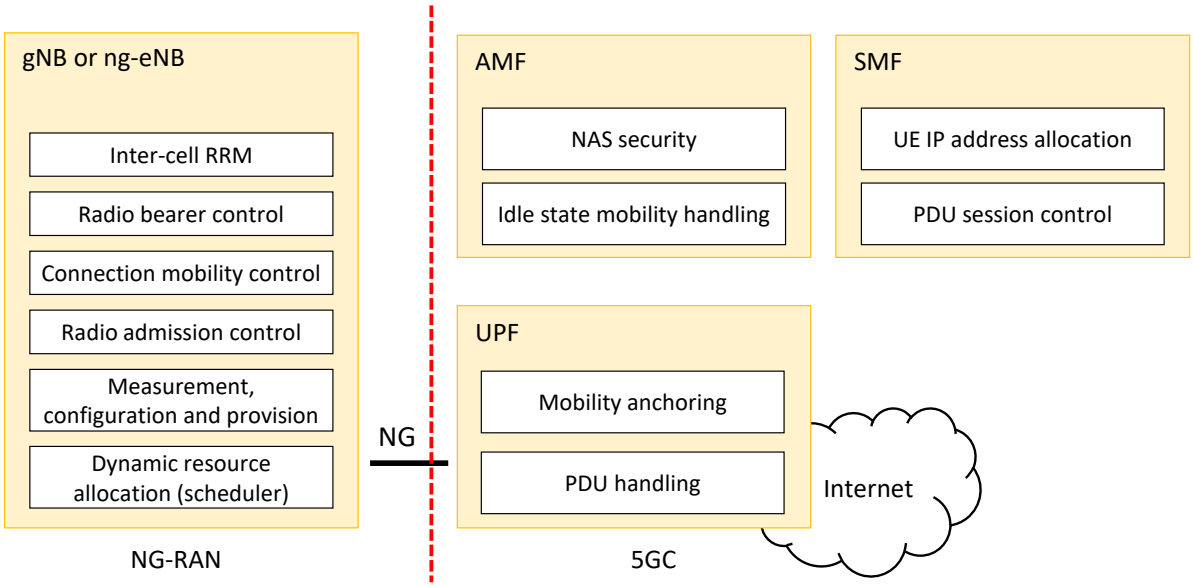


Figure 2.7: Functional split between NG-RAN and 5GC [4].

functions described in Section 2.1.1 for the network, link and physical layers. These protocols have been updated to enable the novel functionalities of 5G NR described so far: flexible resource shaping, millimeter wave frequency bands, massive MIMO and enhanced connectivity schemes, like MC, among others. The latter is described in the next section as one of the RRM functionalities of 5G NR to provide a highest degree of performance and flexibility, both for eMBB and URLLC services.

Multi-connectivity

Given that, during its deployment, the 5G NR will coexist with the current LTE network in a non-standalone manner, as of today, the concept of MC is described by 3GPP as an extension of the Release-12 DC [3,57] in the shape of a multi-RAT dual connectivity (MR-DC) [58], in which one of the nodes is an LTE eNB and the other one is a gNB, connected to the former by means of a non-ideal backhaul. In both connectivity schemes (DC and MR-DC), for a given user, one of the nodes assumes the role of the master node (MN), carrying the signaling between the UE and the core network and determining the UE RRC state, and the other one assumes the role of a secondary node (SN).

Regarding the signaling between the nodes and the UE, this is exchanged by means of RRC messages along SRBs. In Release-12 DC, SRBs can only be direct bearers, terminated at the MeNB. These may embed RRC messages from the SN, exchanged between this and the MN through the X2 interface. With MR-DC, direct SRBs over the SN and split SRBs are additionally allowed, improving the mobility robustness. Similarly to Release-12 DC, user plane data are exchanged through DRBs, which, may also be split or direct bearers.

Up until Release 15 (5G phase 1), MC only considers the aggregation of radio resources from two different nodes. It will not be until future releases that the concept of MC is expected to

encompass the simultaneous connection of a UE to more than two network nodes, all of which may be NR nodes (gNBs) in a standalone deployment.

The growing interest of standardization bodies, industry and academia towards the exploitation of MC functionalities arises from its multiple benefits, extending those of DC. One of them is the improvement of the connection robustness, due to the radio link diversity. This is especially relevant in high mobility scenarios [57], in which a high number of radio link failures (RLFs) would imply high HO failure rates. Besides, and related to this, MC enables reliability (understood as opposed to a packet loss rate) to be noticeably improved through data duplication at the PDCP sublayer [4] by sending packet duplicates through different logical channels. This feature is especially relevant for URLLC services, allowing both latency and packet loss rates to be reduced [59]. On the other hand, if the data flow is not duplicated but split among logical channels, managed by different network nodes, a throughput boost may be achieved from the UE side, with the additional benefit of a possible load balance among the network nodes involved in the process throughout the X2/Xn interface. In this case, the eMBB services benefit from this feature. Besides, both the reliability and throughput boosts may be further increased by making use of CA. This leads to a more general concept of MC, in which the ability of a UE to hold connections with a number of network nodes is expanded through the bandwidth aggregation provided by CA in each node, which may be seen as an additional degree of freedom from the network management point of view.

However, despite its benefits, MC also carries some drawbacks. First, an increase in the signaling load, both over the air, between the UE and the network, and among the nodes involved, over the backhaul network. Besides, as a big amount of resources from different network nodes are assigned to a relatively small number of users (the MC-capable UEs, which, in early stages will represent a minority), heterogeneous traffic densities might appear from one network node to another, possibly leading to an eventual load imbalance. Due to these facts, the resource allocation should be carefully done.

2.2 Self-organizing networks

In this section, the concept and different functions of SON are described. First, following the classical idea of SON in Section 2.2.1, as they were envisaged by the NGMN Alliance, and then, continuing with the concept of NG-SON in Section 2.2.2.

2.2.1 Classical SON

The complexity of cellular networks has significantly increased over the last years. This is mainly due to the integration of novel functionalities to give response to the ever-increasing demands of users for enhanced services and performance. Examples of these may be CA, as a means to increase the UE accessible bandwidth; DC, as a means to improve reliability, reducing the number of RLFs, or, in the coming years, MC. As a consequence, the management and infrastructure costs (often referred to as the OPEX and CAPEX, respectively) that MNOs must face have also noticeably increased.

The key to deal with such complexity arises from the automation of the network management tasks. This is the objective of SONs [60, 61]. Specifically, the main benefits of SON can be summarized as follows:

- To reduce installation time and costs.
- To reduce OPEX by reducing manual efforts when monitoring, optimizing, diagnosing and healing the network.
- To reduce CAPEX due to a better use of the network infrastructure and spectrum resources.
- To improve the network performance.
- To improve the user experience.

Now, depending on the location of the algorithms for the network management automation, three different SON architectures may be distinguished: distributed SON, centralized SON and hybrid architecture SON. In a distributed architecture, the SON mechanisms are located at the lowest level of the network: the network nodes (i.e., the eNBs or gNBs). This approach guarantees a fast response to any performance issue that may arise, at the expense of having a myopic vision of the network. With this approach, every network node is only responsible for itself, not having information from other nodes. If a wide area SON mechanism was to be deployed, every node should continually exchange performance and configuration data with its neighbors, which would make the signaling load rise without even guaranteeing a stable network management, given its inherently individualized operation mode and its possibly resulting back-and-forth operation. In a centralized architecture, SON algorithms are located in a central entity for network management, towards which network nodes periodically send configuration and performance data. This alternative allows to optimally manage the network, since full snapshots of the system can be taken. However, the huge amount of nodes that a nowadays network manages at the same time makes these data need to be reported every certain minutes, in order not to incur into an excessive signaling load or overtake the computing power of the SON system. As a result, this solution shows a slower response than that of a distributed approach, not being suitable for fast-changing network issues. In order to overcome these inconveniences, the hybrid SON architecture implements the functions devoted to an agile response at a node level, whereas the more delay-tolerant management functions are implemented in a centralized unit.

For SON functions to be able to make their own decisions, information from the network must previously be gathered. These sources of information are the following [16]:

- Configuration management (CM) parameters. This is the current configuration settings of the network elements.
- Performance management (PM) parameters. This information consists of counters, which reflect the performance of network elements, quantifying the number of times that a certain event takes place in a time window.
- Alarms, also known as fault management (FM) data. This kind of messages are raised whenever a failure takes place in the network.

- Mobile call traces. This source of data consists of very detailed information of events generated by specific UEs, which can be collected from network elements.
- Drive tests. This information is gathered from field measurements, which are taken within a specific area with specialized equipment.
- KPIs. These indicators are calculated by combining other counters or measurements to provide a meaningful performance measure. They are the measurements more frequently used as input of SON functions.
- Context information. This information is not gathered from the network, despite it has an impact on the network. This is information related to the environment, such as type of area (e.g., urban area) or typical weather in the cell (e.g., rainy) or the user location and speed.

Figure 2.8 shows the main SON functions according to 3GPP [7], which are categorized into three function groups:

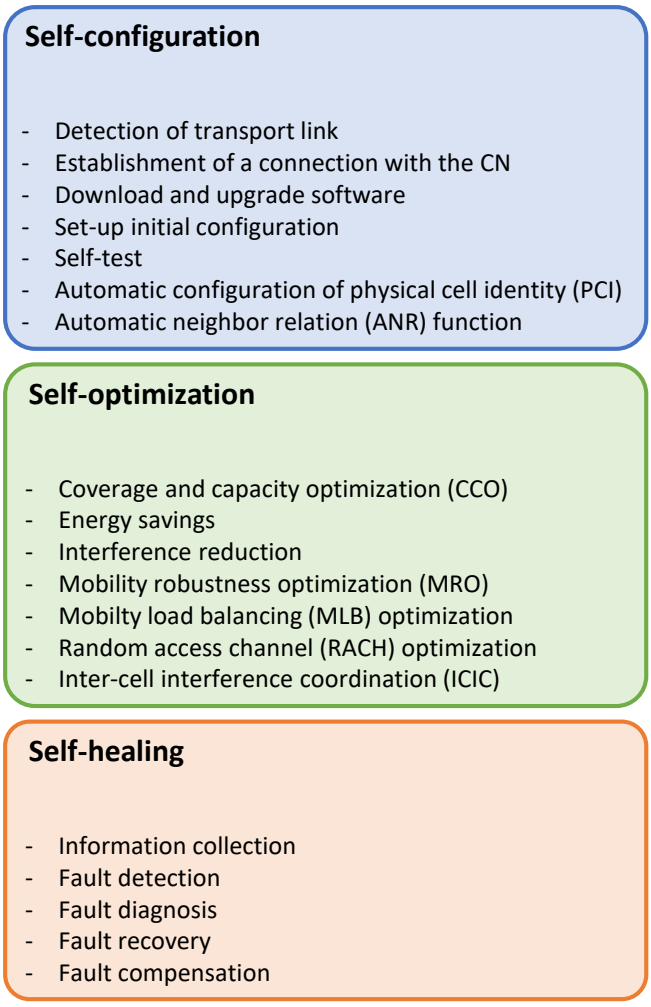


Figure 2.8: Main SON functions, according to 3GPP.

- Self-configuration. Self-configuration is the process of bringing a new network element or network element parts into service with minimal human operator intervention. As a result of this, some tasks of self-configuration may be distinguished: to detect the transport link and to establish a connection with the core network elements, to download and upgrade to the latest software version, to set up the initial configuration parameters including neighbor relations and to set the node to operational mode among others. The main use cases defined in self-configuration are [10]:
  - Automated configuration of the physical cell identity (PCI). This SON use case provides an automated configuration of a newly introduced cell physical identity. When a new node is brought into the field, a physical cell identity needs to be selected for each of its supported cells, avoiding collision with respective neighboring cells.
  - Automatic neighbor relation (ANR) function. The purpose of the ANR function is to relieve the operator from the burden of manually managing neighbor relations. To that end, the ANR functionality instructs UEs to perform measurements which, once reported to the corresponding node, allows it to autonomously determine the best neighboring nodes. These relations are built so that the UE's performance is degraded the least (reducing the probability of HO failures and dropped calls) while preventing the signaling load from excessively increasing for mobility reasons.
- Self-optimization aims at holding an adequate network performance and service quality in opposition to the time-varying issues that may lead to a suboptimal operating point. This is mainly done by readjusting network parameters. 3GPP highlights different use cases of self-optimization [10]:
  - CCO. A typical operational task is to optimize the network based on coverage and capacity criteria. Coverage optimization has usually a higher priority than capacity optimization. To provide an optimal coverage, users should establish and maintain connections with acceptable or default service quality, according to operator's requirements. It implies that coverage is continuous and users are unaware of cell borders. The coverage must therefore be provided in both, idle and active mode for both, uplink and downlink. Coverage optimization algorithms must take the impact on capacity into account. Since coverage and capacity are linked, a trade-off between the two goals may also be a subject of optimization.
  - Energy savings. Energy savings based on enabling the possibility, for a cell providing additional capacity in a deployment where capacity boosters can be distinguished from cells providing basic coverage, to be switched off when its capacity is no longer needed and to be re-activated on a need basis.
  - Interference reduction. Capacity could be improved through interference reduction by switching off those cells which are not needed for traffic at some point of time.
  - Mobility robustness optimization (MRO). The main objective of mobility robustness optimization is to reduce the number of HO-related RLFs and to drive the mobility parameters to an optimal configuration to prevent the degradation of the service performance.

  - MLB operation. This use case aims to alleviate the effects of a geographically uneven traffic distribution, which are the waste of network resources in the least loaded areas and the saturation of the network in the most loaded areas. The objective of MLB is to equally distribute the traffic load among the network nodes in order to make the most of the capacity provided by the network. To steer UEs from the most loaded nodes towards less loaded ones, mobility parameters are tuned.

  - Random access channel (RACH) optimization. This use case addresses the issue of the preamble collision in a call attempt in the RACH. The number of preamble collisions increases with the number of devices contending for access, which makes access delays increase. Thus, this use case aims to reduce the access delay while minimizing the interference of a given node towards its neighbors by adjusting the RACH-related configuration parameters.

  - Inter-cell inteference coordination (ICIC). The objective of this use case is to reduce or avoid the interference between PRBs in uplink and downlink by a coordinated usage of available radio resources in adjacent cells. This coordination is carried out by prioritizing users in the different cells. The main benefit is an improved signal quality and higher user data throughput.

- Self-healing. The target of self-healing functions is to reduce as much as possible the network performance degradation due to a failure. First, and before the current failure is known, fault detection, diagnosis and compensation tasks are carried out. Then, when the failure root cause is known, the corresponding actions for recovery are performed. According to [16], these functions are described as follows:

  - Information collection. The first phase of a SON algorithm is to collect information from the network in order to analyze the current status. The more complete the information is, the faster the failures are identified and solved.

  - Fault detection. The objective of this task is to find cells with service degradation or service outage. This is usually done by comparing each cell state with a cell normal operating behavior.

  - Diagnosis. The next step after identifying a cell as faulty, is determining the root cause of this misbehavior, so corrective actions can be taken. This is the objective of the diagnosis task. In the same way that a doctor diagnoses a patient regarding a number of symptoms, this task aims at determining the ultimate cause of a network failure by assessing a set of features.

  - Fault compensation. In order to reestablish an adequate service quality under the presence of a failure, compensation tasks are carried out. An example of this is the tilt adjustment of cells neighboring a cell in outage in order to absorb its traffic load and prevent a coverage hole.

  - Fault recovery. This function carries out the execution of the identified repair actions to solve the diagnosed failure.

This thesis focuses both on self-healing and self-optimization tasks. Regarding the former, the enhancement of diagnosis tasks is covered in Chapters 3 and 4 by including some of the features of NG-SONs, described next. Concerning self-optimization, this thesis aims at enhancing RRM by describing a general-purpose framework for CC management in a MC-enabled environment in Chapter 5.

2.2.2 Next-generation SON (NG-SON)

Currently, SONs are characterized by the use of mechanisms to automate the RAN management, many of which rely on controller-based approaches [13, 14, 26], or ML-based techniques [24, 28, 29, 34]. In the coming years, however, the advent of the 5G NR and its different and new functionalities as well as the increasingly high number of users of cellular networks (most of which will be machines, under the Internet of Things paradigm) will make these solutions become inefficient, due to the high volume of available information. As a result, NG-SONs should include a number of features in their way to become fully autonomous while providing optimal resource usage. Some of these are:

- The use of **big data techniques**. As the network becomes more complex due to the increasing number of functionalities and users, the volume, variety and velocity of available performance data will become hard to handle. Luckily, in the last decade, different data processing techniques and programming frameworks have been proposed in order to deal with this [8,30]. This is the case of the MapReduce programming model, for example, which gives support to parallel computation and is implemented in the open source framework Hadoop.

- Complementing the above, and in order to express the available performance information in more efficient way, **dimensionality reduction techniques** will be used. This kind of techniques aims to reduce the amount of dimensions (performance indicators, in this case) required to represent the network state, losing the least possible useful information in the process. The application of dimensionality reduction techniques in SON tasks is covered in Chapter 4 in this thesis.

- Related to the above, NG-SONs will be characterized by using a much more **varied amount of sources of information** for network management. Specifically, the scope of NG-SON will broaden both vertically and horizontally compared to the scope of SONs. In the vertical direction, NG-SONs will not only assess the lower layers of the devices communicating through the air interface, but will span to the top layers of the end devices, gathering application-specific performance information and allowing a **user-centric network management**. In the horizontal direction, NG-SONs will span from an end device to another, including the intermediate network elements, thus allowing an **E2E management**.

- Following the latter, information from outside the network will also be used for management tasks. This is the case of **context information**, which will provide insight about external factors that actually have an effect on the network. Context information could then be used for diagnosis [32] or for optimization purposes [62], for example.

- Despite current SONs already work with information in the shape of time series [33, 63], the time resolution of these data usually spans from fifteen to sixty minutes, which imposes this lower bound to the time response of SON mechanisms. This fact is given by the storage constraints imposed in the network OSS. The use of big data and dimensionality reduction techniques for data processing will enable this time resolution to be reduced, allowing a **short-term network management** and **flexible time resolution mechanisms** to be executed.
- Finally, and concerning the above, further studies on time series analysis will conduct NG-SONs to develop **proactive management schemes**. In this way, for example, NG-SONs will be able to avoid network performance degradation before this actually takes place. At this point, context data will also provide a valuable source of information.

Chapter 3

Ensemble method for RCA in cellular networks

In this chapter, a method for the combination of several automatic diagnosis techniques and diagnosis models is proposed, in order to develop an ensemble automatic diagnosis system that outperforms its baseline components.

This chapter is organized as follows. First, Section 3.1 introduces the related work, both regarding automatic diagnosis in cellular networks and the techniques used until now to develop ensemble classifying systems. Section 3.2 presents the problem formulation. Section 3.3 introduces the proposed method to combine multiple baseline diagnosis techniques and models. In Section 3.4, results are analyzed by means of both simulation-based data and data from a live LTE network. Finally, Section 3.5 summarizes the main conclusions.

3.1 Related work

This section aims to provide a brief survey on the most recently proposed techniques for self-healing in cellular networks, as well as some of the most well-known ensemble-based systems for classification, most of which have been used in areas not related to mobile communications.

Nowadays, self-healing has taken advantage of different and advanced automatic mechanisms; mainly for its tasks of detection and diagnosis. Regarding the first, an ensemble of univariate and multivariate methods over a set of KPIs was first proposed in [17] and then implemented as a software tool [18]. With respect to diagnosis, in [20] and [21] diagnosis systems based on Bayesian Networks were proposed. In [22], a scoring system to determine how well a specific network observation fits a diagnosis was used, which was enhanced by [23] by adding subsequent profiling techniques. The method in [28] was based on fuzzy logic and genetic algorithms to develop a rule-based diagnosis system, and [24] proposed an unsupervised diagnosis system based on a variety of neural networks called self-organizing maps (SOM). Up to our knowledge,

diagnosis in cellular networks lacks of ensemble-based methods, which have already been used in other fields, like industry or medicine. For example, in [64], an ensemble of neural networks with cross-validation for fault diagnosis of analog circuits with tolerance is proposed. In this case, the ensemble consists of applying a bagging predictor over the output of the neural networks. In [65], several k-nearest neighbor (kNN) classifiers are put together on a majority-vote ensemble to classify the patterns that several proteins may exhibit when folded.

With respect to ensemble-based methods themselves, one of the earliest ideas consisted in partitioning the feature space (i.e., the vector space in which the features of the cases to be diagnosed are defined) and assigning each part to a different classifier, which is supposed to be the best for this subset of cases [66]. This idea has been widely explored and has given birth to the so-called *mixture of experts* algorithm [67,68], being the paradigm for the *classifier-selection* type of ensemble methods. Under this approach, only one classifier is working at the same time and its selection is determined by the partition that the sample under test belongs.

Conversely, in *classifier-fusion* methods, all classifiers are trained over the entire feature space. The classifier combination process involves merging the individual classifiers to obtain a system that outperforms the standalone classifiers. This is the basis for the widely used bagging and boosting predictors, [69] and [70], respectively. The algorithm AdaBoost is an example of the latter and one of the most known and used algorithms for classifying nowadays. Classifier fusion methods can also be divided into those which work with classification labels only and those which make use of a continuous valued output for each classifier for every class. In this case, the outputs can be seen as the support an expert gives to a class in terms of the class-conditional posterior probabilities [71].

3.2 Problem formulation

3.2.1 RCA in mobile communications networks

In the same way that a patient is diagnosed by a doctor, based on the symptoms he shows, the state of a communications network may be diagnosed based on performance data. This diagnosis task, also called RCA or troubleshooting, is often carried out by human experts, using their knowledge on the underlying relations that the observed indicators and the status of the network have. However, the number of symptoms (counters, alarms, KPIs, call traces, etc.) and possible fault causes the expert has to deal with increases as networks grow in size and complexity, which makes this task to become a very difficult and time consuming issue.

Furthermore, the current manual troubleshooting is a layered task, guided by a Trouble Ticket (TT) system. In this system, a group of specialists tries first to diagnose and solve the problem by performing some simple checks. If they can not find the root of the problem, this is raised to a more specialized team (and so on), which performs a deeper study on the symptoms the case exhibits and resorts to field engineers in case they need to make some on site checks.

As a response to this more and more inefficient procedure, automatic diagnosis systems arose in an attempt of imitating the way of acting of troubleshooters. Figure 3.1a shows the basic

scheme of a system for automatic diagnosis. It is composed of an automatic diagnosis technique and a diagnosis model. The first is an artificial intelligence system that outputs a diagnosis taking a set of symptoms (e.g., KPIs) from a test case as its input. The second represents the knowledge a human expert would have on the underlying relations between the symptoms and the fault causes and may take different forms depending on the diagnosis technique it is destined to work with. For example, for a Bayesian classifier, this model has the shape of prior probabilities and probability density functions (PDFs), whereas for a case-based reasoning (CBR) method, the diagnosis model acquires the shape of sets of rules. As it can be seen in this figure, the diagnosis model may be built from a set of training cases by means of a ML algorithm or by troubleshooting experts by gathering their knowledge. The proposed method aims to combine the knowledge acquired by any number and kind of diagnosis models and automatic diagnosis techniques in an attempt to reduce the errors in fault detection and diagnosis. This is shown in Figure 3.1b, in which a number of automatic diagnosis systems are separately built. Whenever a new test case arrives, each of these systems outputs a different diagnosis, which are finally combined to produce a more correct ensemble diagnosis.

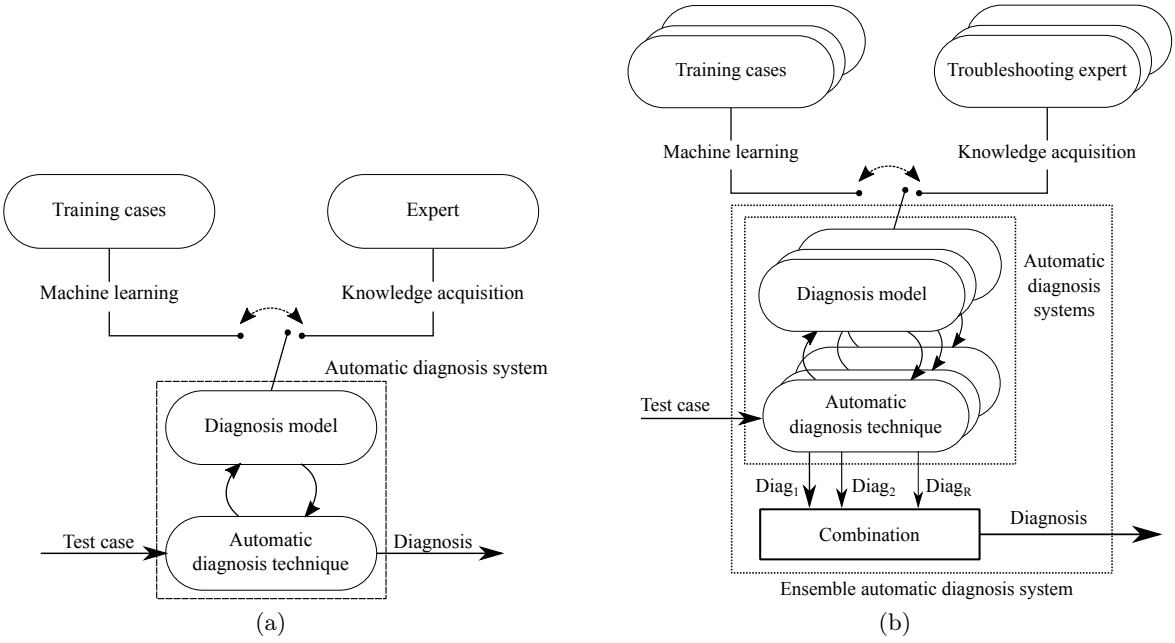


Figure 3.1: Schemes of (a) an automatic diagnosis system and (b) an ensemble automatic diagnosis system.

3.2.2 Automated diagnosis from the classification theory

A diagnosis system is a method that, given a set of indicators or symptoms (called *case* hereafter), intends to infer the cause that provoked them. In this sense, a diagnosis system acts as a classifying system, in which the attributes from the cases to be classified correspond to the symptoms from the case to be diagnosed, and the classes to be assigned correspond to the causes to be inferred. This is an issue long time investigated in data mining theory [72], and many types of classifiers have been developed over the years in an attempt to get the maximum information

the cases under diagnosis could provide. However, no algorithm has proven to be clearly better than the rest for all kinds of input data by now. One reason for the increasing efforts in the related research is that the performance of a classifier normally depends on the nature and distribution of the data it has to work with. For this reason, the present chapter focuses not only on combining different diagnosis models but on offering the possibility to combine multiple classifiers in the form of automatic diagnosis techniques.

Let us assume that we have a set of M fault causes to diagnose and R diagnosis systems (which differ either in their diagnosis model, in the their diagnosis technique or in both) to combine, and that each of these systems can have a subset of these causes as their output, namely, W^r for the system r. In this scenario, the set of causes a diagnosis system can identify may be different from one system to another. This can be seen in Eq. (3.1), where each row stands for a W^r and the element w_m^r stands for the m^{th} fault cause, considered by the r^{th} system. According to this, each row may be different from another.

$$\begin{pmatrix} W^1 \\ \vdots \\ W^r \\ \vdots \\ W^R \end{pmatrix} = \begin{pmatrix} w_1^1 & \dots & w_m^1 & \cdots & w_M^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_1^r & \dots & w_m^r & \cdots & w_M^r \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_1^R & \dots & w_m^R & \cdots & w_M^R \end{pmatrix} \tag{3.1}$$

In a diagnosis system, a case, $\overline{x}$, is characterized by its symptoms, x_n, where $\overline{x}$ may be written as $\{x_1, x_2, \dots, x_N\}$, having a total of N possible symptoms. And similarly to the considered fault causes, each diagnosis system may consider only a subset of these symptoms, namely, N^r for the diagnosis system r.

In the context of diagnosis systems for mobile communication networks a case corresponds to an observation or measurement from the network; a symptom may be an event counter, a KPI, a call trace or an alarm and the causes are seen as the network states, among which the normal and several fault states may be distinguished. In this chapter, some results from theory of classifiers are used, extended and applied in this context in an attempt of combining the knowledge acquired by these R diagnosis systems, developing a more reliable and accurate RCA system for communication networks.

3.3 Method for combining multiple automatic diagnosis systems

In this section, a method to combine the knowledge acquired by any number and kind of standalone automatic diagnosis systems by means of a classifier-fusion scheme is proposed. The proposed method consists of two stages: the construction of behavior models of the automatic diagnosis systems from training cases, Section 3.3.1, and the combination of these models in order to make a more accurate diagnosis over test cases, Section 3.3.2. This can be seen in Figures 3.2 and 3.5. Before this method can be applied, two sets of N-dimensional cases must be distinguished: the modeling set and the testing set, where each of these N dimensions stands

for a working KPI. The modeling set will be used in the first stage and the testing set in the second.

3.3.1 Construction of the behavior models

The baseline diagnosis systems are to be combined by means of mixing their models of behavior, which need to be extracted first.

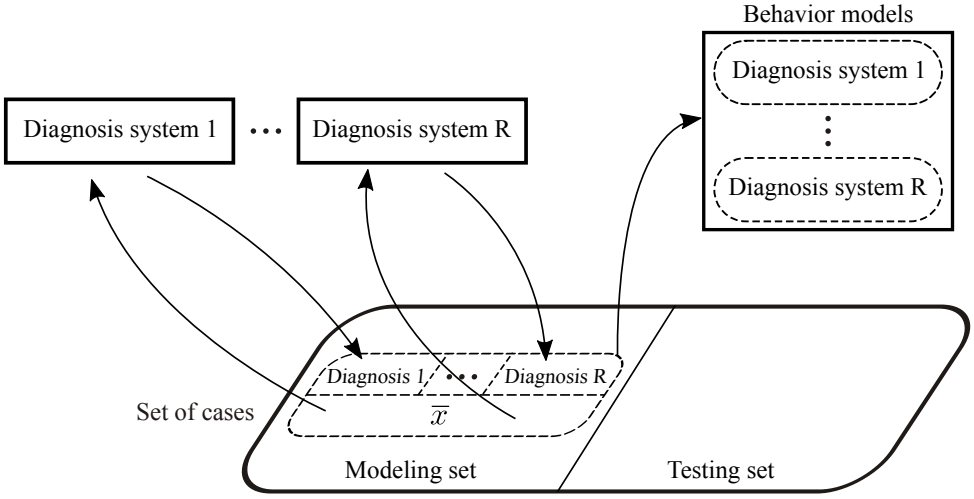


Figure 3.2: Proposed method for combining diagnosis systems. Stage 1: Construction of the behavior models.

Once the diagnosis model from each diagnosis system has been built (see Figure 3.1) either from training cases via a ML method [29], or from the experts' knowledge [24], each diagnosis system can start classifying. In this stage, every case from the modeling set is diagnosed (i.e., classified) by the R systems. That is, each system assigns to each case one of the M possible fault causes; in particular, one of the causes that system can discern. This can be seen in Figure 3.2, where the case $\overline{x}$ acts as the input for the R systems and, in turn, they assign it R diagnosis labels. If the system r diagnoses the case $\overline{x}$ with the cause m, this case receives the label w_m^{r*}. In this way, each diagnosis system makes a different partition of the modeling set into $|W^{r*}|$ disjoint subsets, whose maximum is $|W^r|$, that is, the number of causes that system considers, Figure 3.3, where $|A|$ is the number of elements in the set A. This leads to finally identify M^* different causes, being M^* the union of W^{r*} over r, with $M^* \leq M$. According to this, a new matrix from Eq. (3.1) may be written, substituting every row (i.e., every W^r) by its corresponding W^{r*}. Each row would represent one of the partitions of the modeling set and each column would represent how a cause "is seen" by each diagnosis system regarding the KPIs that the cases labeled as w_m^{r*} exhibit.

It should be noticed that each of these M^* subsets contains a number of $|N^r|$-dimensional cases. At this point, the behavior of the diagnosis system r is modeled through the estimation of the statistical distributions of the $|N^r|$ KPIs for the cases belonging to W^{r*}. That is, the behavior of each diagnosis system is modeled by means of $|N^r| \times M^*$ PDFs. The estimated

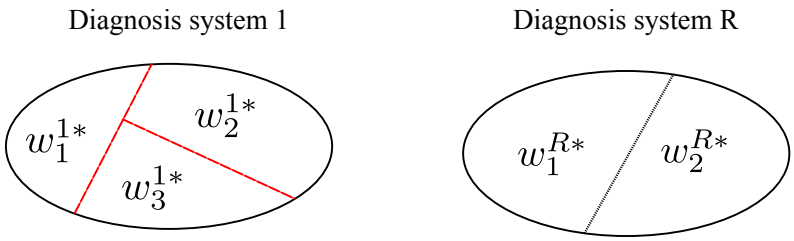


Figure 3.3: Modeling set divided into different subsets by means of two different partitions: on the left, the partition the first diagnosis system makes, having $W^1 = \{w_1^1, w_2^1, w_3^1\}$ with $|W^1| = |W^{1*}| = 3$; on the right, the partition the diagnosis system R makes, having $W^R = \{w_1^R, w_2^R, w_3^R\}$ and $W^{R*} = \{w_1^{R*}, w_2^{R*}\}$. In this last case, the diagnosis system R only diagnosed the causes 1 and 2 although being able of also identifying the fault cause 3.

statistical distribution of the n^{th} KPI for the subset of cases diagnosed as m by the diagnosis system r is $p(x_n|w_m^{r*})$. The choice of the PDF that estimates each one of these distributions is done according to the maximum likelihood criterion. To do so, some families of PDFs are considered in the fitting procedure, Table 3.1. In a first step, the distribution of the KPI x_n from the cases labeled as w_m^{r*} is fitted according to the maximum likelihood criterion with each one of the considered families of PDFs. This results in a set of candidates for estimating its distribution. These PDFs are then sorted by their likelihood and the one with the maximum value is chosen to be the estimation for the KPI.

The reason for considering these families of PDFs is to get the better estimation of the distribution of the KPI x_n given its belonging to w_m^{r*}. As an example, Figure 3.4a shows a normalized histogram of the KPI *95th percentile RSRP* from the cases labeled as w_m^{r*}. In this figure, two families of PDFs have been used in an attempt of fitting the underlying histogram: the normal and the generalized extreme value (GEV) functions. As it can be seen, the latter fits it better, resulting in a higher value in a likelihood-ratio test.

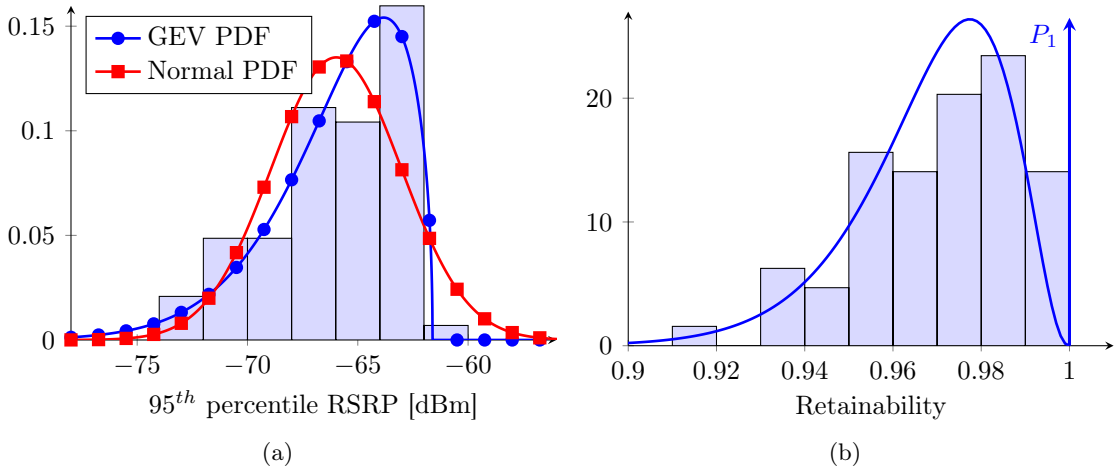


Figure 3.4: (a) Normalized histogram for the KPI 95th percentile RSRP and two fitted PDFs: a GEV PDF in blue and a normal PDF in red. (b), Normalized histogram for the KPI Retainability and a β' PDF estimation.

Table 3.1: Families of PDFs considered for the estimation of $p(x_n|w_m^{r*})$.

Distribution	**PDF**	**Parameters**
Beta	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	a, b
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ, σ
Log-normal	$\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln(x-\mu)}{\sigma}\right)^2\right)$	μ, σ
Exponential	$\lambda\exp(-\lambda x)$	λ
Gen. Extreme Value	$\frac{1}{\sigma}t(x)^{\xi+1}\exp(-t(x)),$ $t(x)=\begin{cases}\left(1+\left(\frac{x-\mu}{\sigma}\right)\xi\right)^{-\frac{1}{\xi}} & \xi\neq 0\\ \exp(-(x-\mu)/\sigma) & \xi=0\end{cases}$	μ, σ, ξ
T-location	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}$	ν, μ, σ
Nakagami	$\frac{2m^m}{\Gamma(m)\Omega^m}x^{2m-1}\exp\left(-\frac{m}{\Omega}x^2\right)$	m, Ω
Gamma	$\frac{1}{\Gamma(k)\theta^k}x^{k-1}\exp\left(-\frac{x}{\theta}\right)$	k, θ
Logistic	$\frac{\exp\left(\frac{x-\mu}{s}\right)}{s\left(1+\exp\left(-\frac{x-\mu}{s}\right)\right)^2}$	μ, s
Log-logistic	$\frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{\left(1+(x/\alpha^\beta)\right)^2}$	α, β
Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left(-\frac{x}{\lambda}\right)^k$	λ, k
Rayleigh	$\frac{x}{\sigma^2}\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right)$	σ
Rice	$\frac{x}{\sigma^2}\exp\left(-\frac{x^2+\nu^2}{2\sigma^2}\right)I_0\left(\frac{x\nu}{\sigma^2}\right)$	ν, σ

While some KPIs are counters and they do not have an upper limit, there are others that are inherently bounded, usually between zero and one, as they are defined as a ratio. Normally, the beta PDF is used to fit these KPIs [73]. KPIs like the retainability or the accessibility often reach these extreme values making the resulting fitted beta function present asymptotes in these values. To avoid this issue, a modified beta function (β') is used instead of that of Table 3.1. In particular:

$$\beta'(x) = (1 - P_0 - P_1)\hat{\beta}(x) + P_0/h_\beta\delta(x) + P_1/h_\beta\delta(x-1), \tag{3.2}$$

where $\hat{\beta}(x)$ stands for the distribution fitted to a set with no extreme values; P_0 and P_1 stand for the relative frequency of cases with value 0 and 1, respectively; δ stands for the Dirac's delta

and h_β stands for the step (the resolution) when computing β'. This can be seen in Figure 3.4b, where a normalized histogram for the KPI retainability is shown.

3.3.2 Combination of behavior models

This stage uses the cases from the testing set. In the previous stage the estimated functions have been seen as conditional PDFs. That is, functions that express how the KPIs are distributed over the cases diagnosed with a given cause by a given system. However, this set of functions may be seen as likelihood functions by just changing the approach. From this point of view, the function depends on w_m^{r*} given that an observation of the random variable x_n (that is, the n^{th} KPI) has taken place.

At this point, some diagnosis system may have not diagnosed a given cause, as seen in Figure 3.3. Assuming that the fault cause m was not diagnosed, the function $p(x_n|w_m^{r*})$ would have a value of 0 $\forall x_n$. That is, it is impossible for the fault cause m to take place according to the diagnosis system R no matter what value the KPI n takes.

Now, assuming that the KPIs are independent among each other, a joint likelihood function of w_m^{r*}, that is, $p(\overline{x}|w_m^{r*})$, may be written as

$$p(\overline{x}|w_m^{r*}) = \prod_{n \in N^r} p(x_n|w_m^{r*}). \tag{3.3}$$

Given Eq. (3.3), and assuming that the prior probability of w_m^{r*} ($P(w_m^{r*})$) is given as the relative frequency of the cases labeled as m by classifier r, the *a posteriori* probability for a diagnosis system r to diagnose a case with the cause m given that its KPIs equal $\overline{x}$ (i.e., $P(w_m^r|\overline{x})$) can be calculated by just applying the Bayes' theorem. That is,

$$P(w_m^{r*}|\overline{x}) = \frac{p(\overline{x}|w_m^{r*})P(w_m^{r*})}{\sum_{w_i^{r*} \in W^{r*}} p(\overline{x}|w_i^{r*})P(w_i^{r*})}. \tag{3.4}$$

At this point, $M^* \times R$ *a posteriori* probabilities may be distinguished. Figure 3.5 shows this when a case $\overline{y}$ from the testing set is to be diagnosed. As it can be seen in this figure, the KPIs from case $\overline{y}$ act as input values in the behavior models of the R diagnosis systems; that is, the likelihood functions $p(\overline{y}|w_m^{r*})$ for $w_m^{r*} \in W^{r*}$ and $r = 1, \ldots, R$. Then, the *a posteriori* probabilities $P(w_m^{r*}|\overline{y})$ are computed using these together with $P(w_m^{r*})$ by means of the Bayes' theorem.

Now, these $M^* \times R$ *a posteriori* probabilities together with the prior probabilities can be combined over R using some algebraic functions, producing M^* probabilities of the kind $P(w_m^*|\overline{y})_{Rule_t}$ per function used, where m again stands for the cause and t is an index for the rule used in the combination. That is:

$$P(w_m^*|\overline{y})_{Rule_t} = f_{Rule_t}\left(P(w_m^{1*}|\overline{y}), \ldots, P(w_m^{R*}|\overline{y}); P(w_m)\right), \tag{3.5}$$

where $P(w_m)$ stands for the relative frequency of the cases in the modeling set whose ground truth (i.e., its actual label) is m.

Some rules for the combination of *a posteriori* probabilities given by several classifying systems are proposed in [74] and studied further in [71]. In the first, those rules are derived from a maximum *a posteriori* estimation in a multiple random variable scenario in an attempt of lightening the efforts of computing several joint PDFs. These rules are summarized in Table 3.2.

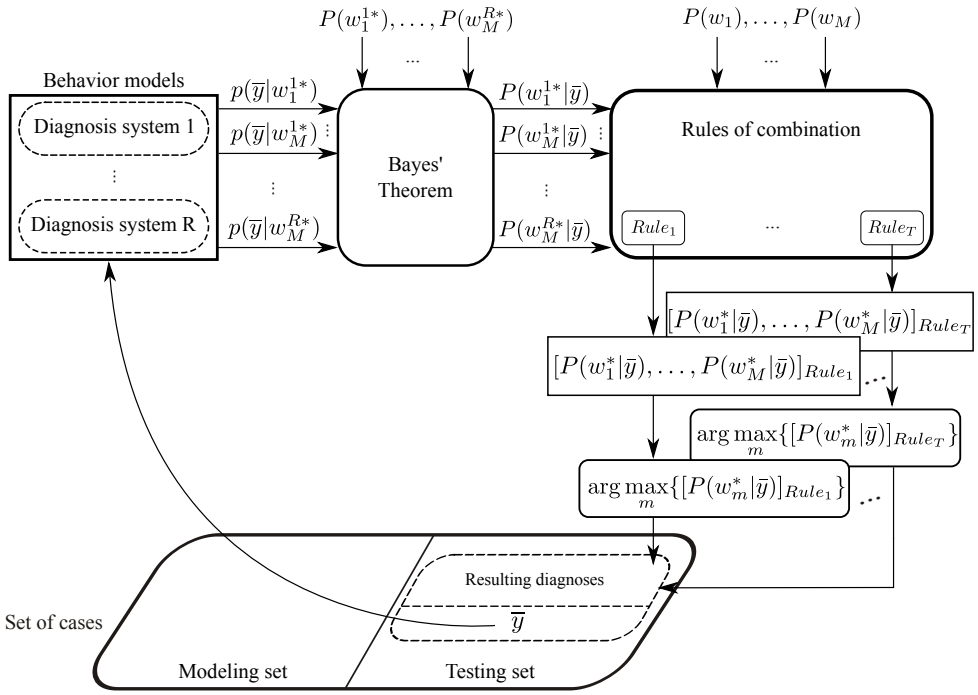


Figure 3.5: Proposed method for combining diagnosis systems. Stage 2: Combining the behavior models.

As this point, the fault cause with the maximum *a posteriori* probability is taken as the final diagnosis per each rule of combination, d_t. That is,

$$d_{Rule_t} = arg\max_m\{P(w_m^*|\overline{y})_{Rule_t}\}. \tag{3.6}$$

Note that a situation with $M^* < M$ means that there is at least one fault cause that have not been identified by any system. In this case, although $P(w_m)$ is not null, $P(w_m^*|\overline{y})$ will be zero at the end of the computation and it would be impossible for it to be finally diagnosed in consequence.

3.4 Performance analysis

In this section, the proposed method is assessed by combining two different diagnosis models, sharing a common automatic diagnosis technique. In the first test, which is carried out using

Table 3.2: Algebraic rules for the combination of *a posteriori* probabilities.

Rule	$P(w_m^*\|\overline{y})$
Product rule	$P(w_m)^{-(R-1)} \prod_{r=1}^{R} P(w_m^{r*}\|\overline{y})$
Sum rule	$(1-R)P(w_m) + \sum_{r=1}^{R} P(w_m^{r*}\|\overline{y})$
Max rule	$(1-R)P(w_m) + R \max_{r=1}^{R}\{P(w_m^{r*}\|\overline{y})\}$
Min rule	$P(w_m)^{-(R-1)} \min_{r=1}^{R}\{P(w_m^{r*}\|\overline{y})\}$
Median rule	$\underset{r=1}{\overset{R}{med}}\{P(w_m^{r*}\|\overline{y})\}$

simulation-based data, each model is provided by different troubleshooting experts. The second test is performed using data from a live cellular network, and both diagnosis models are built from a common set of training cases, using different ML techniques for knowledge acquisition.

The proposed method has been evaluated and compared to the baseline (standalone) diagnosis systems by means of the following figures of merit:

- Diagnosis error rate (DER): it is the ratio of problematic cases diagnosed as a fault cause different to the real one (misclassified cases), N_{MPC}, to the total number of problematic cases, N_{PC}. That is:

$$DER = \frac{N_{MPC}}{N_{PC}}. \tag{3.7}$$

- False positive rate (FPR): it is the number of normal cases diagnosed as problematic cases, (N_{FP}), to the total number of normal cases, (N_{NC}). That is:

$$FPR = \frac{N_{FP}}{N_{NC}}. \tag{3.8}$$

- False negative rate (FNR): it is the number of problematic cases diagnosed as normal cases, N_{FN}, to the total number of problematic cases, N_{PC}. This is the most critical metric, as it gives an idea on how often the diagnosis system interprets there is no problem when actually some cells are suffering from malfunctioning. That is:

$$FNR = \frac{N_{FN}}{N_{PC}}. \tag{3.9}$$

Given these definitions, an overall error rate (OER) may be defined as

$$OER = P_N \cdot FPR + P_{PR} \cdot (FNR + DER), \tag{3.10}$$

where P_N stands for the relative frequency of the normal cases and P_{PC} stands for the relative frequency of the problematic cases. This metric is useful to assess every method at a single glance. Since these figures of merit require the true cause to be known, the used testing set will include the real diagnosis.

3.4.1 Combination of diagnosis models devised by multiple experts

Scenario

In this test, cases are provided by an LTE RAN simulator [75]. This simulator considers an LTE network composed of 57 macro-cells evenly distributed in space and grouped into 19 three-sector-sites and the configuration parameters for simulating a normal cell functioning can be seen in Table 3.3.

Table 3.3: Simulation parameters for cells normal functioning.

Parameter	Configuration
Cellular layout	Hexagonal grid, 57 cells, cell radius 0.5 km
Transmission direction	Downlink
Carrier frequency	2.0 GHz
System bandwidth	1.4 MHz, 6 PRB
Frequency reuse	1
Propagation model	Okumura-Hata with wrap-around, Log-normal slow fading, $(\mu_{sf}, \sigma_{sf}) = (0, 8\,\text{dB})$, correlation distance = 50 m
Channel model	Multipath fading, model: ETU[1]
Mobility model	Random direction, 3 km/h
Service model	Full Buffer, Poisson traffic arrival
Base station model	Tri-sectorized antenna, SISO[2], $P_{TX_{max}}$ = 43 dBm, Downtilt = 9°, azimuth beamwidth = 70°, elevation beamwidth = 10°
Scheduler	Time domain: Round-Robin, frequency domain: best channel
Power control	Equal transmit power per PRB
Link Adaptation	Fast, CQI[3]-based, perfect estimation
HO	Triggering event = A3, HO margin = 3 dB, Measurement type = RSRP
Radio Link Failure	SINR[4]< -6.9 dB for 500 ms, [76]
Traffic distribution	Evenly distributed in space
Time resolution	100 ms, 100 TTI
Epoch & KPI time	100 s

[1] Extended typical urban
[2] Single input single output
[3] Channel quality indicator
[4] Signal-to-interference-plus-noise ratio

With this simulator, 1196 cases have been obtained. In this case, training cases are not needed since the diagnosis models have been defined by experts. It is assumed that a detection system is placed before the input of the diagnosis system, so that only the faulty cases are put under test, putting aside the cases belonging to a normal cause of functioning. Therefore, in this test, only the DER is taken into account.

In this scenario, six typical RAN fault causes have been considered ($M = 6$):

- *Excessive downtilt:* This situation takes place when the coverage area for a cell is too small, making the signal level in the edge of the cell to be too weak and causing a high number of HO failures. The quality of the signal in the surroundings of the cell is also decreased.
- *Coverage hole:* A cell has a coverage hole in some point inside its area when the power received by the user at this point from any cell is not enough to hold the service. This excessive attenuation can be caused by either obstacles or a bad radio frequency planning and it mainly produces a high number of call drops.
- *Inter-system interference:* This fault cause may occur due to other cellular networks, like UMTS. It is not always an easy issue to solve, since the fault usually comes from an outer system. This fault normally causes both the SINR and the average throughput decrease.
- *Too late HO:* a too late HO takes place if a radio link failure occurs while the UE is moving from one cell to another and the corresponding HO between these cells has not taken place yet. In that case, the UE will request the second cell a connection re-establishment using the PCI of the first cell and its common radio network temporary identifier (C-RNTI) in that first cell, which will alert the second cell a too late HO has occurred.
- *Excessive uptilt:* A cell suffers from excessive uptilt when its coverage area is larger than necessary, normally because of a bad configuration of the antennas. This situation can result in the overlapping of coverage areas from possibly non-adjacent cells, producing a high number of HOs and call drops in this cell and its neighbors.
- *Lack of coverage:* A user suffers from weak coverage when the SINR measured in the cell is below the minimum level needed to maintain a planned performance requirement because the received power is low.

The simulation parameters used to model these degradations are shown in Table 3.4, from [26], as well as the *a priori* probability of these causes to take place, given by the experts. As it can be seen, several values have been used for modeling a single fault cause, according to lighter and more severe degradation.

In this test, seven observable features or KPIs ($N = 7$) have been used to discern among this set of causes:

- *Retainability*, given as a percentage. This KPI quantifies the ability of the cell to hold the service once accepted by the admission control. It gives an idea on how often a user experiences a call drop.

Table 3.4: Parameters used for modeling fault causes in Section 3.4.1 and *a priori* probabilities for each cause, from [26].

Fault Cause	**Configuration**	$P(\omega_m)$
Excessive downtilt	Downtilt = [14, 16]°	0.18
Coverage hole	μ_{sf} = [49, 53] dB	0.09
Inter-system interf.	$P_{TX_{max}}$ = 33 dBm	0.1
(interfering source)	Downtilt = 15°	
	Azimuth beamwidth = [30, 60]°	
	Elevation beamwidth = 10°	
Too late HO	HOM = [6, 8] dBm	0.23
Excessive uptilt	Downtilt = [0, 1]°	0.21
Lack of coverage	$P_{TX_{max}}$ = [7, 10] dBm	0.19

- *HO success rate (HOSR)*, given as a percentage. This KPI measures the ability of the network to provide mobility to a user without losing its connection. It can be calculated as the ratio between the number of successful HOs and the total number of HOs.
- 95^{th} *percentile RSRP*, given in dBm.
- 5^{th} *percentile RSRQ*, given in dB.
- 95^{th} *percentile SINR*, given in dB. The SINR is defined as the ratio between the power of the desired data signal and the sum of the powers of all inter-cell interferences and the noise.
- 95^{th} *percentile distance*, given in km. This KPI measures the distance between users and their serving cell, expressed in km. It can be estimated attending to the transmission delay between them and gives an idea of the cell coverage area.
- *Average throughput for user k,* T_k, given in kbps. In LTE systems, the user throughput depends on the SINR experienced by the user through the following equation, [77],

$$T_k = (1 - BLER(SINR_k)) \cdot \frac{D_k}{TTI}, \tag{3.11}$$

  where the BLER is obtained from the users' SINR and D_k is the TB payload in bits of user k.

In order to show the impact that a proper modeling may have in the diagnosis performance of the proposed method, the proportion of cases used for the modeling to the testing set has been varied from 25% to 75%. To obtain more reliable results when the number of cases are scarce either in the testing or in the modeling set, a stratified Monte Carlo cross-validation (MCCV) of 50 repetitions has been performed per each step of the modeling-to-testing ratio. That is, the samples assigned to each set have been randomized in each repetition preserving the relative frequency of each cause in these subsets. Then, the resulting DERs have been averaged over the 50 repetitions.

The standalone classifiers

This test represents the usual case in cellular networks by which different troubleshooting experts have different beliefs on how a fault cause manifests itself in the network through a set of symptoms, thus having different diagnosis models. When deploying a diagnosis system in a network, according to the proposed method, instead of choosing one single model, the knowledge from both experts is fused by combining two diagnosis models. Furthermore, both diagnosis models comprise the six fault causes and the seven different KPIs described above. That is, $W^1 = W^2$ with $|W^1| = M$ and $N^1 = N^2$.

The artificial intelligence technique used for these tests is based on a fuzzy logic controller (FLC) [28]. FLCs first compute how much every feature of a given sample resembles or belongs to a set of predefined categories. In this process, called *fuzzification*, every sample gets an input label for each one of its features, which represents their corresponding feature-specific category, as well as an input scalar, which represents the degree of belonging to such category. The functions used for such transformation are called input *membership functions*. The labels obtained in this way are used to make up the antecedent of a rule-based *if ..., then ...* system. The consequent of these rules are output labels, which allow recovering a continuous-valued output following a process of *defuzzification*. In the defuzzification, a set of output membership functions is used, together with the output labels and the output scalars, which are computed from the input scalars. The (de)fuzzification processes allow addressing complex systems in a similar way as piecewise functions are used to model complex behaviors. These processes allow a rule-based mechanism to control a system which usually has a high number of continuous-valued inputs and outputs. These rules aim to encompass the way of thinking of a human, and thus, are one of the most intuitive mechanisms to integrate policies in a control system. In the context of self-healing, a FLC may be used as a diagnosis system by neglecting the defuzzification process, considering the output labels as the diagnoses. This way, the input membership functions and the rules of a FLC form a diagnosis model.

Table 3.5 shows the threshold values of the membership functions used for the KPIs assessed in this test. The lower threshold stands for the value below which a KPI is considered to be low and the upper limit stands for the value above which a KPI is considered to be high. On the other hand, Table 3.6 shows the *if ..., then ...* rules that complete each diagnosis model, given by each expert. These rules are built with an AND operator, meaning that a given rule only applies if the conditions for all the KPIs are simultaneously fulfilled. From left to right, each column below "KPI" in Table 3.6 corresponds to the KPIs shown in Table 3.5. H stands for a high value in that KPI and L for a low value. Regarding the numbering of the diagnoses, 1 means excessive downtilt; 2: coverage hole; 3: inter-system interference; 4: too late HO; 5: excessive uptilt and 6: lack of coverage.

Results

Table 3.7 shows the DERs computed when the *Max rule* is used for the combination (Table 3.2). In Table 3.7, the average DER and the rate of improvement are shown. This last rate represents the amount of repetitions (among the 50 that have been performed) in which the DER from the

Table 3.5: Diagnosis models for the diagnosis systems used in test 1: used thresholds.

KPI	**Thresholds**
Retainability	$[0.973, 0.996]$
HOSR	$[0.899, 0.989]$
RSRP [dBm]	$[-76.9, -72.4]$
RSRQ [dB]	$[-18.8, -18.2]$
SINR [dB]	$[13, 14.5]$
Throughput [kbps]	$[96.2, 111.67]$
Distance [km]	$[0.838, 0.88]$

Table 3.6: Diagnosis models for the diagnosis systems used in test 1: used rules.

Diagnosis model 1								**Diagnosis model 2**							
KPI							**Diag.**	**KPI**							**Diag.**
L	L	H	L	-	H	L	1	-	-	H	L	-	H	L	1
H	H	-	L	L	H	L	1	L	-	H	H	-	L	H	2
L	-	-	H	H	-	H	2	L	-	H	H	H	-	H	2
L	-	-	H	L	L	H	3	L	-	-	H	L	L	H	3
L	L	H	-	L	L	H	3	L	-	H	-	L	L	H	3
-	-	H	H	H	H	-	4	L	-	H	H	L	L	-	3
-	H	H	-	H	H	-	4	L	H	-	H	L	L	-	3
H	-	H	-	H	H	-	4	-	H	H	H	L	L	H	3
-	-	H	-	H	H	H	4	L	L	-	-	-	H	H	4
H	H	-	-	H	-	H	4	L	L	-	L	-	-	H	4
H	H	-	L	-	-	H	4	L	L	H	-	H	L	-	4
H	H	-	-	-	H	H	4	L	L	H	-	H	-	H	4
H	H	H	-	-	-	H	4	L	L	L	L	L	L	-	4
-	-	H	L	H	-	H	4	-	-	L	H	-	L	H	5
-	-	H	L	-	L	H	4	H	H	-	H	-	L	H	5
L	L	H	L	-	-	H	4	H	H	L	-	L	L	H	5
-	-	L	-	L	L	H	5	-	-	-	L	L	H	L	6
-	-	L	-	L	H	L	6								
L	-	-	L	L	H	L	6								
-	H	-	L	L	H	L	6								
H	H	-	-	L	H	L	6								
L	L	L	L	L	-	L	6								

ensemble method is lower than the best one provided by the baseline diagnosis systems. With a 25% of modeling-to-testing ratio only 60% of the iterations shows a better ensemble DER than the ones from its base diagnosis systems, showing, therefore, little improvement in the average DER. This result highlights how the scarcity of cases for modeling impacts on the classifying performance of the ensemble. However, if the number of cases used for modeling is doubled, 98% of the iterations shows a better DER, which results also in a lower average DER. In case the modeling-to-testing ratio is set to 75% every DER provided by the ensemble method is lower

than the lowest provided by its components, reaching a 5.34% of average DER.

Regarding the DER of the standalone diagnosis systems, it can be seen how these are held over the modeling-to-testing ratio, given the preservation of the relative frequency of each cause in the modeling and testing sets.

Table 3.7: Results of test 1.

	Modeling-to-testing ratio		
	25%	50%	75%
Diagnosis syst. 1, average DER	13.81%	13.7%	13.65%
Diagnosis syst. 2, average DER	16.34%	16.13%	16.3%
Ens. Method: Max rule average DER	8.29%	5.92%	5.34%
Rate of improvement	60%	98%	100%

3.4.2 Combination of different diagnosis systems on a live network

Once the proposed method has been tested with cases provided by a simulator, a second test with cases from a real live LTE network has been performed. In this test, the diagnosis models built from two different ML algorithms have been combined.

Scenario

A real LTE network composed of more than 8000 different cells providing coverage to almost 4 million people has been analyzed. Its vastness makes many different cells to coexist and also a wide variety of problematic causes to come up. Table 3.8 summarizes the main parameters of the network. Among all the available candidates, 45 random cells have been chosen to represent the network behavior. These cells have been monitored for almost 6 days on average and their KPIs have been stored in an hourly basis. Taking into account that the state of a single cell varies substantially throughout the day due to the traffic fluctuation, several cases have been stored from each cell at different hours, resulting in a total of 14692 cases. Once these cases were gathered, they were all labeled by the experts, distinguishing four groups of cases ($M = 4$): three kinds of problematic patterns and the normal cell functioning. The causes of malfunctioning that were found are:

- *Overload*: This fault cause is mainly distinguished by a high number of RRC connections in the cell, which makes the CPU processing load and the number of HO attempts raise consequently. The accessibility and retainability KPIs also hold values quite below the ones for a cell with normal functioning.
- *Lack of coverage*: This issue can be identified based on the number of bad coverage evaluation reports, which should be noticeably high.

Table 3.8: Main parameters of the real LTE network used in test two.

Parameter	Configuration
Network Layout	Urban area
Number of cells	8679
System bandwidth	10 MHz (50 PRBs)
Frequency reuse factor	1
Max. Transmitted Power	46 dBm
Max. Transmitted Power of UE	23 dBm
Horizontal HPBW (half-power beam width)	65°
HO margin	3 dB
KPI Time Period	Hourly
Number of observed cells	45
Number of days under observation	6 days per cell (on average)
Size of the dataset	14692 labeled cases

- *Non-operating cell*: In this case, and only if the cell is reporting any KPI measurement, most of the reported measurements should be near zero: the retainability, the accessibility, the number of performed HO, the number of RRC connections or the number of coverage reports.

The *a priori* probability of occurrence of each class has been computed as the relative frequency of the cases within this selection, Table 3.9. From this table it should be noted that there are more faulty cases than healthy ones. This is because a previous non-perfect faulty cases detecting stage has been applied, which bypassed some normal cases that now are to be diagnosed as such.

At this point, 20% of the total number of cases (holding the proportion shown in Table 3.9 between them) were used as a training set for the ML algorithms and the rest were used to conform the modeling and testing sets in a ratio that, as in Section 3.4.1, was varied along the test.

Table 3.9: Prior probability of occurrence for the causes considered in test two.

P(Overload)	P(Lack of cov.)	P(Non-operating)	P(Normal)
0.01	0.22	0.47	0.3

In this test, six of the most representative KPIs in an LTE network have been chosen to discern between the possible diagnoses, having $N = 6$:

- *Retainability*: described in Section 3.4.1.
- *Accessibility*: it is used to show the percentage of connections that have got access to that cell over the KPI time period. A low value in this KPI means that many connections have been blocked during the access procedure.

- *Number of RRC connections*: it is the number of successfully established RRC connections. Related to the *Accessibility* KPI, it gives an idea of the amount of users served by the cell.
- *Number of ping-pong HOs*: this KPI counts the number of ping-pong HOs that takes place in the cell over the measurement time period. A high value in this KPI may mean a bad configuration in the HO policy, as the number of connections that goes back and forth over a cell and its neighbors is high for a single call.
- *Number of bad coverage reports*: it counts the number of times a cell is notified that the UE measured a signal level in which the requirements for the event A2 takes place [56]. That is, the measured signal level is under a certain threshold.
- *CPU average load*: it is the average CPU load due to the processes carried out by the cell over the KPI time period.

The standalone classifiers

In this test, the two used standalone classifiers share a similar diagnosis system, an FLC, which diagnoses the cases according to *if* ... , *then* ... rules. The difference resides in the algorithms used to learn the rules that they apply during the diagnosis process. The first is a genetic algorithm [28]. In genetic algorithms, three main processes may be distinguished: reproduction, by means of which new individuals are created by either mutation or combination of the previously existing; evaluation, or the calculation of the probability of each individual to survive and reproduce, and selection, a process in which some individuals are chosen to survive and reproduce based on the results from the evaluation stage. The second algorithm for rule learning is proposed in [29]. This ML technique first take a case from the training set and derives the rule that covers it. Then, they look for the cases covered by this rule and score the rule according to the number of covered cases. New incoming cases are taken until the training set is completely explored. Provided this set of scored rules, the algorithm then fuses them into a lower number of rules in an attempt of maximizing the number of cases (and therefore, the score) covered by the resulting fused rules. In these tests, it is assumed that not only faulty cases, but also some normal cases are inputs for the diagnosis stage. This can happen when there is no detection system before the diagnosis system or in the realistic situation in which the detection system has a given probability of error. As in Section 3.4.1, both systems take as a possible output all the presented diagnoses making use of the six KPIs shown above. Table 3.10 shows the thresholds used for these KPIs to consider them high or low and Table 3.11 shows the rules that each ML algorithm has derived from the testing set. As in Table 3.6. The KPIs are sorted in the same way as in Table 3.10 and the numbering of the diagnoses are 1: CPU overload; 2: lack of coverage; 3: non-operating cell and 4: normal functioning.

Results

In this test, the modeling-to-testing ratio has been varied from 10% to 90% in steps of 10. In this case, a 10-iteration MCCV has been done in each of these steps. The resulting metrics (DER, FPR, FNR and OER) have been then averaged. Table 3.12 shows the metrics that result of

Table 3.10: Diagnosis models for the diagnosis systems used in test 2: used thresholds.

KPI	Thresholds
Retainability	$[0.99, 0.997]$
Accessibility	$[0.992, 0.998]$
Number of RRC Connections	$[5846, 20703]$
Number of ping-pong HO	$[18, 83]$
Number of bad cov. reports	$[217, 1070]$
CPU average load [%]	$[22.5, 34.45]$

Table 3.11: Diagnosis models for the diagnosis systems used in test 2: used rules.

Diagnosis model 1: use of [28] for rule learning							**Diagnosis model 2: use of [29] for rule learning**						
KPI						**Diagnosis**	**KPI**						**Diagnosis**
H	H	-	-	L	L	1	H	H	L	-	L	L	1
H	-	H	H	L	L	1	H	H	-	H	L	L	1
-	H	H	L	-	L	2	H	-	H	H	L	L	1
H	-	-	L	-	H	2	-	L	H	L	-	H	2
H	-	-	L	H	-	2	-	H	H	-	H	H	2
H	-	H	L	-	-	2	-	-	H	L	H	H	2
-	L	H	L	-	H	2	L	-	H	-	H	H	2
H	H	-	-	H	-	2	H	H	H	-	H	-	2
-	H	H	-	H	-	2	H	L	-	L	L	H	2
-	-	H	L	H	-	2	H	-	H	L	-	H	2
H	-	H	-	H	H	2	H	-	H	L	H	-	2
H	-	H	-	H	H	2	-	L	L	L	L	L	3
-	L	L	L	L	L	3	L	L	L	L	H	-	4
L	L	-	-	H	H	4	-	L	L	L	H	H	4
L	L	L	-	H	-	4	L	L	L	-	H	H	4
L	L	H	L	L	L	4	L	-	L	L	H	H	4
L	L	L	-	-	H	4							
L	-	L	-	H	H	4							
-	-	L	L	H	H	4							
L	L	L	H	-	-	4							

using a proportion of 60% in the modeling-to-testing ratio. This ratio has proved to minimize the values of all the metrics in this test.

As it can be seen in Table 3.12, in most cases, the combined diagnosis system outperforms the standalone diagnosis systems. Concretely, the median rule achieves the lowest OER with a 5.39%, in comparison with the 8.16% and the 11.43% of the base systems. This achievement does not come directly from its DER, which is nearly the same than the one from the genetic algorithm, but comes from a more accurate identification of the normal state. This entails a significant reduction in the FPR, and especially, in the FNR, which is less than half the FNR provided by the genetic algorithm.

These results can also be seen in the normalized confusion matrices from the diagnosis methods. Figure 3.6a shows the normalized confusion matrix for the FLC using [28] for rule learning; Figure 3.6b shows the confusion matrix given that the technique from [29] was used to learn the rules and Figure 3.6c shows the matrix from applying the median rule with a 60% of modeling-to-testing ratio in the ensemble method. In these matrices, the elements from the fourth column (excluding the main diagonal) account for the false negatives and the elements from the fourth row account for the false positives. It can be seen how the elements from the main diagonal are reinforced in the ensemble method and how only those diagnoses which are mistaken by both baseline systems are slightly inherited by the latter. Figure 3.6c also shows graphically how the FPR and FNR decreased with respect to those from the standalone systems.

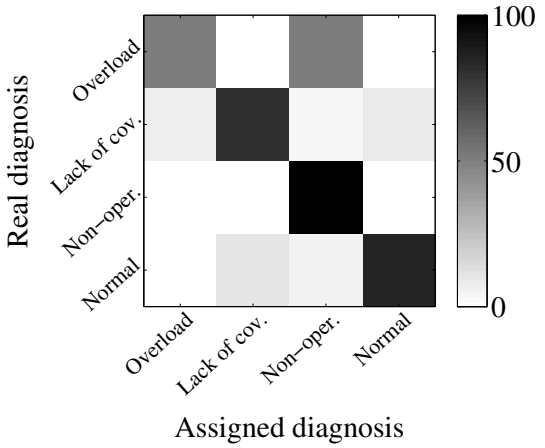
Table 3.12: Results of test 2.

	DER	**FPR**	**FNR**	**OER**
Training: [28] for rule learning	2.62%	16.91%	6.47%	11.43%
Training: [29] for rule learning	1.87%	16.61%	2.68%	8.16%
Ensemble method				
Product Rule	2.6%	12.21%	1.32%	6.2 %
Sum Rule	**1.78%**	11.55%	**1.25%**	5.59%
Max Rule	**1.78%**	11.51%	**1.25%**	5.57%
Min Rule	2.05%	11.42%	1.4%	5.84 %
Median Rule	**1.78%**	**10.67%**	1.34%	**5.39%**
Majority Vote Rule	**1.78%**	11.23%	**1.25%**	5.49%

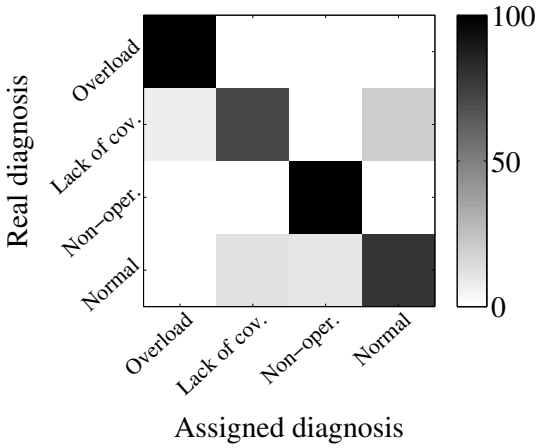
3.5 Conclusions

A method to combine fault diagnosis systems has been presented and tested in a cellular network environment. The proposed method allows to merge systems based on different artificial intelligence techniques in order to enhance the results of the individual systems. In addition, multiple diagnosis models designed by different troubleshooting experts can be fused into a model that combines the knowledge of all the experts. Likewise, models based on learning algorithms using different training datasets can also be combined into an enhanced model. This combination of multiple systems and models to generate a single one not only improves the diagnosis accuracy, but also solves the main problem that operators have in the selection of diagnosis systems.

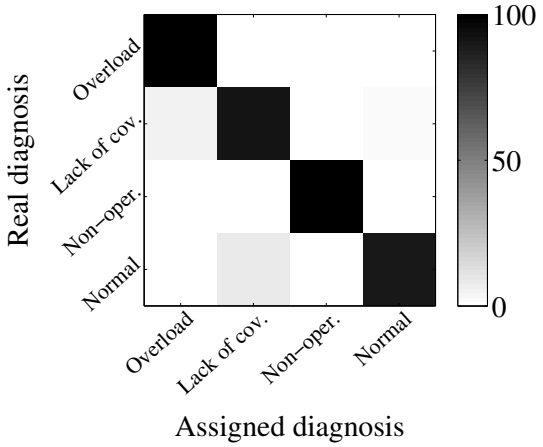
The method has been tested in two scenarios: cases provided by an LTE RAN simulator and cases gathered from a live LTE network. Likewise, two use cases have been assessed: the combination of models designed by two different diagnosis experts and the combination of two diagnosis systems that use different learning algorithms. The proposed method has proved to outperform the behavior of its base components in both tests in terms of the DER, and even more, in terms of the FNR.



(a) Normalized confusion matrix of the baseline diagnosis system using [28] for rule learning.



(b) Normalized confusion matrix of the baseline diagnosis system using [29] for rule learning.



(c) Normalized confusion matrix of the proposed method for combination using the median rule and a modeling-to-testing ratio of 60%.

Figure 3.6: Normalized confusion matrices for the second test.

Chapter 4

Dimensionality reduction for self-healing

In this chapter, dimensionality reduction is described and assessed as a means to overcome some of the most pressing needs of self-healing, like its full automation, relieving troubleshooting experts of identifying which KPIs to take as their inputs or the reduction of the storage needs in network databases. To that end, different families of techniques are explored, showing the benefits and trade-offs that their usage entails in current cellular networks.

This chapter is organized as follows. In Section 4.1, the current work on dimensionality reduction in the field of cellular network management is described, also briefly summarizing its impact on other fields. Section 4.2 outlines the problem formulation. Next, section 4.3 describes the two families of techniques for dimensionality reduction; namely, feature selection and feature extraction, and contextualizes them in the scope of self-healing in cellular networks. Then, in Sections 4.4 and 4.5, an unsupervised and a supervised method for feature selection are proposed and assessed, respectively. Following, section 4.6 elaborates on the insights of different techniques for feature extraction. Following the analysis of the techniques until this point, in Section 4.7, a general framework for dimensionality reduction devoted to the enhancement of automated cellular network management is described and analyzed. To conclude, Section 4.8 summarizes the main conclusions of this chapter.

4.1 Related work

In recent years, and due to the explosion of available data in different research and industrial activities, dimensionality reduction has become an essential task of data preprocessing. This is due to the fact that with an increasing number of dimensions, data become more and more sparse as the number of features approaches the order of magnitude of the number of samples, which is a problem for any method that requires statistical significance. For example, in the field of computer vision, an image may be seen as a heavy container of scarce useful information and a lot

of useless data. This led to the usage of dimensionality reduction techniques in this field [78]. In medicine, it is often needed to identify and isolate the subset of genes behind a certain pathology among the thousands of possible candidates. Again, techniques for dimensionality reduction are often used for this task [79, 80]. Also, in industrial manufacturing processes, dimensionality reduction is used to help diagnosis techniques to identify possible faults in the products [81]. In the field of communication networks, dimensionality reduction has also been used to complement some tasks, like traffic flow classification. This is the case of [82] and [83], in which ML techniques for classification are used to categorize Internet traffic.

Regarding cellular network management, some steps have been taken towards the full automation of SON through dimensionality reduction, despite until now, few works have been reported in this field. In [84], network optimization benefited from a supervised correlation-based technique for KPI selection, which is used to help identifying different traffic patterns in a UMTS (universal mobile telecommunications system) network. In [85], a supervised technique based on a genetic algorithm is proposed for KPI selection in a problem of automatic diagnosis, thus being devoted to self-healing tasks. However, given their supervised nature, the works in [84] and [85] rely on the availability of a network status label attached to every network observation, which seldom is present. Actually, most of the performance management data stored in nowadays cellular networks are unlabeled, due to the high amount of time required for network experts to analyze and document every network fault, together with the pressing need for a fast response. Besides, the technique described in [85] for KPI selection implements a *wrapper-type* method for dimensionality reduction. This kind of techniques embed a classification algorithm in their processing chain and aim at determining the set of features that minimizes its misclassification rate. Thus, the resulting selection is classifier-dependent, providing poorer results when a different classifier from the one it is embedded is subsequently used. Consequently, in this chapter, different *filter-type* techniques for dimensionality reduction are proposed. Contrary to *wrapper-type* techniques, *filter-type* methods for dimensionality reduction are particularly efficient, being less prone to over-fitting and also providing a KPI selection suitable for any kind of subsequent classification or prediction technique.

Furthermore, feature extraction is described and applied in the field of cellular network management in this chapter, which, to the author's knowledge, was a field still to be explored. First, in a standalone manner, and then, together with feature selection techniques, feature extraction is included as part of a novel framework for dimensionality reduction in self-healing tasks.

4.2 Problem formulation

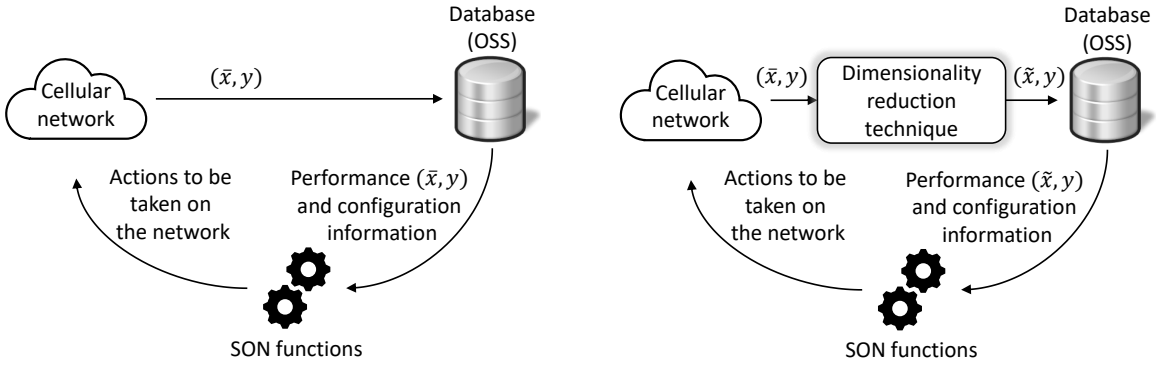
This chapter aims to delve into how self-healing, and particularly, RCA, can benefit from dimensionality reduction techniques in its way towards its full automation and its performance improvement.

In a cellular network, performance information is monitored and periodically stored in a centralized database: the OSS database (see Figure 4.1a). This information usually takes the shape of metrics (KPIs) derived from network event counters and may be expressed as a set

of N-dimensional samples, $\bar{x} = \{x_1, x_2, \ \ldots, x_N\}$, being N the number of the monitored KPIs; that is, the number of features. A feature could be, for example, the number of user connection attempts registered by a base station in a given time period. Usual values for N (several thousands), together with the high number of network elements simultaneously monitored, entail the storage of a huge amount of performance data every day and pose a performance issue for subsequent RCA functions. Together with $\bar{x}$, a possible additional label, y, often referred to as the ground truth label, may be attached. This label corresponds to the network state under which each sample was gathered. A ground truth label could be a coverage hole, a problem of interference or a network overload. The availability of this label allows using supervised techniques for dimensionality reduction [84, 85] and supervised techniques for automatic diagnosis [28, 29]; whereas its lack forces network experts to only use unsupervised techniques for both dimensionality reduction and RCA functions [24].

In the context of RCA, self-healing functions take the shape of classifying systems, Figure 3.1a, also called automatic diagnosis systems. Specifically, in a classification problem, a dimensionality reduction technique is a procedure that either finds or generates a set of q features (with $q << N$) that represents a sample with the minimum loss of useful information. That is, after dimensionality reduction has taken place, the classifier takes $\tilde{x} \in \mathbb{R}^q$ as its input, instead of $\bar{x} \in \mathbb{R}^N$ (see Figure 4.1b).

Thus, in the same way that dimensionality reduction techniques allow finding the set of symptoms (or genes) behind a particular disease, they pursue narrowing down the amount of features that automatic diagnosis systems have to deal with, where each feature stands for a network KPI. In particular, reducing the number of symptoms to be evaluated to diagnose a certain disease, identifying the truly relevant ones, has two great advantages. On the one hand, it accelerates the identification of this disease, having to only evaluate a smaller number of symptoms, and on the other hand, it reduces the likelihood of mistaking it for another disease. These two benefits appear in the context of RCA functions in terms of a reduced computing time for diagnosis and a reduced DER.



(a) Typical gathering of performance information in a cellular network (SON scheme).

(b) Gathering of performance information in a cellular network including a stage for dimensionality reduction (NG-SON scheme).

Figure 4.1: Comparison of schemes for network performance information gathering and use in SON functions.

In the next section, the different families of techniques for dimensionality reduction are described, together with their benefits and drawbacks in the field of self-healing functions.

4.3 Overview of dimensionality reduction techniques in the context of Self-Healing

Dimensionality reduction does not encompass a single family of ML techniques, but two: feature selection and feature extraction, where each of them presents its own benefits and drawbacks. These are explained in this section, as well as the implications that these families of techniques have when included to complement RCA functions.

4.3.1 Feature selection

Within dimensionality reduction, feature selection consists in identifying the subset of features which are considered as the most relevant according to a certain criterion or target. It is usual that this kind of techniques are implemented as a feature scoring algorithm with a given threshold. In this way, a feature selection technique acts as a filter: only those attributes above the threshold pass through the filter, whereas the others are discarded and considered as non relevant. Usually, these techniques are differentiated regarding how these scores are defined. For some of them, the score is based on the proportion of information contained in a given feature in relation to the total amount of information in the data. This is the case of the F-test, in which the amount of information is usually measured in terms of the variance of the data [86]. Other techniques, however, rely on the ability of a given feature to preserve the multi-cluster structure of the data; that is, its ability to preserve the separability of groups of samples making up differentiated clusters. This is the case of novel feature selection techniques like the Laplacian score [87].

Besides, within feature selection techniques there is a wide variety of both supervised and unsupervised methods. The first use the information contained both in the features and in the ground truth label. Supervised methods try to find the features whose behavior varies the most depending on these labels (e.g., [85]). Conversely, unsupervised methods only rely on the information contained in the features themselves. The usage of one or another approach in practice depends on the criteria followed by network experts to manage the network. In case that no ground truth labels were available (i.e., given an unlabeled dataset), only an unsupervised approach could be followed. However, if ground truth labels are available, different paths may be followed. If network experts simply want the DERs to be minimized, supervised techniques for feature selection are usually preferred. According to [80], despite the performance of unsupervised techniques is close to that of supervised techniques, the latter currently outperform the former. Nevertheless, unsupervised techniques may allow finding features which reveal underlying network states that were not initially considered among the set of ground truth labels. This is especially relevant for networks currently being extended, in which new functionalities are included, leading to new possible network states/issues. Following this reasoning, the usage of supervised techniques would then be recommended in stable and mature networks. In such case, network experts could safely prioritize the minimization of DERs at the expense of the discovery

of undetected faults.

A scheme for a feature selection technique is shown in Figure 4.2. This technique takes as its input the pair $(\bar{x}, y)$, where the ground truth label may be present (to be used by a supervised technique) or not. As a result, a list of selected KPIs is provided and used to filter $\bar{x}$, leading to $\hat{x}$. Now, $\hat{x}$ is made up of features x_i to x_j, being a subset of the original N features.

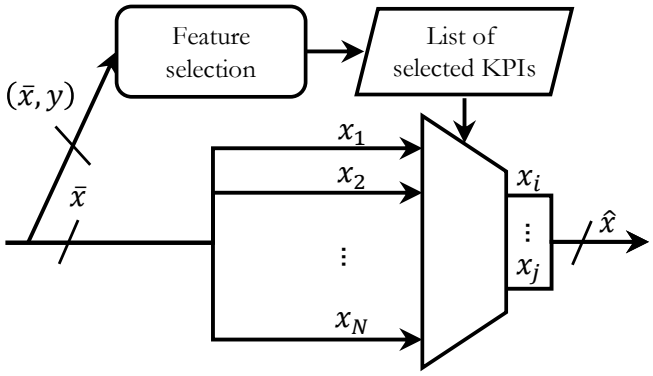


Figure 4.2: Scheme for feature selection in the context of RCA functions. The upper branch represents the model creation during the learning phase, whereas the lower branch represents the filtering process, according to the KPIs that have been selected.

The fact that the selected features make up a subset of the original feature set brings a big advantage from the point of view of human experts: the resulting selected KPIs preserve the meaning of those at the input. This way, they are still comprehensible by troubleshooting experts. Related to this, automatic selection of KPIs entails a second advantage. If the most relevant KPIs for a certain purpose are known in advance, then the monitoring of all the indicators considered as non relevant might be disabled. This would lead to a significant reduction in the storage needs of the network databases, even though the selection should be reassessed every certain time.

4.3.2 Feature extraction

Unlike the selection approach, with feature extraction a new set of features is built from the original feature set, so that the number of the resulting new features is lower than the number of the original ones. From a mathematical point of view, feature extraction is a tool that allows projecting the original features onto a more convenient and reduced basis. New features are built in such a way that they retain as much information as possible from the original feature set. In the context of cellular networks this means that a new set of synthetic KPIs are built upon the combination of the original ones.

Feature extraction techniques may be classified into linear and non-linear techniques. The first group is made up of techniques whose output are weights to be directly applied to the original feature set, so that the new features result from a linear combination of the first. In non-linear feature extraction techniques, however, a non-linear transformation is applied to the original feature set. These groups of techniques for feature extraction are further described in Section 4.6.

Feature extraction poses an interesting advantage when compared to feature selection. Given that the new synthetic features are computed based on the combination of those from the original feature set, such indicators could contain a much higher amount of useful information than if a technique for feature selection was used. That is, the original feature set could be compressed into a smaller set of rich information synthetic KPIs. This is specially useful for many ML techniques, which are commonly used as SON functions, and which usually suffer from issues related to the sparsity of data along the features they use.

Feature extraction, however, poses some non-negligible drawbacks. First, the synthetic KPIs generated this way are not comprehensible by troubleshooting experts, making it difficult to relate them to a given network state. And second, every time a new sample for a synthetic KPI is to be computed, a value for all of its components is needed, needing all of them to be permanently monitored.

A scheme for feature extraction is shown in Figure 4.3. In this case, given that the vast majority of feature extraction techniques is unsupervised, only $\bar{x}$ is used as its input. The result is a model for KPI transformation. This model may differ depending on the particular technique being used. For example, it may consist of a matrix describing how the indicators at the input are linearly combined to produce the synthetic ones. It may also describe non-linear functions to be applied over the former to get the latter. In any case, the resulting output takes the form of $\tilde{x}$: a vector with less dimensions that $\bar{x}$ and which is comprised of features $\tilde{x}_1$ to $\tilde{x}_E$. The tilde in $\tilde{x}_1$ and $\tilde{x}_E$ highlights the fact that these features are different from those at the input, not being a subset of the latter.

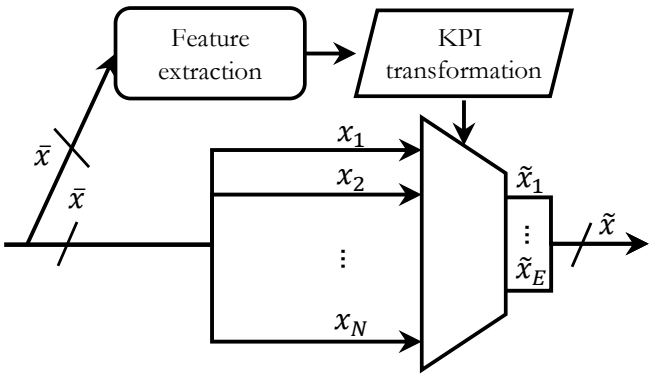


Figure 4.3: Scheme for feature extraction in the context of RCA functions. The upper branch represents the model creation during the learning phase (in this case, the KPI transformation to be applied), whereas the lower branch represents the application of such KPI transformation.

4.4 Feature selection: unsupervised approach

Despite the number of collected KPIs tends to be very high, it is quite unlikely that the network state under which a certain case was observed (i.e., the ground truth) is saved as well in network databases. Hence, developing an unsupervised technique for feature selection to identify the most relevant KPIs appears as a pressing need.

In this section, an unsupervised technique for feature selection is proposed and assessed with data gathered from a live cellular network, complementing RCA tasks and showing its ability to both reduce the network storage needs and subsequent DERs while fully automating the diagnosis task.

4.4.1 Proposed method

The unsupervised method for feature selection proposed in this section is made up of two main steps: a clustering stage and a supervised feature selection stage (see Figure 4.4). Raw data are gathered in an OSS database, containing a set ($\mathbf{Z}$) of M unlabeled samples of dimension N. Within the clustering stage, the first step is the data pre-processing. In this case, a standard score normalization is applied, resulting in $\mathbf{X}$. A standard normalization stands for the removal of the average of each feature and the normalization of its standard deviation to one.

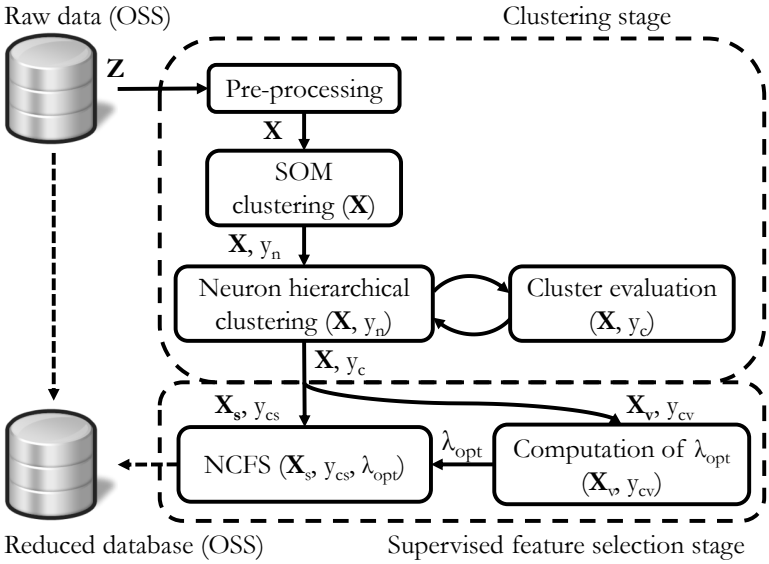


Figure 4.4: Proposed method for unsupervised KPI selection.

The next step consists of clustering the normalized data, so that each of the samples $x_i \in \mathbf{X}$ (where i is the sample index) gets a cluster label, $y_{ic} \in y_c$, which will be used in the next step. In this case, a neural network approach has been followed [24], given its excellent results when used in pattern recognition and diagnosis tasks in cellular networks. The method in [24] consists of a clustering stage followed by a diagnosis stage. In this thesis, only the clustering phase has been performed, not dealing with the interpretation of the cluster labels from its subsequent diagnosis phase. The clustering technique in [24] consists, in turn, of two steps. First, a regular SOM clustering is performed, using a sequential training after a fast batch training. After this step, every sample x_i gets a label $y_{in} \in y_n$, representing which neuron got a higher degree of activation. Next, a hierarchical Ward clustering is iteratively performed over the neurons, starting from a number of clusters given by the best Davies-Bouldin index and being merged if the Kolmogorov-Smirnov test over the features along the clusters does not show enough dissimilarity. As a result, every sample x_i gets a final cluster label, $y_{ic} \in y_c$.

At this point, the labels in y_c are a pattern identifier, which does not need to be interpreted, unlike in [24]. Thus, it is not necessary to relate these patterns to a given network state, and thus, no intervention from an expert is needed. The labels in y_c, together with their corresponding data samples $\mathbf{X}$, can now be forwarded to a supervised feature selection technique, which will not care about the actual meaning of y_c. The subsequent supervised technique will select those KPIs that best differentiate the cluster labels y_c.

In this section, a state-of-the-art technique has been chosen due to its validity for multi-class problems and its insensitivity in the number of irrelevant features. The neighborhood component feature selection (NCFS) [88] is a method which learns a feature weighting vector, w, maximizing the expected leave-one-out classification accuracy in a nearest neighbor (NN) classification using a regularization term. Regularization (controlled by λ) contributes to reduce over-fitting, making the parameters of the model (in this case, w) shrink, lowering the expected variance of the predictions and thus, improving the generalization of the prediction model. The optimal value for λ, λ_{opt} (Eq. (4.1)), is the value of λ which leads to a w that minimizes the average misclassification error of the inner NN classifier in NCFS. Despite several methods may be used to compute λ_{opt} [89], a classical k-fold cross-validation approach is followed in this paper. To do so, the misclassification error of the inner NN classifier in NCFS is computed (Eq. (4.2)) and averaged over the k folds throughout a range of values of λ (Eq. (4.1)). λ_{opt} is chosen as the value of λ for which this average error is minimized. In the method proposed in this section, the pair $(\mathbf{X}, y_c)$ is partitioned into two disjoint sets: $(\mathbf{X_v}, y_{cv})$ and $(\mathbf{X_s}, y_{cs})$, where the subindex v stands for validation and s stands for selection. The first pair is used for the computation of λ_{opt} via cross-validation. The classification loss (or penalty) for fold k conditioned to a certain λ, $L_{k,\lambda}$, is shown in Eq. (4.1), where $y_{cv_{k,t}}$ stands for the labels of the samples used in the test stage in fold k in the cross-validation procedure; $\hat{y}_{i,\lambda}$ stands for the predicted label of sample i conditioned to a certain value of λ; y_i is the actual label of sample i, which is eventually contained in y_{cv}; F equals 1 when its argument is true and 0 when it is not, and w_i is the weight of sample i so that they sum to their corresponding prior class (cluster) probability, which are also normalized to sum 1.

$$\lambda_{opt} = \arg\min_{\lambda} \left(\frac{1}{k} \sum_{k} L_{k,\lambda} \right) \tag{4.1}$$

$$L_{k,\lambda} = \sum_{\substack{y_i \in y_{cv_{k,t}} \\ y_{cv_{k,t}} \subset y_{cv}}} w_i F\left\{\hat{y}_{i,\lambda} \neq y_i\right\}. \tag{4.2}$$

Finally, the NCFS technique is applied over the pair $(\mathbf{X_s}, y_{cs})$ using λ_{opt} as the regularization parameter, obtaining w as the outcome. The highest n weights of w correspond to the best n features of $\mathbf{X_s}$, and therefore, of $\mathbf{X}$.

4.4.2 Performance analysis

Experiment setup

To evaluate the performance of the proposed method in the field of RCA within SONs, a test has been carried out using a 359-sample dataset, gathered from a live cellular network [90]. Given the lack of 5G commercial deployments at the moment of writing, an LTE RAN has been assessed instead, without loss of generality. Each sample is composed of 286 RAN KPIs and a ground truth label, accounting for the network state under which the sample was collected. In particular, four different labels are differentiated: high traffic (referred to as C_1), no traffic (C_2), high CPU utilization (C_3) and low coverage (C_4). These labels are only used in order to compute the DER, as a means to quantify the validity of the chosen indicators. Thus, the database is treated as *unlabeled* during the feature selection procedure. Regarding the available KPIs, they range from mobility, accessibility and CPU load, to counters related to retainability and throughput.

In this test, the performance of different feature selection techniques is compared by evaluating the DER resulting from using a diagnosis technique which takes as its input only the KPIs chosen by such selection techniques. The KPIs considered in this test are: all of them, to set a baseline, showing the case when no selection is made; a subset of them, given by troubleshooting experts' recommendation (abbreviated hereafter as TE), and a subset of them, given by three unsupervised techniques for feature selection, among which the proposed method (abbreviated hereafter as UP, for unsupervised technique for feature selection) is present.

As an NN classifier is embedded in NCFS, it is expected that the performance of a subsequent kNN diagnosis algorithm is specially enhanced. Thus, to show the capabilities of the proposed method in a more general scenario, an alternative diagnosis technique has been used instead: a multi-class linear discriminant analysis (LDA) classifier [91].

In order to show the validity of the whole proposed method (Figure 4.4), it has been compared against two state-of-the-art unsupervised feature selection techniques [87, 92], which, as of today, have not been tested in the field of self-healing in cellular networks. In [87], a method called Laplacian score (abbreviated hereafter as LS) is computed for each feature to reflect its locality preserving power; in [92], the selected features are those that better preserve the multi-cluster structure of the data (abbreviated hereafter as MC), making use of manifold learning (see Section 4.6.1) and L1-regularized models.

The dataset has been split into three subsets, following a 0.4, 0.4, 0.2 proportion. The first, the *selection set*, is devised for the KPI selection (being $\mathbf{X}$ in Figure 4.4); the second, the *training set*, is used to train the diagnosis method, and the third, the *test set*, is used for the computation of the DERs. Note that only the indicators selected after using the first subset have been used in the training and test subsets. Given the small amount of samples in the original dataset, a stratified MCCV with 100 iterations has been performed, shuffling the samples assigned to each subset in each iteration while preserving the relative frequency of each label in each split.

For this test, the SOM clustering used a 5×5-neuron grid, with 200 epochs for the batch training and 500 epochs for the subsequent sequential training. Regarding the NCFS algorithm, in each iteration λ_{opt} was computed with λ ranging from 1 to 10 by means of a 5-fold cross-

validation. The parameters used for NFCS were: $\sigma = 1$ (kernel width) and $\eta = 10^{-4}$ (gradient tolerance). For LS and MC, 5 neighbors were considered for the computation of their weighting matrices through a kNN algorithm.

Results and discussion

The results of this test are summarized in Figure 4.5. Each box plot represents the first, second and third quartile (Q_1, Q_2 and Q_3, respectively; from the bottom, the blue, the red and the next blue horizontal lines) as well as the lower and upper adjacent values (i.e., $Q_1 - 1.5 \cdot (Q_3 - Q_1)$ and $Q_3 + 1.5 \cdot (Q_3 - Q_1)$, respectively) for each method of selection of KPIs throughout the 100 iterations. Outliers are shown as crosses. The number of indicators that have been considered are: 286 for the first case (all of them), and 20 when the unsupervised selection methods are used (LS, MC and UP). Regarding TE, two troubleshooting experts (TE_1 and TE_2) were asked to select the most reliable KPIs to their knowledge. Table 4.1 shows the KPIs that these experts selected. TE_1 chose 20 out of the 286 available and TE_2, 5. Results (Figure 4.5) show that the proposed method could relieve the expert from analyzing and selecting the most reliable KPIs, even when no prior knowledge on the amount and variety of underlying fault causes in the network is provided. It can be seen how the selection of TE_2 left out some useful information when compared to *All*, leading to a higher DER. The reason for UP to outperform the results of TEs is that, given an unknown underlying fault cause, it is able to select not only KPIs directly related to that fault, but to select the indicators that (within that database) best allow to identify such fault, avoiding human bias. For example, instead of using KPIs like the uplink and downlink traffic (KPIs 4 and 8 in Table 4.1) to analyze the amount of traffic, KPIs related to the PRB utilization were selected.

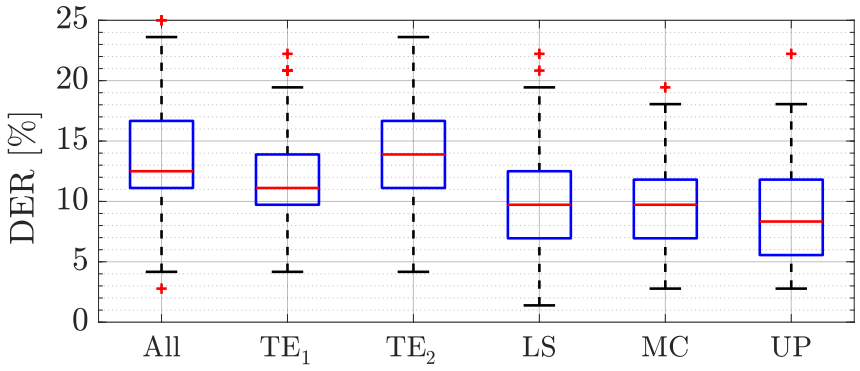


Figure 4.5: DERs for an LDA classifier given different methods for the selection of the KPIs. TE = Troubleshooting expert; LS = Laplacian score [87]; MC = Multi-cluster feature selection technique [92], and UP = Unsupervised selection of KPIs (the proposed method).

Regarding the case with all the KPIs, a lower median DER for UP than for *All* demonstrates the ability of the proposed method to reduce the volume of the OSS database by a 93% (as only 20 out of the 286 available KPIs have been considered) while improving the accuracy of the diagnosis method. To show this into more detail, two normalized confusion matrices are presented, comparing the case *All* (Table 4.2a) versus the case UP (Table 4.2b). In these matrices, C_i stands for the ground truth label, and C_i', for the diagnosed fault cause. As it can be seen in

Table 4.1: KPIs selected by the troubleshooting experts.

1. Accessibility	2. Retainability
3. #Bad cov. eval. rep.	4. Uplink traffic
5. CPU load 60%-80%	6. HOSR
7. Dropped call rate	8. Downlink traffic
9. #E-RAB attempts	10. Average CQI
11. IRAT handover rate	12. Avg. UE session time
13. #E-RAB succ. connections	14. #Random access attempts
15. #Succ. random access conn.	16. #RRC conn. attempts
17. #Best cell eval. report	18. Uplink data volume
19. Downlink data volume	20. #RRC conn. succ. estab.
$TE_1(20)$: 1-20	$TE_2(5)$: 1, 3, 5, 7, 8

Table 4.2a, only the *no traffic* case (C_2) was clearly identified. In the other cases, non relevant information, contained in some of the KPIs, led to a severe misclassification. On the contrary, if only the KPIs selected by UP are used in the diagnosis, C_1 is correctly diagnosed 85.7% of the time, and C_3, a 51.9%. C_3 misclassification might be due to the insufficient variety of KPIs related to the number of active connections within a given time window. Indicators like the average number of active E-RABs (E-UTRAN radio access bearer) would be decisive. In contrast to indicators like the number of E-RABs successfully established, which would be low for both C_2 and C_3 (most calls would be blocked). It can also be seen how C_4 (*low coverage*) could not be properly diagnosed regardless the selection method and despite having a counter for bad coverage evaluation reports (KPI 3 in Table 4.1). This may be due to the fact that the test database did not contain direct information about measurements of RSRP or RSRQ, which would be crucial. Also, an additional source of inaccuracy may be the ground truth labeling task. This is related to the ability of network experts to first differentiate the underlying fault causes given a set of unlabeled observations. If, due to their similarity, two abnormal situations were confused, or even if an abnormal situation remained hidden (not leading to its own class label) in the manual labeling process, the DER could be noticeably degraded. This could be the case of C_2 and C_3 and a possible hidden fault cause provoking similar symptoms to that of C_2 or C_3, like a *software failure*. Thus, one of the benefits of the proposed method is that, when used together with a subsequent unsupervised diagnosis technique [24], it does not require network experts to label future observations, and thus, to possibly introduce this source of error.

Table 4.2: Normalized confusion matrices (shown as percentages) after diagnosis for the selection methods: (a) all, (b) UP.

	C'_1	C'_2	C'_3	C'_4
C_1	14.3	14.3	42.9	28.6
C_2	0	98.3	1.7	0
C_3	47.2	52.8	0	0
C_4	100	0	0	0

(a)

	C'_1	C'_2	C'_3	C'_4
C_1	85.7	0	0	14.3
C_2	0	100	0	0
C_3	0	48.1	51.9	0
C_4	0	100	0	0

(b)

Finally, the proposed method is shown to outperform the KPI selection made by other state-of-the-art unsupervised techniques [87, 92].

4.5 Feature selection: supervised approach

In this section, a supervised technique for feature selection is proposed. As described in Section 4.3.1, supervised techniques for feature selection usually provide better results that unsupervised ones at the expense of needing labeled samples and, in the worst case, possibly neglecting network states not initially considered in these labels. This is shown in Section 4.5.2, where the supervised method for KPI selection proposed in this section is compared to that of Section 4.4.1.

4.5.1 Proposed method

The proposed method relies on how differently KPIs behave in the presence of different network states. This concept can be quantized for KPI p and network states i and j as the overlapping area of the PDFs of p conditioned to i and j: $f_p(x|i)$ and $f_p(x|j)$, respectively. This overlapping area can be mathematically expressed as Eq. (4.3) (from [93]), where $OVL(i, j, p)$ stands for the overlapping area for KPI p when the network states i and j are considered. As an example, Figure 4.6 shows $OVL(i, j, p)$ (in light gray) and $OVL(i, k, p)$ (in dark gray), which is almost negligible. In light of this, KPI p would be useful to discern between network states i and k, given its different behavior under both network states, but a bad choice in order to differentiate between i and j, given its similarity.

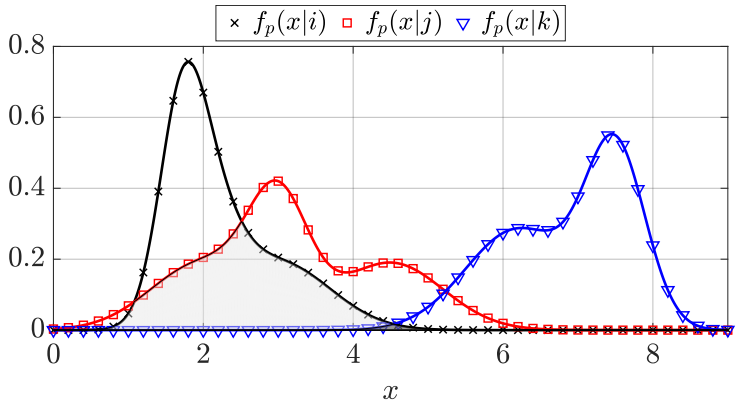


Figure 4.6: PDFs of KPI p conditioned to network states i, j and k, together with the overlap regions of $f_p(x|i)$ with both $f_p(x|j)$ and $f_p(x|k)$.

$$OVL(i, j, p) = \int \min(f_p(x|i), f_p(x|j))dx \tag{4.3}$$

Algorithm 4.1 describes the proposed method for KPI selection to discern among a set of network states. This procedure computes the overlapping area for each KPI for each pair (i, j) of network states with $i \neq j$. After $OVL(i, j, p)$ is computed for every p, a filter is applied: only those indicators with an overlap area lower or equal to a user-defined threshold (T) are retained,

in order to discard all the indicators that behave in a similar way under the network states i and j. Then, the resulting indicators are sorted by $OVL(i,j,p)$ in an ascending way, taking the n first. This way, the original set of KPIs (of size N) may be reduced to a set of up to $n \cdot C(I,2)$ selected indicators, where I stands for the total number of network states and $C(\cdot)$ stands for a binomial coefficient. As, at the most, N KPIs should be evaluated for every pair of causes, the computational complexity of the proposed method is bounded by $\mathcal{O}(I^2N)$.

Algorithm 4.1: Automatic selection of KPIs

```
for i ∈ list of states do
    for j < i do
        for p ∈ list of PIs and ∉ list of selected KPIs do
            Compute OVL(i, j, p) (Eq. 4.3);
        Keep only those KPIs with OVL(i, j, p) ≤ T;
        Sort OVL(i, j, p) in ascending order;
        n first KPIs → list of selected KPIs;
```

Now, for Eq. (4.3) to be computed, the PDFs for each KPI and network state must be estimated first from a set of M_f samples. The PDF estimate of KPI p under the presence of the network state i is $\hat{f}_p(x|i)$, contrary to the true PDF of such indicator, being $f_p(x|i)$. In order to make the proposed method as autonomous as possible, a non-parametric technique for PDF-estimation is used: the kernel density estimation (KDE). In KDE, the PDF estimate is computed as the weighted sum of kernel smoothers:

$$\hat{f}(x) = \frac{1}{M_f \cdot h} \sum_{m=1}^{M_f} K\left(\frac{x - X_m}{h}\right) \tag{4.4}$$

Kernel smoothers are non-negative real-valued integrable functions and are usually expressed as $K\left(\frac{x-X_m}{h}\right)$, where x stands for the random variable whose PDF is to be estimated; X_m is the actual value of the m^{th} sample of the random variable x, and h is a smoothing factor, often referred to as the bandwidth. That is, this PDF-estimating method relies on assigning each sample of a random variable a probability density in form of a baseline function rescaled by some factor and centered at the value of the sample. Some of the most widely used kernel smoothers are shown in Table 4.3, where u replaces $\frac{x-X_m}{h}$ in order to show them in a more compact way. The domain of these functions is also shown in this table. As it can be seen, most of these domains are bounded, but dependent on the value of u, which is in turn dependent on h. The higher the value of h, the smoother the PDF estimate. Thus, high values for h result in high bias and low variance for the estimate $\hat{f}(x)$ and vice versa. Besides, if h was sufficiently small compared to the spacing among the samples, then the overlap would tend to zero. Its impact in both the accuracy of the estimation of $\hat{f}(x)$ and the trade-off between bias and variance has led to a wide variety of bandwidth selection techniques. All of them rely on finding the value of h which minimizes the asymptotic mean integrated squared error (AMISE) of the estimate [94]. However, computing the AMISE for a given value of h implies knowing in advance the true PDF for x, $f(x)$. In order to deal with this, in this work the value for h for kernel smoother s, h_s,

is computed and optimized (leading to h_s^*) through a leave-one-out cross-validation (LOOCV) procedure, prior to the PDF fitting. h_s^* is chosen from a range of values for h_s, following the maximum likelihood (ML) criterion (Eq. (4.5)). The likelihood for a given value of h_s, $L(h_s)$, is computed with Eq. (4.6), using the LOOCV [95]. In this equation, u stands for the index of the left-out sample; v, for the index of the remaining samples, and M_{cv}, for the number of samples devoted to the cross-validation procedure. Note that M_f and M_{cv} sum up to the total number of samples used by the selection technique.

Table 4.3: Most widely used kernel smoothers.

Function	$\mathbf{K(u)}$	**Domain**
Normal	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$	$(-\infty, \infty)$
Uniform	$\frac{1}{2}$	$\lvert u\rvert \leq 1$
Triangle	$1 - \lvert u\rvert$	$\lvert u\rvert \leq 1$
Epanechnikov	$\frac{3}{4}(1-u^2)$	$\lvert u\rvert \leq 1$

$$h_s^* = \arg\max_{h_s} \left[L(h_s)\right] \tag{4.5}$$

$$L(h_s) = \frac{1}{M_{cv}} \sum_{u=1}^{M_{cv}} \log \left[\frac{1}{(M_{cv}-1)h_s} \sum_{v \neq u} K_s \left(\frac{X_u - X_v}{h_s} \right) \right] \tag{4.6}$$

4.5.2 Performance analysis

Experiment setup

In this section, the same dataset as in Section 4.4.2 has been used. That is, four different labels are differentiated: high traffic (C_1), no traffic (C_2), high CPU utilization (C_3) and low coverage (C_4). As in Section 4.4.2, the performance of different feature selection techniques is compared by evaluating the DER resulting from using a diagnosis technique which takes as its input only the KPIs chosen by such selection techniques. Again, an LDA classifier is used as the diagnosis tool.

Seven different situations for feature selection are distinguished. First, to set a baseline, all the KPIs are used, representing the situation when no selection is performed. Second, two troubleshooting experts (abbreviated hereafter as TE_1 and TE_2) are asked to select those KPIs that, to their knowledge, best represent the set of network states contained in this dataset. Next, the technique for feature selection proposed in Section 4.4.1 (abbreviated as UP onwards) is used, as an example of unsupervised technique for feature selection. Then, two well known supervised techniques for feature selection are used: the ReliefF algorithm [96] (abbreviated as RL onwards) and a sequential feature selection technique [97] (abbreviated hereafter as SQ). Finally, the proposed technique (abbreviated hereafter as OV, for overlap index-based technique for feature selection) is assessed.

Same as in Section 4.4.2, the dataset is split into three subsets: the *selection set*, the *training set* and the *test set*, following a 0.4, 0.4, 0.2 proportion, over which a 50-iteration MCCV is performed.

For OV, the Epanechnikov kernel smoother has been used, given its ability to optimally reduce the AMISE [98]. Besides, before comparing its performance with those of the other selection methods, a preliminary test has been performed to adjust T and n for a minimum DER. Both in this parameter tuning test and in the final test, the *selection subset* has been further subdivided into two equally sized subsets (having $M_{cv} = M_f$): one for the computation of h^*_{Ep} using Eqs. 4.5 and 4.6, and one for the PDFs fitting and KPI selection, following Algorithm 4.1. T has been ranged from 0.01 to 0.3, and n, from 2 to 4. For each sampling point, a 30-iteration MCCV has been performed on the original dataset, following the same 0.4, 0.4, 0.2 split. The resulting DER versus T and n is shown in Figure 4.7 with an error bar plot. On the one hand, low values for T lead to an increment in the DER, as it is quite likely that only one KPI is chosen to discern between a given pair of network states. On the other hand, high values for T may degrade the DER (specially for $n = 4$), given the increasing probability of some low-relevance indicator to be included for a certain pair of network states. For the final test, n has been set to 2 and $T = 0.03$, due to the stability of the DER with T and the high data volume reduction achieved in that case: at most, up to 12 KPIs may be selected, out of a total of 286. To make a fair comparison in the final test, the number of KPIs being selected by UP, RL and SQ has also been limited to 12.

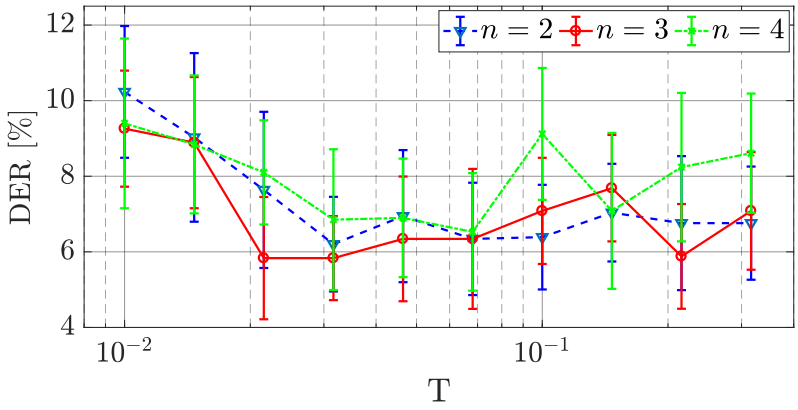


Figure 4.7: DER versus T for different values of n using the Epanechnikov kernel.

Results and discussion

Figure 4.8 shows the resulting DERs of this test over the 50 repetitions by means of a box plot for each case of KPI selection. In light of these results, TE_1 managed to get a similar DER to that of *All* by only using 12 KPIs (KPIs 1 to 12 from Table 4.4), meaning that a manual troubleshooting can be made without noticeably impacting the diagnosis performance. However, TE_2 only selected 6 KPIs (KPIs 1, 3, 8, 12, 13 and 14 from Table 4.4). In this case, the use of so few KPIs leads to a degradation of the DER when compared to both TE_1 and *All*. It can be seen how the lack of relevant information has a bigger impact on the DER than the degradation

due to the noise-like contribution of many of the indicators in *All*. Next, it is shown how UP yields a similar performance to that of supervised techniques like RL and SQ, all of which are better than both TE_1 and TE_2, due to the avoidance of human bias.

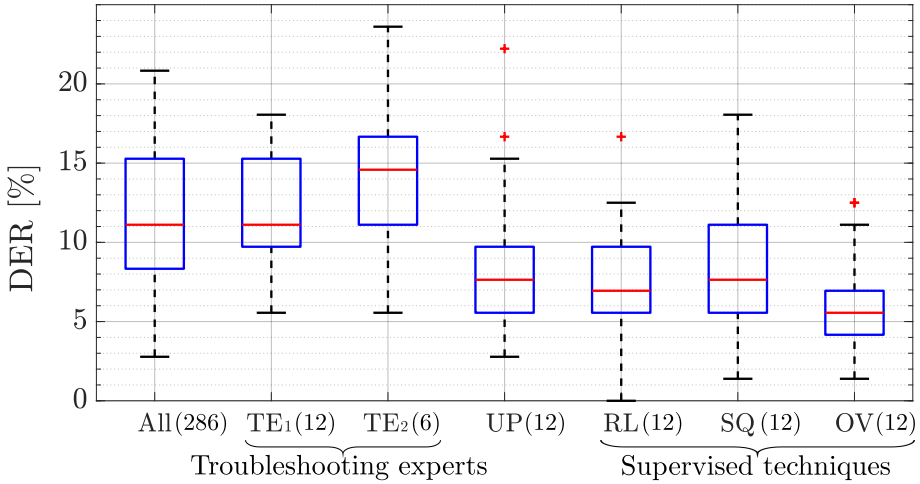


Figure 4.8: DERs for an LDA classifier given different methods for the selection of the KPIs. The number of KPIs being used in each selection case is shown between parentheses.

Table 4.4: KPIs selected by the troubleshooting experts.

1. Dropped call rate	2. #Random access attempts
3. Average channel qual. indic.	4. Retainability
5. #E-RAB succ. connections	6. HOSR
7. Uplink data volume	8. Downlink traffic
9. Uplink traffic	10. Accessibility
11. #Bad cov. eval. rep.	12. CPU load 60%-80%
13. Avg. UE session time	14. IRAT handover rate

Finally, it can be seen how OV clearly outperforms the remaining cases in terms of the median DER. In this way, when compared with the case *All*, the proposed method is shown to be capable of reducing the dataset size at least a 96%, while lowering the DER a 50%. In particular, the misclassification rates are shown on a network state basis for the cases *All* (Table 4.5a) and OV (Table 4.5b) with two normalized confusion matrices. According to these matrices, only the *no traffic* (C_2) and *high CPU utilization* (C_3) cases have been almost perfectly identified in both situations. Regarding Table 4.5a, the case *All* shows that C_1 and C_4 (*high traffic* and *low coverage*, respectively) are often confused. This may be due to the assessment of KPIs whose statistical behavior does not differ enough from one network state to another. For example, C_1 may be confused with C_4 when counters like the *number of RRC connection attempts* are assessed, being relatively high in both cases. Table 4.5b shows how this confusion has been fully removed for (C_4, C_1') and noticeably reduced for (C_1, C_4'), highlighting the benefits of the proposed method. Besides, Table 4.5a shows another source of confusion, the element (C_4, C_2'): a prediction of *no traffic* given a problem of *low coverage*, due to the assessment of KPIs like the *number of successful RRC connections*. This KPI presents low values for both C_2 and C_4. In

this case, Table 4.5b shows how this confusion still persists after the KPI selection. The reason for the proposed method to have chosen such KPI is that, despite it behaves in a similar way according to C_4 and C_2, it behaves differently according to the pairs (C_1, C_2) and (C_1, C_4), being a good option to discern between these network states. This issue could be addressed using more stringent criteria in Algorithm 4.1. For example, by bounding the admissible overlap for a given KPI for every pair of network states. However, this would noticeably reduce the number of KPIs (similar to the effect of a small T in Figure 4.7), which would lead to an eventual increase in the DER.

Table 4.5: Normalized confusion matrices (shown as percentages) after diagnosis for the selection methods: (a) all, (b) OV.

	C'_1	C'_2	C'_3	C'_4
C_1	41.7	8.3	0	50
C_2	0	94.5	0	5.5
C_3	0	0	100	0
C_4	33.3	33.3	0	33.3

(a)

	C'_1	C'_2	C'_3	C'_4
C_1	83.3	0	8.3	8.3
C_2	0	100	0	0
C_3	0	0	100	0
C_4	0	33.3	0	66.7

(b)

As a final remark, and in order to cope with the time-varying nature of the network behavior, periodical reselections of KPIs should be performed, including recently labeled samples in the sample set devoted to the KPI selection.

4.6 Feature extraction

To complement Sections 4.4 and 4.5, in which feature selection techniques are assessed in the context of self-healing, in this section, the benefits of applying feature extraction techniques in a task of automatic diagnosis are presented. To that end, two well-known families of methods for feature extraction are first presented and then compared in the task of reducing the dimensionality in the scenario and dataset presented in Section 4.4.

4.6.1 Overview of feature extraction techniques

A feature extraction technique is a process that transforms an element $\bar{x} \in \mathbb{R}^N$, defined by the features $\{x_1, \ldots, x_N\}$, in an element $\tilde{x} \in \mathbb{R}^E$, defined by the features $\{\tilde{x}_1, \ldots, \tilde{x}_E\}$, where each new feature $\tilde{x}_j$ comes from the combination of several features of $\bar{x}$. Even though $\tilde{x}$ may have any number of dimensions, in general, it is preferred that $E < N$, thus making $\tilde{x}$ contain the information of $\bar{x}$ in a lower number of dimensions.

The reason why feature extraction techniques get to retain a higher amount of useful information that feature selection techniques given a same number of resulting features is that not all the information retained in the *selected* (contrary to the *extracted*) KPIs is *useful* information. This fact is mainly due to the nature of the magnitude being quantified and its suitability to represent the performance of the underlying process. At this point, the utility of a feature,

or KPI, can be quantified as its variance with respect to the variance of the whole data set, considering all the KPIs. Often, the techniques for feature extraction have as a criterion that the per-feature variance is maximized, thus reducing the number of necessary features to retain the whole variance. This is the case of the techniques based on principal component analysis or PCA [99].

Next, some particular techniques for feature extraction are briefly described, as part of two of the most studied families of methods for dimensionality reduction: techniques based on component analysis and techniques based on manifold learning.

Component analysis

The techniques based on component analysis define the features transformation using information from their statistic behavior. Within this group, one may find techniques implying both a linear transformation over the original features (e.g., PCA) and techniques implying non-linear transformations, like the kernel PCA (kPCA).

- **PCA**: Principal component analysis is one of the most used feature extraction techniques in a wide variety of fields of science, due to its high effectiveness and ease of implementation. Given an original space of N dimensions, PCA determines the E dimensions or hyperplanes that, being orthogonal among them and a linear combination of the former N features, maximize the variance of the projection of the original samples.
- **Kernel PCA (kPCA)**: this non-linear feature extraction technique applies PCA over the features resulting from a non-linear transformation of the original features [100]. The non-linear function applied on the original space is known as the *kernel*. This way, simple hyperplanes, resulting from the application of PCA, defined over the transformed space, result in complex structures in the space of the original features.
- **Independent component analysis (ICA)**: Unlike PCA, which only seeks for the orthogonality of the resulting features, ICA targets the statistic independence of these; generally by means of minimizing the mutual information among the features [101]. There exist linear and non-linear variants of this technique.

Figure 4.9 shows, as an example, the resulting features of applying PCA and ICA over a set of test samples. On the one hand, the orthogonality of $\tilde{x}_1$ and $\tilde{x}_2$ can be seen as a result of applying PCA; on the other hand, the statistic independence of the distributions of $\tilde{x}'_1$ and $\tilde{x}'_2$ appears with the possibility of expressing its joint PDF as the product of their marginal PDFs in the transformed space (i.e., the space defined by $\tilde{x}'_1$ and $\tilde{x}'_2$).

Manifold learning

The methods based on manifold learning are a set of non-linear techniques for feature extraction. They rely on the premise that a set of samples of high dimensionality is indeed a body with a set of a lower number of dimensions, whose shape has been manipulated to result in the latter.

Figure 4.10 shows an example of this. In Figure 4.10a, a 2D plane which has been rolled up to result in a 3-dimensional structure is depicted. Figure 4.10b shows the result of unwrapping

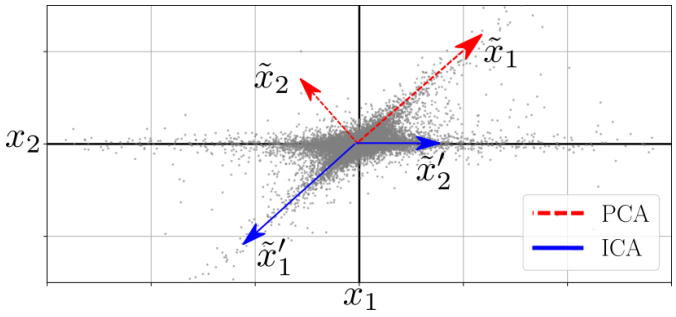
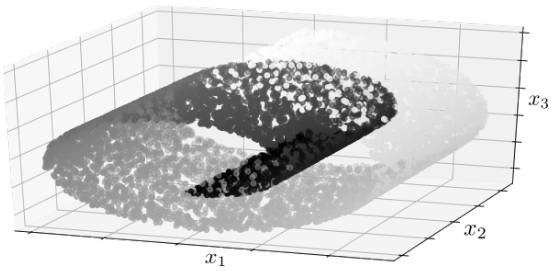
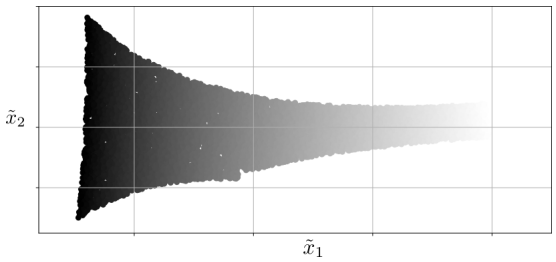


Figure 4.9: Resulting features of the application of PCA (dashed red lines, $\tilde{x}$) and ICA (solid blue line, $\tilde{x}'$) over a two-dimensional set of samples [102].

the plane after applying a manifold learning technique. In a problem of classification, it would be difficult to define the decision boundaries over a structure as the one shown in Figure 4.10a; however, it would be easy to do this with the structure shown in Figure 4.10b.



(a) Samples $\bar{x}$ in the N-dimensional original space.



(b) Samples $\tilde{x}$ in the E-dimensional transformed space.

Figure 4.10: Example of feature extraction using LLE, a technique based on manifold learning, from [103].

Some of the most well-known feature extraction techniques based on manifold learning are the following:

- **Locally-linear embedding, LLE**: this technique allows determining the E-dimensional space which better preserves the distance between the projection of each sample and their neighbors' [104]. Figure 4.10 shows an example of the application of this technique.
- **Spectral embedding, SE**: this technique, as LLE, uses the concept of neighborhood among samples. In this case, to define the graph whose spectral decomposition allows defining the E-dimensional space onto which projecting the original space [105].

4.6.2 Performance analysis

Some of the feature extraction techniques presented in the previous section are assessed next regarding their ability to reduce the number of features at the input of a method for RCA while preserving as much useful information as possible.

Experiment setup

In this section, the same dataset as in Section 4.4.2 has been used. That is, four different labels are differentiated: high traffic (C_1), no traffic (C_2), high CPU utilization (C_3) and low coverage (C_4). As in Section 4.4.2, the performance of different feature extraction techniques is compared by evaluating the DER resulting from using a diagnosis technique which takes as its input only the KPIs chosen by such selection techniques. Again, an LDA classifier is used as the diagnosis tool.

Seven different situations for feature extraction are distinguished. First, to set a baseline, all the KPIs are used, representing the situation when no selection is performed. Then, four techniques based on component analysis have been assessed: PCA, kPCA using a Gaussian kernel (kPCA_1 onwards), kPCA using a sigmoid kernel (kPCA_2 onwards) and ICA. Finally, two feature extraction techniques based on manifold learning are used: LLE and SE. The same number of synthetic KPIs are considered in all these situations, 10.

As in Section 4.5.2, the dataset has been partitioned following a 0.4, 0.4, 0.2 split, devised for the computation of the KPI transformation model for the feature extraction techniques, the training of the LDA classifier and the testing of this classifier, respectively. Again, a 50-repetition stratified MCCV has been performed. Finally, a standard normalization is applied to each split.

Results and discussion

Figure 4.11 shows the resulting DERs for this test. In light of this, all the component analysis-based techniques except for kPCA_1 provide a lower DER than the case with all the KPIs. The reason for the good results of PCA comes from the nature of the KPIs being monitored and the relation they have among each other, being fundamentally linear. In turn, this linearity appears from the nature of the processes being quantified. For example, for several network states, indicators like the number of call attempts and the call success rate are related by means of a linear expression. In the case of kPCA_2, the normalization of the KPIs and the distribution of many of them around zero makes the sigmoid kernel to approximate to a linear function, providing an optimum final DER, similar to the one provided by PCA. On the other hand, the resulting DERs of both kPCA_1 and the techniques based on manifold learning provide poorer results, even below the baseline case for LLE. The reason is that these techniques assume a strong non-linearity in the dataset, which is contrary to the linear character of the relations among the considered KPIs.

4.7 Dimensionality reduction-based self-healing framework

In this section, a framework including both feature selection and feature extraction techniques for dimensionality reduction in the field of self-healing functions is described. The framework has been designed in a way that it overcomes the individual limitations of the families of techniques composing it, which have been described in the previous sections. Besides, the proposed framework allows different sources of performance information to be integrated: from information gathered from the network, to UE-reported measurement reports and context information.

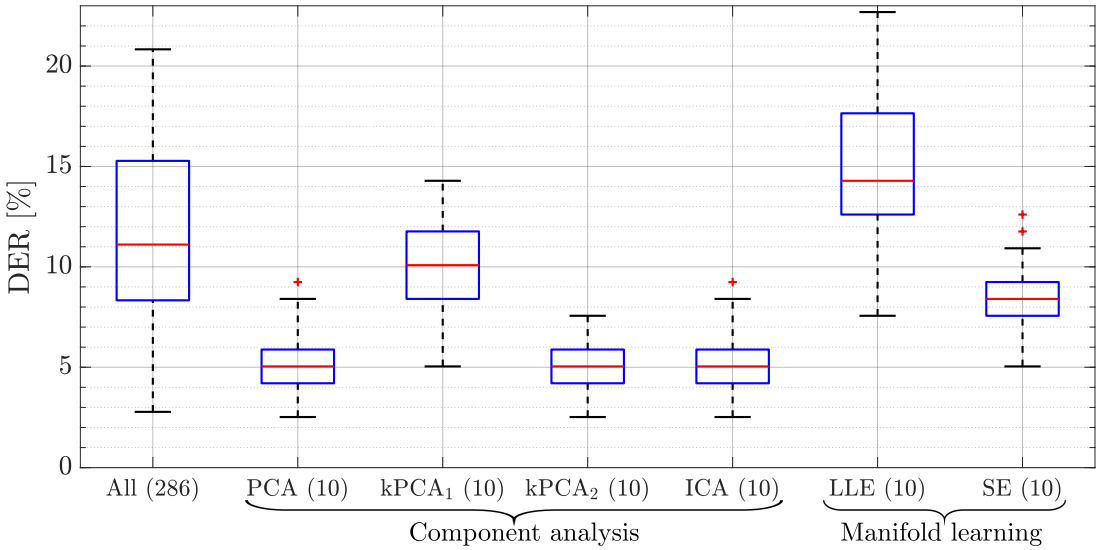


Figure 4.11: DER of diagnosis system based on LDA for different feature extraction families and techniques. The number of indicators used as the input of the self-healing function is shown between parentheses.

4.7.1 Proposed framework

As it was described earlier in Chapter 3, self-healing functions often consist of three phases: a design phase, usually involving a network expert (referred to as troubleshooting expert), followed by a training phase and an operating phase, in which self-healing functions operate autonomously. The design phase is in which troubleshooting experts apply their knowledge. First, they are in charge of selecting the set of network states (e.g., cell outage, weak coverage, etc.) to be eventually identified and repaired. Next, they select the KPIs (e.g., drop call rate, accessibility, etc.) that, to their knowledge, best allow those states to be eventually diagnosed. And finally, they are in charge of providing a knowledge base for self-healing functions, usually in the shape of a set of samples, labeled by them after having been analyzed. In the training stage, self-healing functions learn how network states and KPIs relate to each other. To that end, a system model is built by analyzing the collection of labeled network samples provided by the troubleshooting expert (Figure 3.1). Finally, during the operating phase, self-healing functions predict a network state and even decide which actions to take given a set of new unlabeled samples and the system model from the training phase.

Figure 4.12 shows the proposed framework for self-healing functions, in which the upper block covers the design and training phases, and the lower block covers the operating one. These blocks are further divided into two sections by a dotted vertical line. The left side (which is common in both the upper and lower block) corresponds to a data acquisition and formatting stage. It is in charge of gathering and integrating data coming from different sources. The right side is the one performing the dimensionality reduction and data forwarding to the self-healing function.

The main target of the data acquisition and formatting stage is to provide self-healing functions with information as much varied and detailed as possible: from the RAN- and CN-

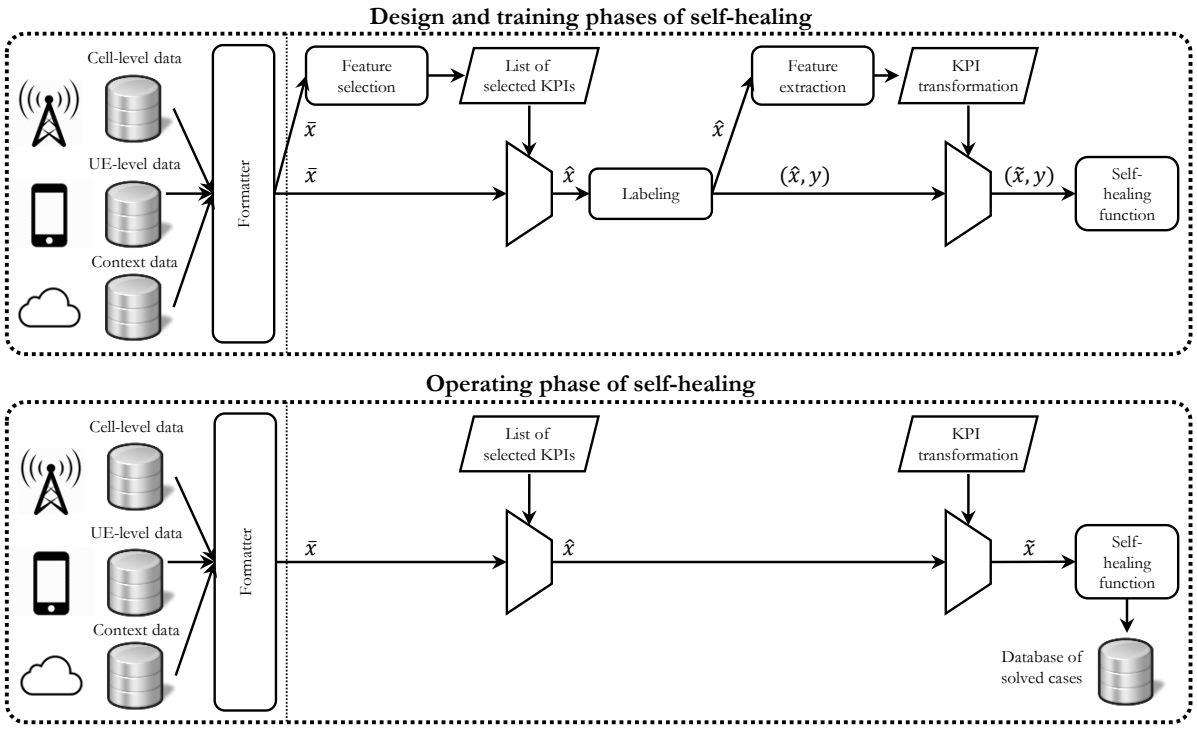


Figure 4.12: Proposed framework for next-generation self-healing networks, including different data sources and dimensionality reduction techniques.

level information that is used in current SON deployments, to UE- and context-level information. As of today, cell-level KPIs may be directly found in the OSS databases, in the form of event counters or more elaborated metrics, computed from the first. The information from the UE may be retrieved in two different ways: in a network-managed way, following the minimization of drive tests (MDT) functionality [106], and by means of specific software running on the UE, acting as an end-user probe. In the first case, this information is stored in base stations as UE logs, consisting of signaling trace records. For example, a register of call trace events collected along the Uu, S1 and X2 interfaces and radio link power and quality measurements for LTE networks. In the second case, these probes allow operators to retrieve service-specific end-to-end performance metrics, providing information about the QoS experienced by end-users. Regarding context information, the most used type is user location and speed, which can be retrieved from MDT logs. Other sources of context information are social networks and over-the-top (OTT) applications. In these cases, network operators and service providers must reach an agreement, so that the former can retrieve the information collected by the latter. Finally, in line with the end-user perceived QoE, information from the users' complaints may be retrieved using a customer experience management (CEM) tool.

Having the information from different sources collected, a formatting process is needed for the resulting indicators to be expressed in an homogeneous format. On the one hand, this procedure includes making these indicators share the same temporal resolution. And, on the other hand, it includes transforming them into quantitative variables in case they were not. For example, in the case of the customers' complaints. As a result, a sample made up of N indicators,

$\bar{x}$, is generated (see Figure 4.12, both in the upper and lower blocks).

Regarding the dimensionality reduction section, a different approach is followed in each block in Figure 4.12. Starting with the upper block (design and training phases), a feature selection technique is proposed to be used over $\bar{x}$. This prevents the troubleshooting expert from having to analyze and select the KPIs for the self-healing function. At this point, and given that no label is attached to the samples, an unsupervised technique for feature selection should be used (e.g., the one described in 4.4.1). This will result in a list of selected KPIs, which will be used to filter $\bar{x}$, resulting in $\hat{x}$. The next step is to provide a network state label y (resulting in the pair $(\hat{x}, y)$) in case that the self-healing function requires labeled samples. In such case, it is necessary to use troubleshooting experts' knowledge in early network deployments for this task. However, assuming a sufficiently mature network, meaning the availability of a populated enough database of past solved cases, the troubleshooting expert could be relieved of analyzing new cases for new re-trainings. Having such database, these labels could be provided by a CBR tool. This tool would verify how similar the new cases under evaluation are with respect to those that have already been evaluated in the past by the self-healing function. That would also be useful for the feature selection technique, which could be supervised instead.

Next, in order to optimize the way how information is provided to the self-healing functions, the usage of a feature extraction technique is proposed. It intends to further reduce the number of KPIs at the input of self-healing functions, in an attempt of compacting the indicators already selected into a reduced number of new synthetic indicators. This allows self-healing functions to be more time efficient and less prone to over-fit, one of the most common issues in ML. As a result, a model for KPI transformation is provided. This transformation is applied over the pair $(\hat{x}, y)$ (only affecting $\hat{x}$), resulting in the pair $(\tilde{x}, y)$, which is used for the training of the self-healing function.

Once that the self-healing function has been trained (i.e., the model relating $\tilde{x}$ and y has been built), it can be used in the operating phase. In this phase, the new unlabeled samples to be evaluated, $\bar{x}$, must be first reduced to the form $\tilde{x}$ (Figure 4.12 below). Thus, a filter using the list of selected KPIs derived in the prior phases is used, without having to compute the list again. In order to reduce the storage needs of the network databases, this filter could be implemented by only monitoring and storing the KPIs in the list. In the same way, the KPI transformation derived in prior phases is applied next to get $\tilde{x}$, without having to compute it again, feeding the self-healing function. Finally, the resulting output of this function is stored in a database of past cases, so that it helps automating future re-designs and re-trainings. At this point, it is the operator's decision whether to store $\bar{x}$, $\hat{x}$ or $\tilde{x}$ together with the resulting output, being a trade-off between the capability to improve future feature selection and labeling and their storage requirements, respectively.

4.7.2 Performance analysis

A proof of concept has been carried out to assess the proposed framework. To this purpose, a diagnosis tool consisting in an LDA classifier has been used. Due to the unavailability of a dataset from a live network simultaneously containing cell, UE and context data, this proof of

concept has been divided into two tests. In the first, a high-dimensional dataset only containing cell-level data from a live network is used to prove the benefits of the proposed framework in terms of dimensionality reduction [90]. In the second, a medium-dimensional simulation-based scenario containing cell, UE and context data is used to prove its benefits in terms of data integration [31].

Test 1: Dimensionality reduction

Experiment setup For this test, the same dataset as the one in Sections 4.4.2, 4.5.2 and 4.6.2 has been used. Thus, each sample is made up of 286 cell-level KPIs plus one ground truth label, where these may be *high traffic* (C_1), *no traffic* (C_2), *high CPU processing load* (C_3) and *low coverage* (C_4).

In a similar way to prior tests in this chapter, in this test, 40% of the samples are used by the dimensionality reduction techniques to compute the list of selected KPIs and their subsequent transformation (see Figure 4.12, upper block). Another 40% are used in the training phase of the self-healing function, and the remaining 20% are used to test the performance of the classifier in terms of its DER. This test is repeated 100 times in a stratified MCCV, given the small sample size.

Nine different situations are assessed in order to test different schemes for dimensionality reduction in self-healing. The first situation, shown as a baseline, results from taking all the available indicators from the database and using them as the input for the diagnosis algorithm. In the second case, a troubleshooting expert (abbreviated as TE onwards) is asked to select the subset of indicators that, to his knowledge, better represent the variety of underlying fault causes. In this case, 20 out of the 286 available indicators were selected. Next, two different unsupervised techniques for feature selection have been used: the Laplacian score [87] (already presented in Section 4.4.2 and abbreviated as U1) and the unsupervised technique for feature selection described in Section 4.4.1 (abbreviated as U2 onwards). This represents the case of an early deployment of a cellular network, when no knowledge from past cases is available. Following, there are three cases of dimensionality reduction through supervised techniques for feature selection (abbreviated as S1, S2 and S3 onwards): a sequential feature selection technique [97] (already presented in Section 4.5.2), the technique NCFS for feature selection [88], which was presented earlier in Section 4.4.1 as the basis for U2 and the supervised technique for feature selection proposed in Section 4.5.1. These cases correspond to a situation in which a certain amount of past solved cases is available. Next, a feature extraction technique is assessed, in order to compare its standalone performance with that of feature selection techniques. In this case, PCA has been used, due to the high performance it showed in Section 4.6.2. In U1, U2, S1, S2 and PCA, 20 indicators have been used, so the comparison among these and TE can be fair. In the case of S3, a maximum of 24 KPIs (setting n to 4 and T to 0.03) was fixed, finally leading to 17 selected KPIs. Finally, the proposed framework (FR) has been tested. To that end, S2 (NCFS) has been used, followed by a module for feature extraction. Again, PCA has been used, being the less computationally expensive technique (when compared to other techniques for feature extraction, like the non-linear ones) while providing the best results for RCA, as it is shown in Figure 4.11 in Section 4.6.2. In this test, the first four principal components have been

taken, retaining 90% (on average) of the total variance of the indicators selected by S2 (Figure 4.13).

An additional test is carried out to show the trade-off between the number of synthetic indicators provided by the feature extraction module and the resulting performance of the self-healing tool. To that end, a PCA module is used following S2 using an increasing number of considered principal components and assessing the subsequent DER.

Results and discussion In Figure 4.13, each box plot shows Q_1, Q_2 and Q_3 as well as the lower and upper adjacent values for each of the situations for dimensionality reduction throughout the 100 iterations. Outliers are shown as crosses. As it can be seen, the worst values for DER correspond to the usage of all the indicators, since many of them may contribute as noise sources in the diagnosis process. The selection of TE shows a slightly better performance compared to that of the *all indicators* case, at the expense of having spent a long time identifying which indicators could be more related to the underlying fault causes. From here on out, all the cases for dimensionality reduction show their capabilities to relieve the TE from making this selection, showing S3 and U2 as the best techniques for supervised and unsupervised feature selection, respectively. Following any of these approaches, the management costs can be reduced by a decrease of a 93% in the volume of the OSS databases (20 out of 286 indicators). Figure 4.13 also shows the performance comparison of feature selection techniques versus feature extraction techniques. As stated in Section 4.3.2, extraction techniques allow a higher performance of subsequent RCA functions when compared to that of selection techniques at the expense of providing a set of synthetic KPIs that are not comprehensible by troubleshooting experts. Finally, FR achieves the best results in terms of Q_1 and the lower adjacent value while only using four KPIs. That is, further achieving a reduction of 98.6% in the databases size.

Table 4.6 shows 10 out of the 20 cell-level KPIs that TE selected (upper side). The lower side of the table shows some of the indicators that S2 selected. As it can be seen, the KPIs that TE selected tend to be magnitudes more comprehensible or interpretable to a human, like the HOSR or the accessibility. However, these might not be the best to identify a given set of network faults. S2 avoids this human bias and, instead, selects a set of indicators that despite being less directly interpretable by a human expert, carries the kind of information needed for the purpose. For example, several specific histogram-like counters are selected, like the number of times that the downlink PRB utilization ranges from 10% to 20%, rather than less specific indicators, like the downlink data volume.

Figure 4.14 shows the results (averaged over the 100 iterations) when the synthetic KPIs generated at the output of the feature extraction technique are used at the input of the diagnosis tool. The left axis shows in blue bars in descending order the *explained variance* of the synthetic KPIs (i.e., the principal components). The *explained variance* is a ratio: the variance calculated on a given subset of features with respect to the variance calculated on the whole feature set. The cumulated *explained variance* is shown in a blue solid line on the same axis. The right axis shows the corresponding subsequent DER. Nine synthetic KPIs are enough to reduce the DER to a 0%.

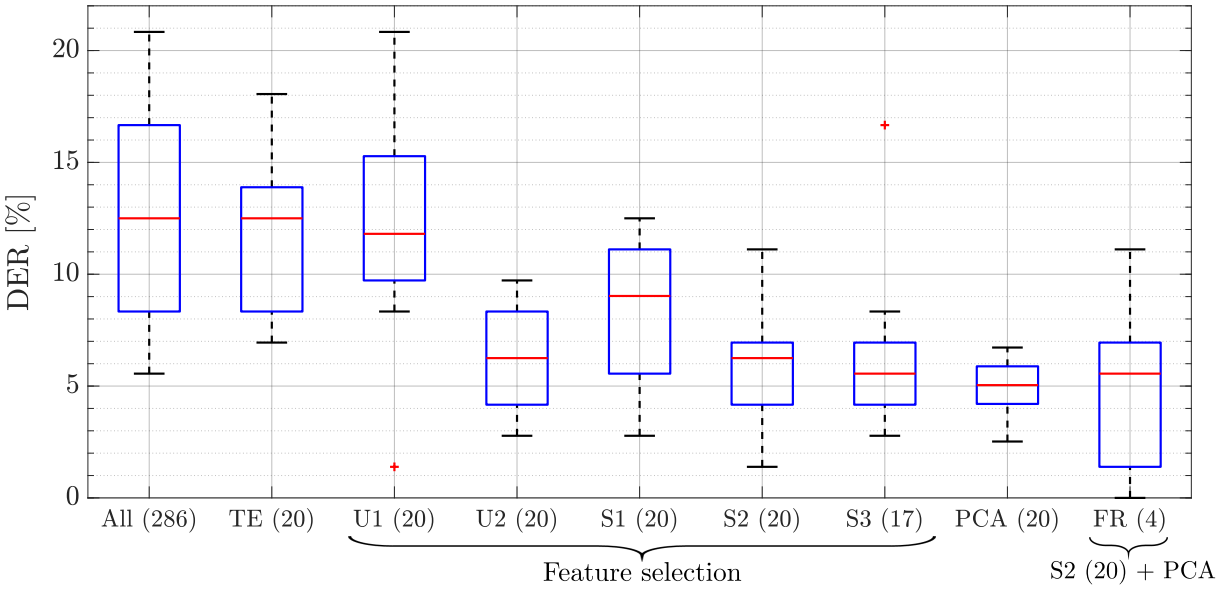


Figure 4.13: DERs for an LDA classifier given different methods for dimensionality reduction over KPIs. TE = Troubleshooting expert; U1 [87]; U2, unsupervised technique for feature selection, proposed in Section 4.4.1; S1 [97]; S2 [88]; S3, supervised technique for feature selection, proposed in Section 4.5.1 and FR (the proposed framework). The number of indicators used as the input of the self-healing function is shown between parentheses.

Table 4.6: Selection of KPIs.

Troubleshooting expert	
1. Dropped call rate	2. Average CQI
3. CPU load 60%-80%	4. Retainability
5. E-RAB estab. succ. rate	6. HOSR
7. Uplink data volume	8. Downlink data volume
9. #Bad cov. eval. rep.	10. Accessibility
Supervised feature selection technique S2, [88]	
1. #-2 < $\text{SINR}_{PUSCH} \leq$ 2 dB	2. #10% < PRB downlink utiliz. $\leq$ 20%
3. #Bad cov. eval. rep.	4. #-9 < $\text{SINR}_{PUCCH} \leq$ -6 dB
5. #HARQ failure uplink QPSK	6. Traffic volume PDCP DRB DL
7. UE average session time	8. #RRC conn. establ. succ.
9. #RRC conn. estab. att. MOS[1]	10. #(0 < UE uplink throughput (PDCP) $\leq$ 1 Mbps)

[1] Mobile-originated signaling

Test 2: Data integration

Experiment setup This simulation-based dataset contains 574 one-minute cell-level samples and 574×600 UE-level 0.1-second samples. The states *overshoot*, *interference*, *weak coverage* and *normal state* are independently forced on several cells, gathering their indicators (e.g., call drop rate or PRB utilization) and those from two of their neighbors. Only the UEs served by these cells have been monitored. As for them, measurements over the radio link (RSRP, RSRQ, SINR)

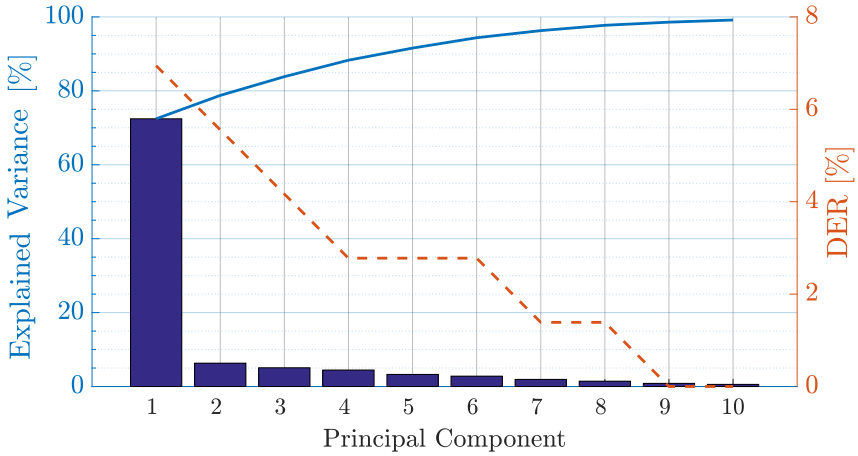


Figure 4.14: On the left axis: individual (bars) and cumulated (solid line) percentage of the *explained variance* by the first to the tenth principal component of PCA. On the right axis, dashed line: resulting DER from taking an increasing number of principal components as the KPIs at the input of the diagnosis self-healing function.

and throughput make up UE-measured data, including their position as context information. This leads to a 78-dimensional dataset when all these features are considered jointly.

Three different situations for data integration for diagnosis have been assessed: using only cell-measured data, using only UE-measured data, and using both together. Each situation has been further divided in two (leading to six situations), applying or not the proposed framework over the integrated data. In this test, the formatting stage (Figure 4.12) consists in averaging 600 samples of UE data each minute. For each situation, the same split, shuffling and number of repetitions as in Test 1 has been performed. In this test, FR consists in S2 selecting 20 indicators followed by PCA taking as many component carriers as to gather 95% of the *explained variance* (five of them, on average).

Results and discussion Figure 4.15 shows that the best case, FR(Cell+UE), corresponds to applying the proposed framework over the integrated data sources. In such case, the diagnosis tool benefits both from the richness of the data provided and from their low dimensionality, two conditions that are not present in any of the other situations.

4.8 Conclusion

In this chapter, different families of techniques for dimensionality reduction have been first described, together with the benefits and drawbacks that their usage implies for self-healing tasks. Next, two techniques for feature selection (one, unsupervised, and the other, supervised) have been proposed, showing their ability to relieve a troubleshooting expert of having to analyze and select the KPIs that best allow network states to be subsequently identified and their ability to reduce the storage needs of the network databases. Whereas the proposed unsupervised technique is specially suited for recently deployed cellular networks, in which the set of network states may still not be known, the proposed supervised technique is specially suited for more stable

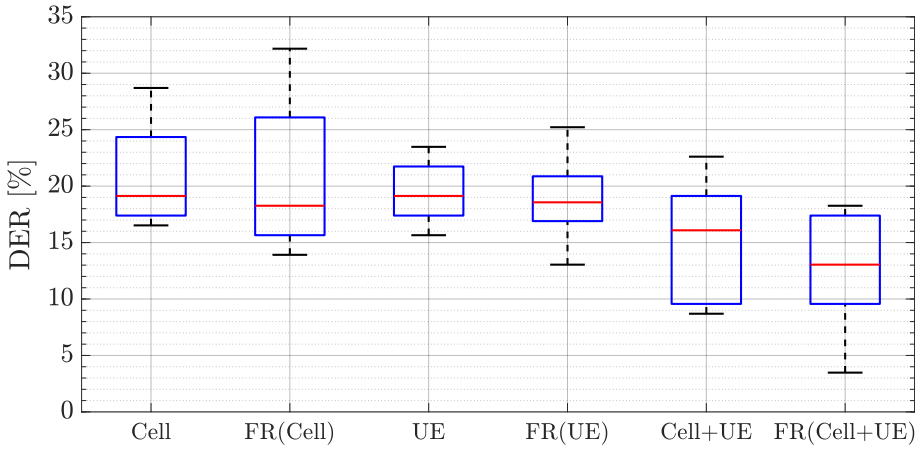


Figure 4.15: DERs of the diagnosis tool given different sources of data and integration methodologies. FR(·) stands for the application of the proposed framework over a source of data.

cellular networks, allowing the diagnosis performance to be further enhanced at the cost of previously knowing the set of underlying network states. Then, different types of feature extraction techniques have been assessed in the field of RCA. Finally, based on the insight acquired from the analysis of dimensionality reduction techniques in the rest of the chapter, a novel framework for self-healing tasks has been proposed, including both feature selection and feature extraction techniques, allowing a number of different sources of performance information to be integrated and eventually enhancing the performance of the subsequent self-healing functions.

Chapter 5

Self-Optimization for 5G NR

In this chapter, some advancements for self-optimization in the upcoming 5G NR are presented. Chapter 5 is divided in two main sections. In Section 5.1, a new RRM functionality, which relies on the forthcoming MC in cellular networks, is proposed to enhance network and UE performance for eMBB traffic. In Section 5.2, a method to reduce the E2E latency in a V2X environment is described and assessed in the context of low-latency communications. Finally, Section 5.3 outlines the conclusions of this chapter.

5.1 Optimizations for eMBB traffic through multi-connectivity

In this section, a tool to manage the assignment of CCs to UEs in a MC-capable cellular network is described and assessed as a means to enhance the network and UE performance.

This section is organized as follows. In Section 5.1.1, the challenges that MC will face during its implementation and the related work are presented. Section 5.1.2 outlines the problem formulation. Next, in Section 5.1.3, the proposed method for CC management is described. And finally, Section 5.1.4 elaborates on the proof of concept that has been carried out with a simulation tool.

5.1.1 Related work

For the benefits of MC to be fully exploited and its drawbacks to be minimized, a number of challenges should be addressed in the long term. Some of these are the downlink and uplink decoupling, that is, to differentiate the nodes that provide downlink and uplink scheduling grants for a given UE; the control- and user-plane split, meaning the mapping of control- and user-plane data onto different CCs, or the joint usage of both licensed and unlicensed frequency bands.

In this chapter, however, two more immediate challenges of MC are addressed. The first of them relates to the selection of the CCs to be assigned to a certain user. Given that not

all the CCs managed by a node are equally loaded (in terms of the number of users being allocated) or experiment the same channel conditions, different criteria for user allocation might be followed, depending on network operators' policies. Since CA was first proposed in Rel-10, different schemes for CC allocation arose, some of which are outlined next. In principle, two simple mechanisms for this assignment were proposed: the least load assignment, with the aim of an equal load balance among CCs, and the random carrier assignment [107]. These mechanisms were found useful, but unable to cope with current trends in cellular communications, like the pushing user-centric vision of the network management. To that end, QoE-based carrier scheduling schemes were proposed [108], maximizing the mean opinion score (MOS) for different services by periodically selecting the CCs which optimize the performance indicators over which the expressions for the MOS were described. Other examples of criteria for CC selection may be found to address mobility issues. In [109], a user mobility-based mechanism for CC selection is proposed regarding the UE's speed, as a trade-off between load balance and the number of handovers required to keep a good signal quality. Finally, and opposing to the extended trend of CC selection according to downlink metrics, in [110], a CC selection mechanism regarding uplink issues is presented, using the maximum power reduction as an input metric for the CC selection. In either case, these methods aim at selecting the CCs to be assigned to a UE from a single node. With the advent of DC, an additional issue arose: the selection of the node acting as the SN, from which additional CCs may be accessed. This brought attention to the following challenge.

The second challenge to be addressed by MC is the UE-BS association rule, eventually derived from network operators' policies as well. As of today, in commercial cellular networks, a UE relies on metrics related to the received power, like the RSRP, to select/reselect the node in which to camp. However, with DC in LTE, novel UE-BS association rules have arisen, to be used both for the selection of the MN and the SN, after which a CC selection might be performed. For example, in [111], a method is proposed for the UE to camp in the cell with the best expected QoS, relying on a ML technique for traffic prediction. In [112], the association rules do not only rely on channel quality metrics, but also on users' mobility patterns.

With the upcoming MC, independently addressing the challenges of selecting which BSs will act as SNs, and which of the CCs belonging to the former will be assigned to a given UE may lead to a suboptimal selection of CCs and thus, to a suboptimal UE overall performance.

5.1.2 Problem formulation

Throughout this section, it is assumed that the PCell, and thus, the MN, is selected according to 3GPP-standardized mechanisms. That is, in RRC idle mode, the PCell is selected by the UE according to the selection or reselection criteria defined by network operators. In RRC connected mode, the PCell selection is performed in a UE-assisted network-controlled manner, according to HO events defined by 3GPP. In either case, the PCell is assumed to be the strongest (highest RSRP) cell.

This way, Section 5.1 addresses the issue of determining which SCells and PSCells (thus, which SNs) should be allocated to an MC-capable user to enhance his performance.

Each service category quantifies its performance in a different way. In the case of eMBB, throughput is the main performance metric being assessed; in the case of URLLC, it is reliability and latency. MC aims at optimizing these metrics by treating in a different way the data flow supported by the nodes simultaneously accessed by a UE. For an eMBB service, data coming from the core network are split among the serving nodes, increasing the overall amount of radio resources allocated to such user. For URLLC, these data are replicated among the serving nodes.

The different treatment of the data flows according to the service category used by the UE in MC is performed at the NR PDCP and NR MAC layers of the involved nodes. This can be seen in Figure 5.1. Specifically, intra-node data split is performed at the MAC layer (UE_a) by means of CA. The remaining cases: inter-node data split (UE_b) and inter- and intra-node data duplication (UE_b and UE_c, respectively), take place at the NR PDCP layer [4]. In the latter case, the two duplicated logical channels are mapped onto two different CCs (thus, using CA), so that the benefits of the link diversity can be preserved.

Thus, in order to exploit the benefits of MC, a tool to coordinate the nodes involved in a MC scenario, either for eMBB or URLLC traffic, needs to be developed. This tool would impact both the NR PDCP and NR MAC layers, determining which CCs from which node should be used by a UE at a certain time, according to a given operator's policy.

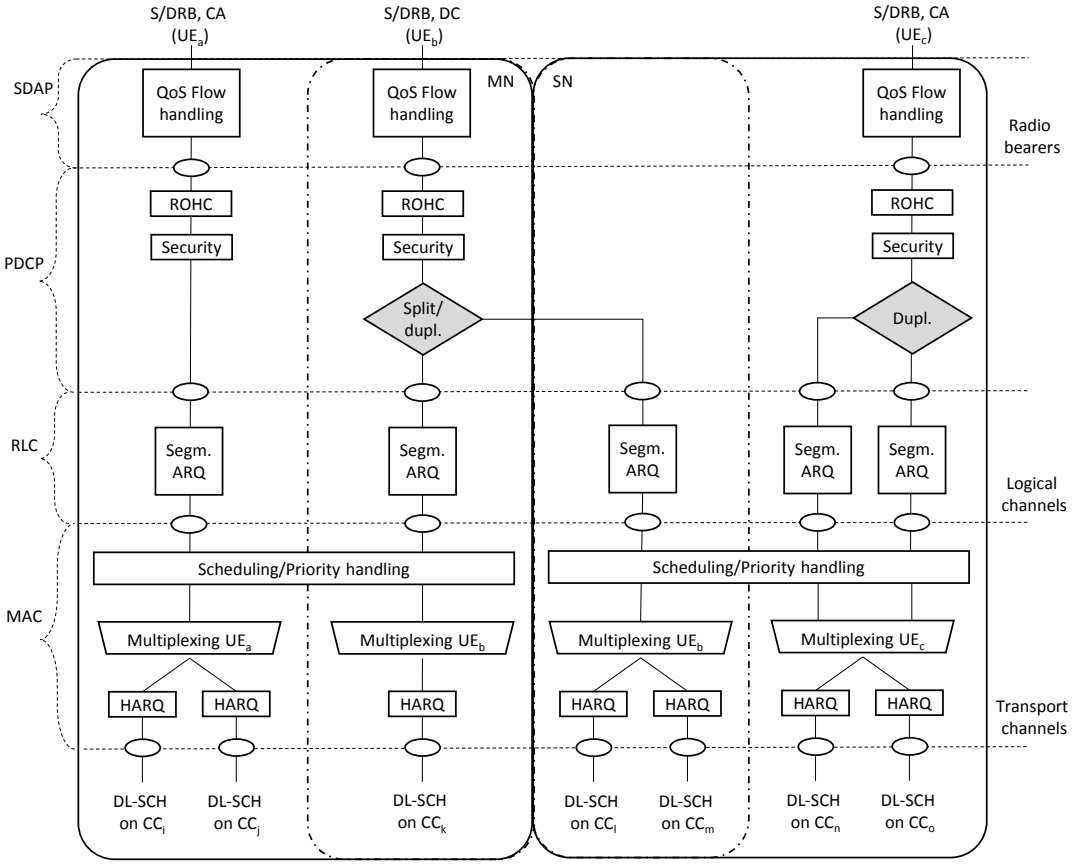


Figure 5.1: Layer-2 structure for downlink NR MC, including traffic split/duplication functionalities at the NR PDCP and MAC layers.

5.1.3 Component carrier management

This section describes a novel RRM functionality, named component carrier manager (CCM), to jointly address the challenges of the CC selection within a given node and the UE-BS association rules for the SNs. In this way, assuming that a UE has previously selected a PCell, and thus, a MN, following the cell selection/reselection criteria defined by network operators, the CCM is in charge of selecting which CCs will host its additional SCells and/or PSCells. The CCM is a functional block to be located at the RRC layer, thus being able to manage and retrieve performance information from the different layer-2 user plane and control plane protocols. One CCM entity operates for each S/DRB for a given user. Specifically, the CCM aims at determining:

- The **number of CCs** to be assigned to a UE. In general, a higher number of CCs for a given user implies enhanced performance metrics. That is, higher throughputs or higher values of reliability.
- The **carrier indices**. In order to fairly share the time and frequency resources among the different users, as well as to fight time-varying fading effects, each user is assigned not only a number of CC, but also their absolute radio frequency channel number (ARFCN).
- The specific **usage of the CCs that were assigned**. In principle, carriers for eMBB users will be used to increase these users' accessible bandwidth, whereas carriers for URLLC users will be used to add redundancy in the form of carriers holding a duplicated data flow. In Figure 5.1, the CCM would be in charge of configuring the NR PDCP and MAC layers to perform the inter- or intra-node data split or duplication.
- The **source nodes providing the CCs**. Each CC is identified by a pair: the index of the node providing it and the actual index of the CC within this node. Thus, the determination of such pair of indices implies establishing the BS that the UE will use as a SN.

To that end, a score is periodically computed for both the CCs managed by the MN and those managed by its neighboring nodes, where each score stands for the suitability of a CC according to previously defined network experts' policies. For these scores to be computed, several sources of performance information are used, like performance information from the CCs themselves (e.g., their load) or metrics derived from UE measurement reports (e.g., RSRP). For the CCM to have a more user-centric vision, service-specific quality metrics may also be taken into account; for example, the initial buffering time or the number and duration of stalls in the case of video services. Currently, these metrics are not directly accessible to the network, since the topmost user-plane data are IP packets at the SDAP layer. These metrics could either be estimated using a model relating SDAP performance indicators to service-specific quality metrics or be periodically reported by the UE to the network, following a procedure still to be standardized. On the other hand, these service-specific metrics could not be reported by a given UE for a given CC unless this UE is already using such service over such CC. This issue could be overcome in the following two ways, which exploit time and space locality, respectively. Regarding time locality, a value for the current service-specific metrics for this UE could be estimated from values reported in the near past by this UE, using regression-based techniques. Concerning space locality, these metrics could be retrieved from neighboring UEs, currently using this service over this CC. At this point, user context, like their location and speed appear as valuable inputs to be considered

in the overall CCM framework. This information may be retrieved from network-specific tools, such as the timing advance, or by means of global navigation satellite systems, whose location reports may be integrated in user call traces.

The current physical location of the CCM depends on the network architecture. In case that a centralized radio access network (C-RAN) is deployed, the CCM will be located at the baseband unit (BBU), taking advantage of its ability to steadily monitor the performance of every node through low-latency and high-capacity backhaul links. On the contrary, if a distributed RAN (D-RAN) is deployed, each node would have its own CCM. In this case, each node should exchange its performance information with its neighboring nodes, so that every node always has an updated vision of its neighbors' performance. Despite performance data are periodically stored in the OSS, the storage period usually ranges from fifteen to sixty minutes, which makes these data to become obsolete for short-term RRM functionalities, like the CCM. The required mechanism for the exchange of performance information should be frequent enough to provide a reliable vision of each node current status, but slow enough not to incur excessive signaling load in the backhaul links or computational cost at the nodes.

Provided that the performance information has been gathered, the score for each CC may be computed. This task may be accomplished using different ML approaches. In order to easily integrate network operators' policies, rule-based systems might be followed; in particular, a FLC may be used, which has been shown as a valuable tool for network management and optimization, [36]. Once every CC (both the CCs of the MN and those of its neighbors) gets a score, they are sorted in a descending way, and only those above a threshold defined by network experts may finally be assigned to a UE, after admission control is performed. Figure 5.2 shows an example of this in a D-RAN. In this figure, BS_a (left) acts as the MN for the UE, being in charge of assigning additional SCells and PSCells to the UE. In the table shown in this figure, CC_{mi} stands for the i^{th} CC provided by network node m; $x_{mi}^{(n)}$ stands for the n^{th} performance metric assessed for CC_{mi}, and S_{mi} stands for the resulting score of CC_{mi}, according to $\{x_{mi}^{(1)}, x_{mi}^{(2)}, \ldots, x_{mi}^{(N)}\}$ and the network operators' policies followed. In the case of Figure 5.2, the CCs above the threshold are CC_{a3}, CC_{b4} and CC_{b1}. Therefore, CC_{a3} will host a new SCell in BS_a for the UE, and CC_{b4} and CC_{b1} will host a PSCell and SCell in BS_b, respectively, setting BS_b as a new SN.

5.1.4 Proof of concept

Experiment setup

Extensive dynamic system-level simulations have been carried out to assess the benefits of the CCM in the context of a 5G NR network. In particular, in this proof of concept, its capability to simultaneously enhance the user experienced performance and integrate network operators' policies for load balance for eMBB services is assessed. For this test, a Matlab simulator based on [75] has been used. Table 5.1 shows the main simulation parameters. The simulated scenario consists of a macrocell D-RAN, composed of 12 tri-sectorial nodes, deployed in a realistic layout, where each sector is made up of five co-located CCs with a bandwidth of 1.4 MHz. All the UEs are MC-capable and the maximum number of CCs that may be allocated per user is five,

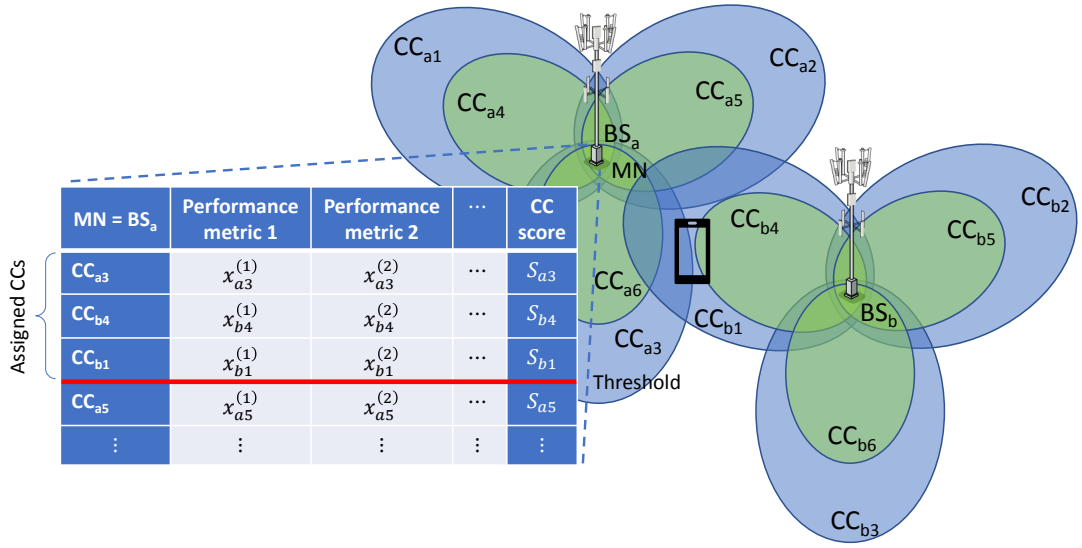


Figure 5.2: Example of operation of the CCM. Having BS_a as the MN for a given UE, its CCs scores and those of its neighbors (BS_b) are computed and sorted, assigning those beyond a certain threshold to the UE.

according to an early deployment of MC-capable UEs.

As a first approach, both the UEs and the selection of the CCs are static. That is, there are no HOs and once the CCs have been selected for a given UE, these are held throughout the duration of each UE's connection, respectively. This implies that CCs are added, but not released or changed along the simulation. Besides, the user plane is split at the gNBs (rather than in the CN), as this provides higher flexibility in simultaneously using the spectrum available in the MN and the SNs [57].

To simulate a load imbalance, part of the users are deployed in a uniform way throughout the scenario and another part of them are deployed in the shape of a hot spot (that is, a region of a high density of users; in this case, with three times the user density of the uniform deployment) along the coverage area of different nodes. As it can be seen in Figure 5.3, the hot spot surrounds site number 5, which will be severely loaded, followed by site number 10. A situation of load imbalance makes the affected users to experiment throughput limitations, due to the scarcity of radio resources for new connections.

Regarding the selection of the CCs, the PCell is the strongest cell (RSRP) [4]. However, for the selection of the PSCells and the SCells, three cases are distinguished:

- **Baseline**: PSCells and SCells are added according to the A4 event (a neighbor cell becomes better than a certain threshold), based on the received power level (RSRP), Table 5.1. This situation represents the baseline case.
- **CCM$_1$**: PSCells and SCells are added using the CCM. In this case, two inputs have been considered:
  - The UE-reported **RSRQ**, in dB.

Table 5.1: Main configuration parameters.

Parameter	Configuration
Scenario	12 tri-sectorial nodes, 5 CCs per sector
Average inter-node distance	2000 m
Direction of transmission	Downlink
Band central frequency	2 GHz
Bandwidth	1.4 MHz (6 PRBs)
Frequency reuse	1
Propagation model	Okumura-Hata Shadowing: log-normal, σ = 8 dB Correlation distance = 50 m
Channel model	ETU model
Mobility model	Static users
Base station model	Tri-sectorial, SISO, P_{TXmax} = 43 dBm
Time resolution	100 ms (1 TTI = 1 ms)
Traffic distribution	Non-uniform distribution of users
Traffic type	Finite buffer Packet size = 2 MB Poisson arrival
Threshold for A4 event	RSRP-based: -120 dBm

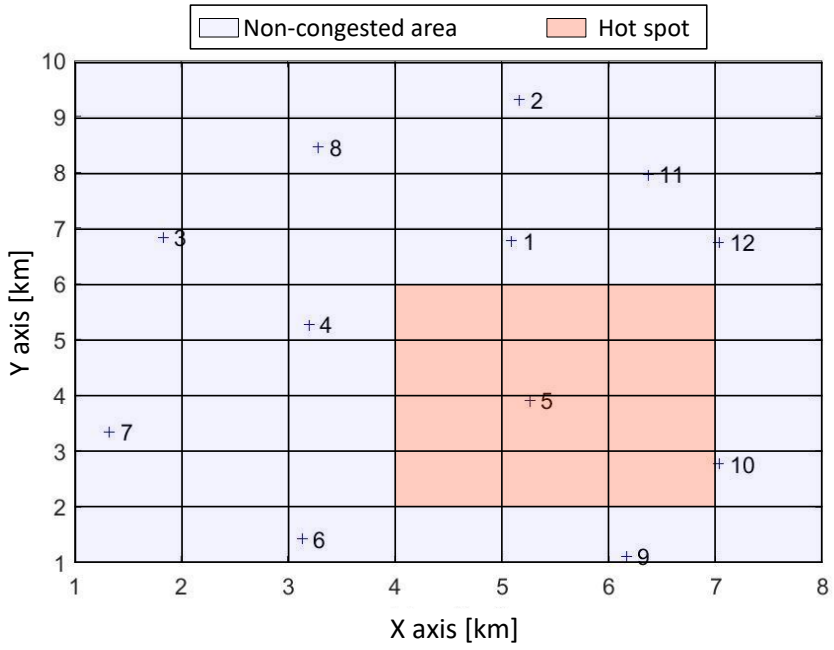


Figure 5.3: Simulation scenario.

– The **CC load**, expressed as a percentage and computed as the free PRBs to the total number of PRBs for a given CC.

- **CCM$_2$**: PSCells and SCells are added using the CCM. Apart from the inputs considered

in CCM_1, an additional input is assessed:

– In case that a CC is provided by a SN, the **usage of the Xn interface** between the SN and the MN due to the user plane load is considered as an input as well. This input metric is computed as the number of UEs currently using resources from these two nodes.

The CCM is implemented as a FLC, whose rules are shown in Table 5.2 for CCM_1 and in Table 5.3 for CCM_2. These rules are built over an AND operator. For example, according to rule 1 in Table 5.2, in order to get a high score for a CC for a given UE, this CC should be scarcely loaded and the UE should have reported high values of RSRQ for this CC, simultaneously. If the CC under evaluation belongs to the MN, the usage of the Xn interface is considered as low. For CCM_1, only the antecedents concerning RSRQ and CC load are assessed; for CCM_2, the Xn usage is assessed as well. Regarding the used threshold, only the CCs scoring medium or better are considered to be finally assigned to the UE.

Table 5.2: Rules implementing network operators' policies in CCM_1.

Rule number	CC load level	RSRQ	Score
1	Free	High	High
2	Medium	High	Medium
3	-	Low	Low
4	Occupied	-	Low

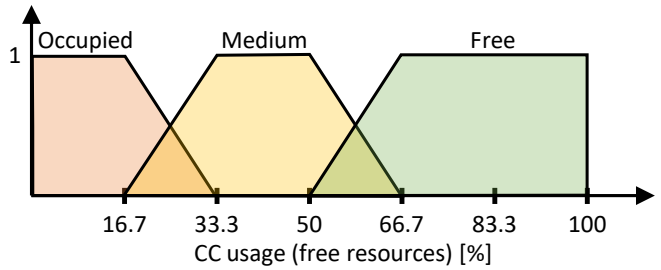
Table 5.3: Rules implementing network operators' policies in CCM_2.

Rule number	CC load level	RSRQ	Xn usage	Score
1	Free	High	Low	High
2	Free	High	Medium	Medium
3	Medium	High	Low	Medium
4	-	Low	-	Low
5	Occupied	-	-	Low
6	-	-	High	Low

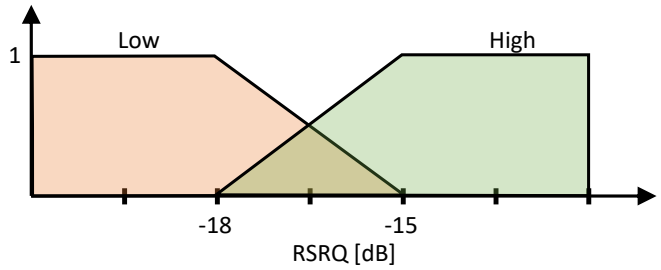
Figure 5.4 shows the membership functions that have been defined over the input metrics. These functions allow mapping crisp values of the input metrics (i.e., CC load, RSRQ and Xn usage) into different categorical labels, which afterwards feed the rule system. These membership functions have been first defined according to expert's knowledge and afterwards heuristically tuned to optimize the throughput response of the system.

Results and discussion

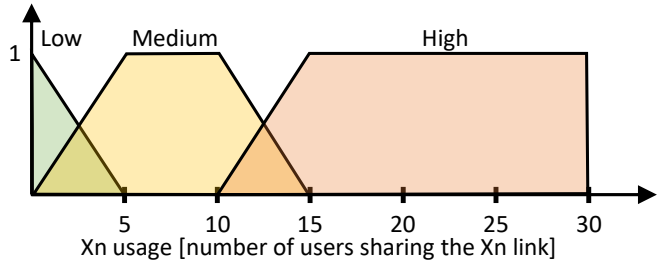
The benefits of using the CCM for the selection of the PSCells and SCells are shown next. Figure 5.5 shows the 5^{th}, 50^{th} and 95^{th} percentile of the UE throughput for both the baseline and the



(a)



(b)



(c)

Figure 5.4: Membership functions for the input metrics used in the CCM.

CCM cases. In this figure, the throughput values have been grouped according to the number of CCs finally assigned to each UE. As it can be seen, an increasing number of CCs implies increasing throughputs in all the three cases. However, when these cases are compared among each other, the selection of the CCs according to the CCM_1 and CCM_2 outperforms the baseline case. Considering both the CC load level and the UE-reported signal quality allows the throughput to be increased, as the RSRP (baseline case) does not give an idea of the available radio resources in a CC. In CCM_2, the additional consideration of the Xn usage provides supplementary information for load balancing, allowing further enhancing the UE throughput by means of a more proper inter-node load sharing.

On the other hand, Figure 5.6 shows the usage of the Xn interface on a node basis, as the sum of the users simultaneously connected to that node and its neighbors (i.e., making use of the Xn interface between these nodes, given the split bearer configuration). Afterwards, results have been normalized by the worst case. This figure shows the effect of the hot spot around nodes 5 and 10, which is especially noticeable in the baseline case, as the CC selection is unaware of both the intra- and the inter-node load level. Only when this information is included with CCM_1

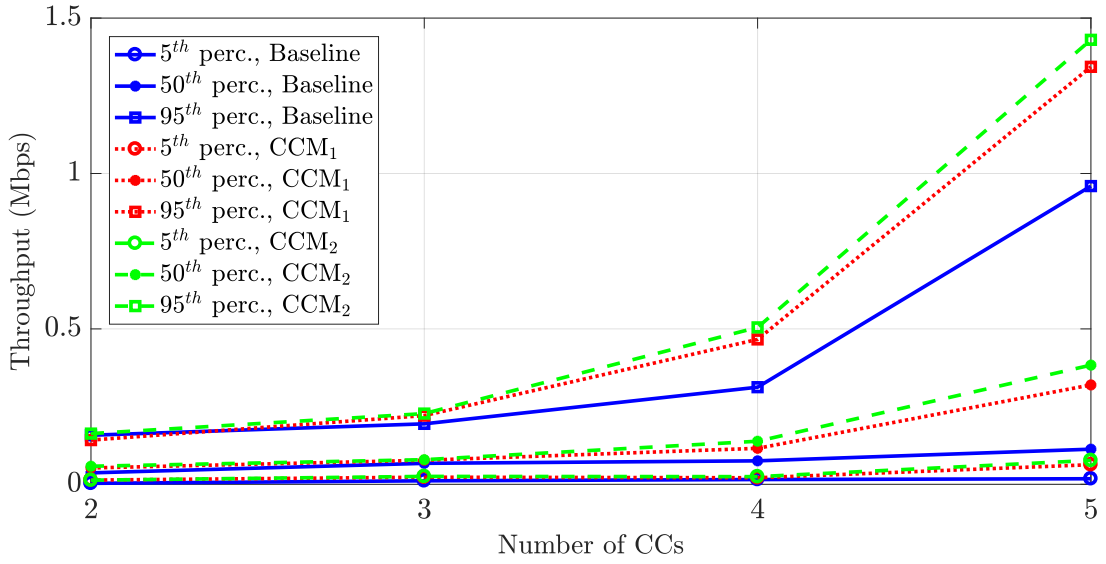


Figure 5.5: Resulting 5^{th}, 50^{th} and 95^{th} percentile for UE throughput, grouped according to the number of CCs finally assigned for the three cases of CC selection: baseline, CCM_1 and CCM_2.

and extended with CCM_2 the traffic load can be efficiently distributed among different nodes. Figure 5.6 also shows how the proposed method does not only allow properly assigning a set of CCs to comply with network operators' rules, but it limits multi-connectivity with specific nodes when the environment is not suitable. For example, whereas the baseline RSRP-based approach for UE-BS association leads to the unaware use of multiple links involving node 5, the proposed method (both CCM_1 and CCM_2) reduces the number of users which use node 5 either as a MN or as a SN. Instead, those UEs will be likely assigned CCs belonging to nodes like 4 or 6, even leading to exclusively using resources from such nodes, using CA. On the other hand, the *mean* case, in which the worst offenders (i.e., nodes 5, 10, 1 and 6) have not been considered, shows how multi-connectivity is scarcely penalized if the situation does not require it.

5.2 Optimizations for low-latency communications traffic through dynamic multi-path connections

In the previous section, a general framework to enhance the user performance is provided, using the concept of multi-connectivity. Despite this framework is suitable for any service category, its benefits have only been tested for the eMBB case. In Section 5.2, however, low-latency communications are addressed. In particular, the latency of the delay-sensitive messages in a V2X scenario is reduced taking advantage of the two interfaces involved in V2X communications in 3GPP: the Uu interface and the PC5 interface.

This section is organized as follows. In Section 5.2.1, the related work is outlined. The problem formulation is stated in Section 5.2.2. Next, in Section 5.2.3, the method to reduce the E2E latency for V2X communications is described. Finally, a proof of concept is carried out in Section 5.2.4 by means of simulations.

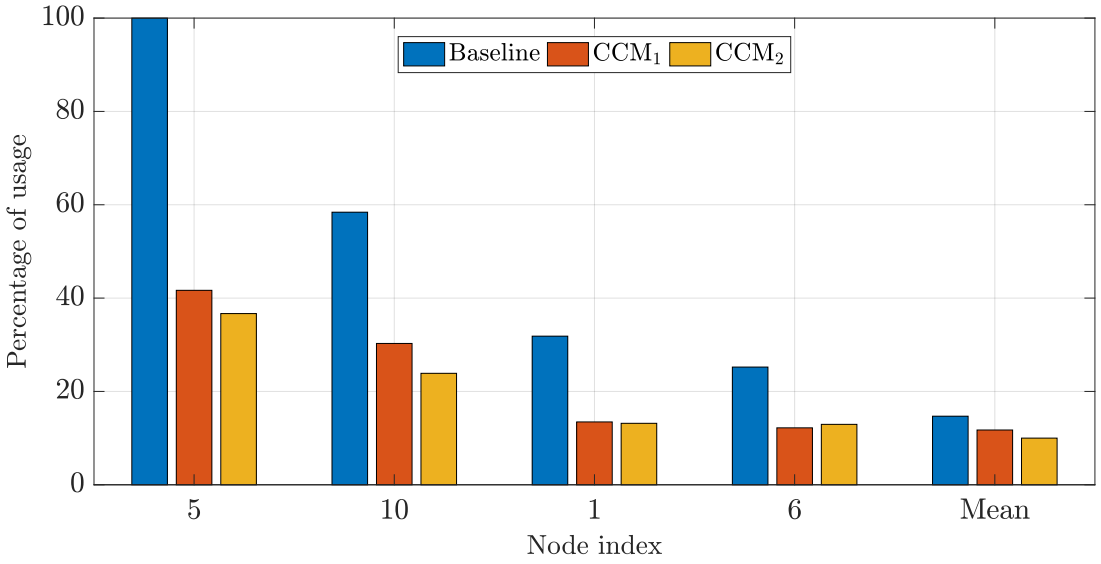


Figure 5.6: Xn usage, computed as the sum of users simultaneously using each node and its neighbors (horizontal axis) and normalized to the worst case. In this figure, only the worst cases are shown. The mean case is computed over the eight remaining nodes.

5.2.1 Related work

In the last years, vehicular communications have taken advantage of several radio access technologies (RATs) to access a variety of services with quite dissimilar requirements. Over time, different approaches for wireless connectivity have been followed; first, in the shape of single-RAT networks and then, as heterogeneous or multi-RAT networks. The latter have drawn special attention in the last five years, mainly by combining dedicated short-range communications (DSRC), like the IEEE 802.11p standard, with cellular communications, like LTE. Some examples of this can be found in [113–116]. In [113], the latency for cooperative awareness messages (CAM) in V2X communications is reduced by performing a handover between LTE and IEEE 802.11p, estimating the expected latency from the channel load. In [114], video traffic is steered between these RATs following a QoS-aware approach. In [115], several design guidelines are given for the implementation of vehicular applications on top of this multi-RAT environment. Finally, in [116], a traffic steering mechanism which relies on a FLC to improve seamless mobility is described.

Despite all these works success in improving QoS, their multi-RAT nature implies some issues when coming to a practical implementation. For example, it is unlikely that both RATs are managed by the same mobile network operator (MNO), which eventually makes a joint network optimization quite difficult or even impossible. Thus, it would be preferable to keep the joint usage of short-range communications and cellular communications under the same umbrella. To that end, the 3GPP standardized a cellular network-supervised set of short-range links for V2X communications [117], among which the PC5 interface stands out. Therefore, Section 5.2 focuses on the joint usage of the Uu (the LTE and 5G NR radio interface) and the PC5 interfaces for the QoS optimization under a 3GPP-compliant framework, making use of the knowledge acquired from past multi-RAT solutions.

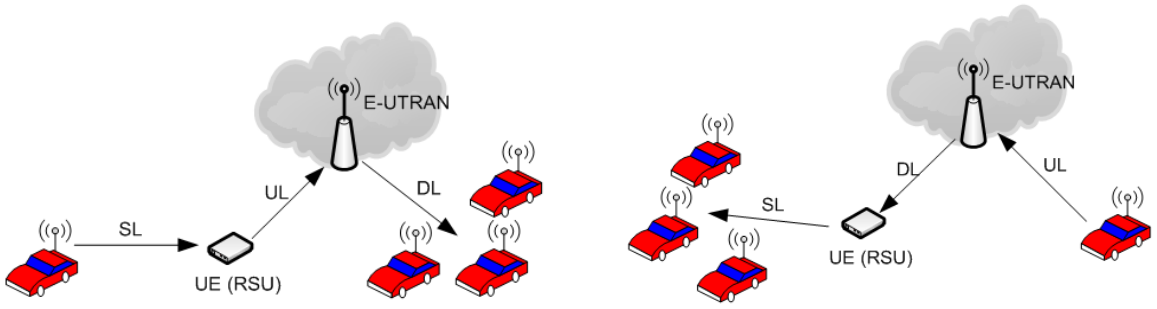
5.2.2 Problem formulation

In [118], two interfaces are differentiated for the vehicles to communicate with other elements: the Uu interface, being the air interface of the LTE and 5G NR networks, and the PC5 interface [117], operating in the band for intelligent transportation systems (ITS) applications, 5.9 GHz. Whereas the Uu interface allows the vehicle to communicate with the network base stations (either macro or small cells), the PC5 interface enables sidelinks (SLs), usually deployed between vehicles or between these and roadside units (RSUs). However, as stated in [118], both interfaces are suitable for the provision of the different kinds of vehicular services; namely, vehicle-to-vehicle (V2V), vehicle-to-infrastructure or network (V2I/N) and vehicle-to-pedestrian (V2P) services, not limiting one of these to a particular interface. This fact, together with the differentiated traffic profiles of V2X, makes necessary an intelligent interface selection in each case.

Depending on the availability of the Uu and PC5 interfaces, 3GPP defines three scenarios for V2X communication [118]:

- Scenario 1: Only the PC5 interface is available. Thus, both the uplink and downlink must rely on SLs.
- Scenario 2: Only the Uu interface is available. In this case, every V2X communication must rely on a direct vehicle-to-eNB link.
- Scenario 3: Both interfaces PC5 and Uu are available (see Figure 5.7). This scenario is, in turn, broken down into two variants. In scenario 3A (Figure 5.7a), the PC5 interface is used for uplink by means of a SL to a UE roadside unit, which eventually interfaces an eNB via the Uu interface. In downlink, the eNB directly interfaces a number of UEs using a broadcast mechanism via Uu. In scenario 3B (Figure 5.7b), a UE directly transmits to an eNB via the Uu interface in uplink, whereas the downlink is supported over a UE RSU, communicating with different UEs through PC5 SLs.

The work described in Section 5.2 focuses on scenario 3. In particular, it focuses on the decision on whether to use a SL (scenario 3A) or use a direct link (scenario 3B) to deliver uplink delay-sensitive V2X messages.



(a) Scenario 3A [118]. In the uplink direction, a UE interfaces a UE RSU via the PC5 link, which eventually interfaces an eNB.

(b) Scenario 3B [118]. In the uplink direction, a UE directly interfaces an eNB via the Uu link.

Figure 5.7: Scenarios supporting V2X operation using both Uu and PC5 interfaces, according to 3GPP.

5.2.3 Proposed method for traffic steering

For the following method to be applied, it is required that UEs have both the Uu and the PC5 interfaces available. Besides, it is assumed that both the RAN and the UE can make their own routing decisions in terms of which interface to use when delivering downlink and uplink messages, respectively, based on the assessment of high-layer metrics. For example, the UE sending an uplink message in Figure 5.7 could send such message through either the PC5 or the Uu interface, corresponding to the scenarios 3A and 3B, respectively.

In order to choose between the different interfaces when a message is to be sent, and in a similar way to how the available CCs are managed in Section 5.1.3, a score is computed for each of them. To that end, a set of performance metrics resulting from the use of one interface or another is evaluated. These metrics may be high-layer performance metrics, like E2E delays, packet-loss rates, etc., and also low-layer metrics, like the measured signal power and quality. The different metrics that are assessed can further be grouped into different sets, which represent different scopes over which optimize the communication. For example: metrics related to a mission-critical behavior, like the delay or the reliability, or metrics which inform about energy issues. Within each of these sets of metrics, every option (PC5 and Uu) can be then characterized formulating a cost function according to the corresponding set or scope. In this way, every interface gets a score regarding each scope. For example, PC5 SL could be the best option in terms of energy consumption, but it could also be the worse regarding the E2E delay.

This grouping into sets, corresponding to high-level scopes, together with the computation of a cost function per interface and optimization scope, allows using a simple and flexible decision algorithm either in each UE or the RAN as the one shown in Algorithm 5.1. Whenever a new packet arrives from upper layers and needs to be sent, Algorithm 5.1 is executed. This is a decision process made up of nested conditions, each of which is related to one of the scopes previously defined by setting a certain criteria over them. If the condition related to the first criterion is fulfilled, then, the interface with the best score in its cost function according to that scope is chosen. If it is not, the second criterion is evaluated. In case that none of the criteria are fulfilled, a default interface is chosen. This type of decision structure enables the proposed solution to be used over a wide range of applications by only setting different conditions over these scopes or by switching these scopes themselves. For example, if the proposed solution is to be used on a mixed-criticality application (e.g., considering both infotainment and delay-sensitive messages), a criticality-related condition should be asked first and cost functions based on delay and reliability metrics should be assessed per each interface in consequence. In such case, other issues like those related to the energy consumption should remain as second order scopes. Moreover, the generality of Algorithm 5.1 and its location under application layers makes it specially proper for the integration of information related to the context.

The cost functions have been introduced as one of the main enablers of the proposed solution. These functions are periodically computed, with a time period entailing a trade-off between energy consumption and the validity of these costs in their attempt to represent the current status of the networks. A periodical update of these costs, performed as a background process, allows the decision algorithm to take into account possible the solution with a dynamic and near

Algorithm 5.1: Decision tree of the proposed solution

```
New packet from upper layers;
if condition related to criteria 1 is fulfilled then
    Send it through the best link regarding scope 1;
else
    if condition related to criteria 2 is fulfilled then
        Send it through the best link regarding scope 2;
    else
        ⋮
        if condition related to criteria n is fulfilled then
            Send it through the best link regarding scope n;
        else
            Send it through the default link;
```

real-time behavior, easily assigning the best interface in each case without the necessity of further assessment.

5.2.4 Proof of concept

Experiment setup

In this proof of concept, an LTE network has been considered, providing the Uu interface. Only one optimization scope has been taken into account, leading to a decision algorithm with a single *if*. In this case, the working scope is the criticality of the messages to be sent, in order to assess the validity of the proposal in a mixed-criticality environment, in an uplink vehicle-to-network/infrastructure (V2N/I) communication. Thus, the corresponding criterion is whether the message from upper layers is labeled as critical or not. In such case, and in order to reduce the impact and degradation of the proposed solution in the overlaying LTE network, the PC5 sidelink has been set as the default link, that is, the link selected for non-critical messages. In this way, the Uu interface is only considered if a message is labeled as critical. In this V2N/I communication, a remote V2X server is assumed to be located beyond the packet data network (PDN) gateway (PGW) of the LTE network.

Thus, it has been assumed that every UE has two network interfaces: a PC5 interface, in order to communicate with the RSU, and an Uu interface, to communicate with an eNB. In this way, two uplink transmission modes can be identified: the direct, by which the packets are sent from a UE to an eNB through the Uu interface (scenario 3B), and the indirect, by which the data packets are sent to the eNB via an intermediate access point (scenario 3A), leading to a PC5 + Uu link.

The cost functions for both direct and indirect links (C_{Direct} and $C_{Indirect}$, respectively) under the scope of criticality have been addressed by means of a delay metric, the round-trip time (RTT), measured over user datagram protocol (UDP). The RTT is included into each cost

function as the moving average over the last three E2E RTT measurements in that link, Eq. (5.1). In order to measure the RTT of the messages, an echo application has been enabled on both the remote V2X server and the UEs, so every incoming UDP packet is sent back throughout the same path whenever it arrives at the V2X server. Thus, in this case, the link (L) over which sending the incoming packet is the one providing the lowest cost value, Eq. (5.2):

$$C_{Direct}[n] = \frac{1}{3}\sum_{n-2}^{n} RTT_{Direct}[n] \tag{5.1a}$$

$$C_{Indirect}[n] = \frac{1}{3}\sum_{n-2}^{n} RTT_{Indirect}[n] \tag{5.1b}$$

$$L = \arg\min\{C_{Direct},\ C_{Indirect}\} \tag{5.2}$$

In order to assess the proposed method, a set of simulations have been carried out using the networks simulator ns-3 [119]. To that end, and as an approximation, the PC5 interface has been simulated using the IEEE 802.11n standard, in the 5 GHz band, using WiFi access points (APs) as the RSUs, and the Uu interface has been simulated using the LTE module of ns3 [120].

To show the benefits of the proposed method, the cumulative distribution function (CDF) of the E2E RTT is assessed in two cases: having a low UE density (leading to a situation of low traffic) and a high UE density. Besides, two baseline situations have been also simulated: the case in which all the messages (critical or not) are sent through the PC5 + Uu link (baseline experiment 1) and the case in which all the messages are sent through the Uu link (baseline experiment 2). In all these cases every node sends 256-byte messages. These messages are labeled as critical with a 5 percent of probability. The packet arrival is given by a Poisson process with four messages per second on average.

The simulated scenario consists in a Manhattan-like deployment in a 100 m × 100 m area. This area is made up of 4 horizontal and 4 vertical 100-m streets, in which a number of UEs have been randomly deployed. In this scenario, the UEs move at a constant speed of 15 km/h and have a 0.5, 0.25 and 0.25 probability of following straight ahead, turning left and turning right at each intersection, respectively. In order to retain the UEs in the simulated area, a wrap-around technique has been used beyond the borders of this area. In this scenario, four RSUs and a macro eNB have been deployed as it is shown in Figure 5.8. The main simulation parameters have been summarized in Table 5.4.

Results

The results of the baseline experiments 1 and 2 are shown in Figure 5.9 (a and b, respectively). Regarding baseline experiment 1 (Figure 5.9a), it can be observed how the increase in the number of UEs, and therefore, the increase in the traffic load, noticeably impacts the performance of the PC5 + Uu link, increasing the 90^{th} percentile of the RTT from 37.4 ms to 529.9 ms. It is worth noting that this increase takes place for both the critical and non-critical messages, since

Table 5.4: Simulation parameters.

Parameter	**Value**
Scenario setup	
Simulation tool	*ns-3*: WiFi and LTE (LENA) modules
Number of UEs	15 (low traffic), 75 (high traffic)
UDP packet size	256 bytes
UDP packet interval, Poisson process, $1/\lambda$:	0.25 s
Number of WiFi APs	4
UE speed	15 km/h
Simulation time	60 s
PC5 interface	
Standard used	IEEE 802.11n
Band	5 GHz
Channel bandwidth	20 MHz
Operating mode	Infrastructure
Uu interface	
Downlink carrier frequency	945 MHz
Uplink carrier frequency	900 MHz
System bandwidth	20 MHz
Number of eNBs	1
Sectors per eNB	1
Interferences	Other cell interf. was not simulated
Shadowing	Log-normal, $\sigma = 8$
PGW $\leftrightarrow$ remote host delay	10 ms

no distinction is made at this point. This increase is mainly due to the usage of a contention-based protocol in the medium access procedure, since the probabilities of finding the medium busy increase quickly with the number of devices. Despite the current PC5 interface does not actually implement a CSMA (carrier sense multiple access) mechanism for medium access (as IEEE 802.11n does), a similar behavior to this may appear when the UEs self-allocate time and frequency resources from a relatively small resource pool, as it is the case in the operation mode 4 for the PC5 interface. This would likely take place in a dense scenario, such as an urban area where resources may be allocated to other users.

Regarding baseline experiment 2, Figure 5.9b shows that the increase in the traffic has a much lower impact in the delay, increasing the 90^{th} percentile of the RTT from 36.8 ms to 43.7 ms. Unlike WiFi, in which, every new packet to be sent needs a contention-based medium access to be performed in advance, in LTE, the contention-based medium access is only performed if frequency and time resources have not been assigned yet to the current device. Once these are allocated, the device holds them during a certain time period, not needing to further perform a contention-based random access procedure and thus, reducing drastically the E2E RTT. In these experiments, the mean time between packet arrivals is 0.25 s, whereas the time window

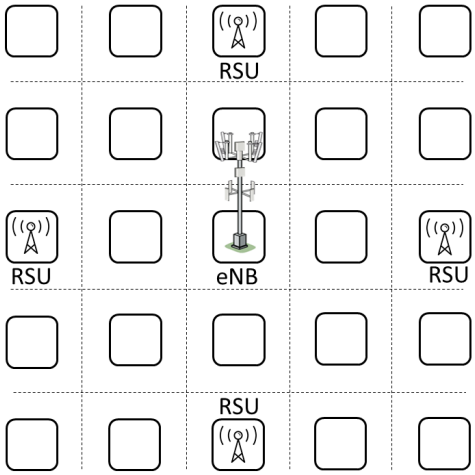


Figure 5.8: Simulation scenario.

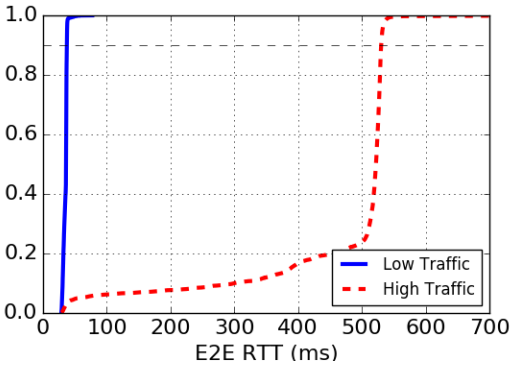
along which the LTE resources are held in a given connection after its activity has ceased is 10 s, which is a commonly used value in commercial networks. In this way, the time delay due to the contention-based random access is only added to the first packet transmitted by a given UE, leading to the short RTTs that are shown in Figure 5.9b. Even though, the effect on the E2E delay of a higher number of devices contending for an access grant can be seen in this figure, as a result of some collisions during the random access procedure. Besides, a higher traffic load at the scheduler in the eNB also contributes to a higher delay. Despite these results, the direct connection through Uu should not be abused, since a congestion can be produced in the access process under certain conditions, impacting not only on the performance of the present proposal, but on the performance of traditional users employing the LTE network as an MBB (mobile broadband) communications system.

The results when the UEs follow the proposed algorithm and both interfaces are available are shown in Figure 5.10 for the four combinations of traffic load and message criticality. For non-critical messages (striped green and dash-dotted blue lines) the behavior of the E2E RTT is rather similar to baseline case 1, as all these messages are sent through the indirect link. However, it should be noted that, despite having a large mean E2E RTT in the non-critical high-traffic case, this delay is noticeably lower than the one obtained in the baseline case 1. This is because in the latter, all the messages, critical or not, are forwarded throughout the same link, leading to a stronger contention in the medium access. In Figure 5.10, however, only the non-critical messages and those critical messages that find the PC5 interface as a faster option are forwarded through the WiFi AP, leading to a lower number of medium access attempts and thus, to a lower delay due to this contention-based mechanism.

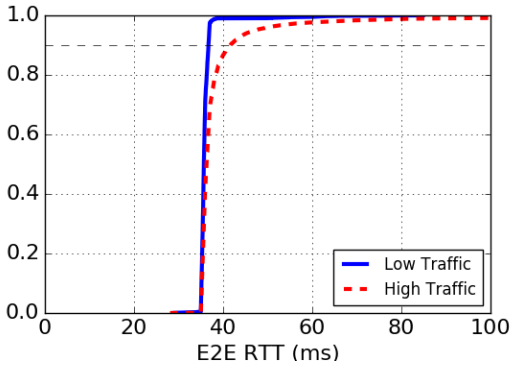
The case with the SL being faster than the direct link arises when the mean time between packet arrivals at the WiFi AP ($1/(\lambda \cdot N_{UE})$, where N_{UE} stands for the number of UEs using the SL towards the RSU at that time) is similar or higher than the LTE inactivity time: 10 s, in this case, and N_{UE} is such, that the delay due to the contention-based access in the WiFi network is shorter that the one in the LTE random access procedure. When the messages are spread enough

in time it is quite likely that every new packet to be sent through the direct link would require a random access procedure to be performed in the Uu interface, increasing the E2E delay. On the other hand, and given that the WiFi gateway holds a single connection which gathers messages from many UEs, the net mean time between packet arrivals at the gateway could be significantly lower than the defined inactivity time in LTE. This would make the allocated resources for the gateway-to-eNB connection not to be released, not needing to perform a random access procedure every time a new packet arrives at the gateway, and thus, not adding an extra term to the RTT of the packets through the indirect link.

Besides, critical messages show a more stable behavior; solid orange and dashed red lines, respectively. Figure 5.10 shows that, with low traffic, both critical and non-critical messages experiment a similar and low-delay profile (being 37.8 ms and 38.4 ms their RTT 90^{th} percentile, respectively) and that, in a high-traffic scenario (dashed red and dash-dotted blue lines), the proposed method provides an effective differentiation for a mixed-criticality service, providing an optimal delay for critical messages (an RTT 90^{th} percentile of 37 ms) and a delay for the non-critical ones much lower than those of the only-PC5 baseline case, having an RTT 90^{th} percentile of 80.2 ms.



(a) Empirical CDF of the measured E2E RTT in the baseline experiment 1: always use the PC5 + Uu links.



(b) Empirical CDF of the measured E2E RTT in the baseline experiment 2: always use the Uu link.

Figure 5.9: Results of the baseline experiments.

In light of these results, and under these assumptions, all of the use cases described in [121] for LTE-based V2N/I and some of the ones described in [122] for 5G-based V2N/I communications could be addressed in terms of expected E2E latency.

5.3 Conclusions

In line with 3GPP Release 15, this chapter has covered the eMBB and low-latency communications service categories, in an attempt to optimize some of their most representative performance metrics through multi-link management.

In particular, the benefits of using MC have been proven for eMBB traffic in terms of an enhancement in throughput. Simulation results in a load-imbalanced scenario show how a CCs assignment according to network- and UE-specific metrics different from RSRP may lead UEs to

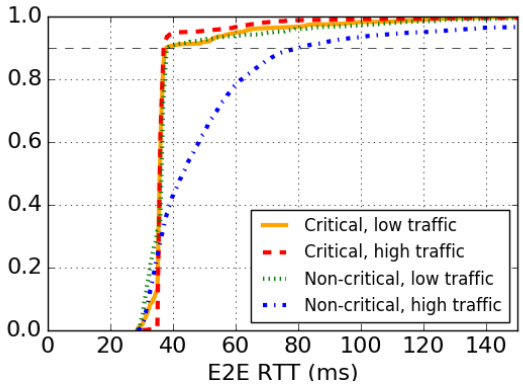


Figure 5.10: Empirical CDF of the measured E2E RTT using the proposed solution to decide to use either the direct Uu link or the indirect PC5 + Uu link regarding past RTT measurements from each path.

increase their 95^{th} percentile for throughput up to 50% when compared to a traditional RSRP-based CC management scheme. Besides, and thanks to the proposed RRM functionality for CC management, the load was successfully balanced.

Next, a method has been proposed in the field of low-latency V2X communications to steer messages over the Uu and the PC5 interfaces in order to provide a minimum delay for messages labeled as critical. To that end, cost functions over delay-related metrics are computed on a per-interface basis as a means to evaluate its suitability at a given time. Results from simulations show how the proposed method allows critical messages to neglect the effect of high traffic loads leading to low delays, while providing a best-effort behavior for non-critical ones.

Chapter 6

Conclusions

This chapter summarizes the main contributions of this thesis. In addition, some future lines of work are suggested. Finally, a list of publications related to this thesis is presented.

6.1 Contributions

This thesis is focused on the enhancement of self-healing and self-optimization tasks within SON, following a ML approach. In particular, regarding self-healing, new proposals for a more accurate RCA are included, improving both the diagnosis techniques themselves and the way that performance information is processed. Concerning self-optimization, mechanisms for performance enhancing are proposed, making use of a multi-link management. The contributions of this thesis are detailed in the following lines:

a) Self-healing:

- **Development of new diagnosis techniques for an enhanced RCA**. The existing literature on RCA in mobile communication networks has proposed some automatic diagnosis systems based on different artificial intelligence techniques, each one with its own pros and cons. When an operator decides to include a tool for RCA in its network, the first decision is which automatic diagnosis technique to use. Once the technique is selected, a diagnosis model should afterwards be chosen. Instead of choosing a single technique and model, in this thesis, a method to combine any number and kind of different standalone diagnosis systems has been proposed, so that the ensemble diagnosis system provides a higher performance than those of its standalone diagnosis systems. In particular, the proposed method allows both combining diagnosis systems based on different artificial intelligence techniques and combining different diagnosis models (i.e., models based on the same technique, but whose parameters are different, because they have been built by different experts or obtained from different training datasets). To that end, an algebraic combination of the statistical behavior model

of the standalone diagnosis systems is carried out. Results have shown that the performance of the proposed method is significantly better in terms of DER compared to its standalone components using both simulated data and cases from a live LTE network, as long as the accuracy of the baseline diagnosis systems is similar and their diagnosis errors are uncorrelated. Furthermore, the proposed method relies on concepts which are not linked to a particular mobile communication technology, but on pattern classification theory, allowing it to be applied either on well established cellular networks, like LTE, or on forthcoming technologies, like 5G NR.

- **Usage of dimensionality reduction to fully automate RCA while reducing CAPEX and OPEX.** SONs aim at automating the management of cellular networks. However, tasks like the selection of the most appropriate KPIs for RCA are still carried out by experts, which is one of their most-time consuming tasks and a cause for suboptimal RCA performance due to the human bias. In this thesis, dimensionality reduction techniques have been studied and used as the enablers for the full automation of RCA in self-healing. In particular, the main contributions in this topic are the following:

  - The development of an *unsupervised* technique for the selection of the most useful KPIs for RCA, consisting in a data clustering stage followed by a supervised procedure for feature selection. Its unsupervised nature allows the method to determine the most suitable KPIs to discern a set of *a priori* unknown network states. As a result, its main benefit is that it allows using databases of unlabeled data for KPI selection, being this the most common situation in recently deployed networks. Results have shown that the proposed method effectively relieves and outperforms both a troubleshooting expert's selection and state-of-the-art unsupervised techniques for feature selection, allowing a drastic reduction of the volume and complexity of both network databases and RCA techniques, respectively, without human intervention.

  - The development of a *supervised* technique for KPI selection for RCA. Despite supervised techniques need additional information in the shape of labels to operate, they provide better performance than unsupervised techniques for feature selection in the field of RCA. This fact makes them specially suitable for stable cellular networks, in which historical databases of past cases can be found. The proposed supervised technique for KPI selection uses the statistical dissimilarity of a KPI under the presence of different network states as a figure of merit, which is measured as the non-overlapping region of the PDFs of this KPI when separately conditioned to these network states. That is, the more differentiated the statistical behavior of a KPI given two underlying network states is, the more valuable this KPI is to accurately distinguish between such states. Besides, in order to properly model the statistical behavior of a KPI, non-parametric techniques for PDF estimation have been used. Results have shown that the proposed method outperforms other state-of-the-art techniques for supervised and unsupervised feature selection and a troubleshooting expert's selection when data from

a live network are used, demonstrating its validity to fully automate diagnosis tasks when used together with a tool for RCA.

- The assessment of different techniques for feature extraction, to be integrated as an intermediate stage between the monitoring of the network KPIs and their usage in RCA. At the expense of losing the meaning of the resulting KPIs, feature extraction techniques allow condensing relevant performance information in a more reduced set of synthetic KPIs than feature selection techniques. The results of using a set of data collected from a live cellular network have shown the benefits of this approach in terms of storage savings and subsequent improvements to the RCA function. The benefits have been found to be specially relevant when linear techniques for feature extraction are used, given the mostly linear dependence of the monitored KPIs.
- Taking advantage of the insight regarding dimensionality reduction techniques up to this point, a self-healing framework for NG-SONs is proposed, using both feature selection and feature extraction techniques to enhance the RCA accuracy. This framework allows using both labeled and unlabeled data, as well as integrating performance information from different sources, like measurements gathered by the UEs, network KPIs or context information. This framework supports and even relieves troubleshooting experts from dealing with such amount of different data (decreasing the OPEX) and enables a reduction of the network storage needs, as well as the eventual complexity of the self-healing mechanisms (decreasing the CAPEX). Results have shown that the proposed framework can effectively manage a high-dimensional environment from different data sources, eventually automating the tasks usually performed by troubleshooting experts while optimizing the performance of the RCA tool.

- **Definition of call trace-based IRAT performance metrics for enhanced detection and RCA.** Despite traditionally, the most widely used source of performance information for self-healing is network KPIs, call traces provide much more detailed performance information. Although this source of information is not always available, it usually provides user-specific performance data, as well as event timestamps and location data, which allows using more accurate detection and RCA tools as well as taking more specific actions to resolve possible issues in the network. In this thesis, a call trace-based method to assess the performance of an IRAT mechanism, the circuit-switched fallback (CSFB), has been defined. To that end, information from the call trace databases of the source and target RATs are analyzed, allowing an E2E performance assessment, in contrast to state-of-the-art methods, which only allow one-end performance evaluations. By using call trace data from live LTE and UMTS networks, results have shown this method as a valuable tool to detect problematic situations which would remain unnoticed by state-of-the-art methods.

b) Self-optimization:

- **CC management in a MC-enabled scenario to optimize the network performance**. To overcome the disparity of the requirements of the different service

categories to be addressed by 5G NR, a number of RRM solutions have been proposed, among which MC may be highlighted. MC allows a number of benefits to be achieved: from increased reliability due to multi-link diversity, to throughput enhancements, due to the usage of additional radio resources among the involved BSs. With MC, two challenges arise: the selection of the BSs to which the UE should be associated and the selection of the CCs belonging to such BSs to be assigned to this UE. While state-of-the-art works address these two issues separately, in this work, these challenges are tackled together, thus leading to enhanced UE performance and network resource utilization. To that end, a FLC-based RRM mechanism is proposed. The proposed solution, which allows network operator's policies to be defined, due to its rule-based nature, accepts a variety of sources of information as inputs for the CC management, like performance data from UE measurement reports, from the network itself or from the UE context. This enables developing a user-centric approach, in line with the current trends for 5G management tasks. Simulations in an eMBB scenario have shown that, in a scenario of load imbalance, the proposed mechanism allows the UE throughput to be increased while improving the network efficiency by means of a proper inter-CC and inter-BS load balance when compared to traditional RSRP-based schemes for CC management, applied in a MC environment.

- **Dynamic traffic steering for low-latency V2X communications**. Low-latency communications stand as one of the communications types to be covered by the forthcoming cellular networks, and within these, V2X communications occupy a prominent place. They are characterized by the interchange of both low-latency messages for control and notification purposes, and delay-tolerant messages, devoted to infotainment services. In this thesis, a mechanism for dynamic traffic steering over the Uu and PC5 interfaces of V2X communications has been proposed to ensure low latency for critical messages, even in high load conditions. As a proof of concept, a simulation has been carried out using LTE macrocells and IEEE 802.11n access points to provide the Uu interface and an approximation to the PC5 interface, respectively. The proposed solution has proven to effectively steer data packets regarding their criticality label, providing a delay-optimized link for critical messages and a best-effort behavior for the non-critical (delay tolerant) messages.

In addition, as part os his research work, David Palacios has coordinated the deployment of a testbed LTE network at the University of Málaga, which plays a major role in two international projects in which UMA is involved. A description of this network can be found in Appendix C.

6.2 Future work

This work might be continued following the next research lines:

- **Coordination of self-healing and self-optimization functions**. Despite SON functions are usually studied separately, in a realistic scenario, and for the management of a commercial cellular network, several SON instances should be simultaneously deployed.

This situation, however, may lead to conflicts between the involved SON functions; specially, between those devoted to self-optimization. This issue has already been addressed in literature. In works like [123] and [124], for example, the problem of mobility parameter conflicts between MLB and MRO instances is tackled. Project SEMAFOUR [42] has also largely studied the issue of conflict avoidance/resolution between SON functions. However, up to now, this issue has been limited to the study of the conflict of self-optimization functions. A self-healing/self-optimization conflict is yet to be explored; particularly, for the case in which self-healing functions resemble an optimization problem. This is the case of the compensation and recovery self-healing functions.

- **Use of advanced tools for data processing**. This thesis has focused on how the management of cellular networks can benefit from ML. Despite techniques for dimensionality reduction contribute to alleviate the problem of the huge amount of performance information in cellular networks, more steps must be taken for a time- and performance-efficient management of the network. These steps go through considering the most recent frameworks and families of techniques for data processing, like big data or deep neural network learning (mostly known as *deep learning*). The first is a framework to deal with high a volume, variety and velocity of data, and some studies have already been reported regarding their usage in the management of cellular networks: [8, 30]. The applications of the latter are yet to be explored in this field. However, the outstanding results of deep learning in pattern classification (with convolutional neural networks), trend prediction (with recursive neural networks) and dimensionality reduction (with autoencoders) allows considering deep learning as one of the most promising families of techniques for future management tasks in communication networks.

- **Use of multi-fault diagnosis in RCA tasks**. Current techniques for RCA rely on the assumption that, at a given time, the network is affected by a single failure, which needs to be identified [16–34]. However, in a system as complex as cellular networks, in which hundreds or thousands of nodes are interconnected among each other, an initial failure can easily propagate throughout the network, making other failures appear, which need not to be of the same nature of the first one. From the network operator's perspective, this would lead to an unpredictable and chaotic situation, in which a variety of apparently unrelated symptoms appear. To deal with this, a multi-fault (or multi-label, if it is seen from the perspective of pattern classification) approach could be followed. Currently, some of the most widely used techniques for pattern classification have a multi-label counterpart. Decision trees or kNN are examples of techniques with such capabilities.

- **Proactive network management**. Until now, the mechanisms for network performance monitoring rely on techniques with a time-limited vision. These techniques are often based on the comparison of the current network state with absolute or relative thresholds, lacking a deeper time analysis. Despite some works have addressed the network management from a time-series perspective (e.g., in [33]), the focus has been put on determining the *current* network state or performance, not wondering about its *future* state. The usage of time-series techniques to forecast the network performance will allow degradations to be anticipated and avoided, thus taking the necessary actions to prevent users from noticing

their effects.

6.3 Publications and projects

The following subsections present the publications related to this thesis.

6.3.1 Journals

Publications arising from this thesis

[I] D. Palacios, E. J. Khatib and R. Barco. "Combination of Multiple Diagnosis Systems in Self-Healing Networks". *Expert Systems with Applications*, vol. 64, pp. 56-68, 2016.

[II] D. Palacios and R. Barco. "Unsupervised Technique for Automatic Selection of Performance Indicators in Self-Organizing Networks". *IEEE Communications Letters*, vol. 21, no. 10, pp. 2198-2201, Oct. 2017.

[III] D. Palacios, I. de-la-Bandera, A. Gómez-Andrades, L. Flores and R. Barco. "Automatic Feature Selection Technique for Next-Generation Self-Organizing Networks". *IEEE Communications Letters*, vol. 22, no. 6, pp. 1272-1275, June 2018.

[IV] D. Palacios, S. Fortes, I. de-la-Bandera and R. Barco. "Self-Healing Framework for Next-Generation Networks through Dimensionality Reduction". *IEEE Communications Magazine*, vol. 56, no. 7, pp. 170-176, July 2018.

[V] D. Palacios, I. de-la-Bandera and R. Barco. "Multi-Node Component Carrier Management for Multi-Connectivity in 5G New Radio". *IEEE Communications Magazine*. Under review.

Publications related to this thesis

[VI] N. Mahmood, D. Laselva, D. Palacios, E. Mustafa, F. Miltiadis, D. M. Kim and I. de-la-Bandera. "Multi-Channel Access Solutions for 5G New Radio". *IEEE Communications Magazine*, Under review.

[VII] S. Fortes, D. Palacios, I. Serrano and R. Barco. "Applying Social Event Data for the Management of Cellular Networks". *IEEE Communications Magazine*, vol. 56, no. 11, pp. 36-43, Nov. 2018.

[VIII] J. Mendoza, D. Palacios, I. de-la-Bandera, A. Herrera and R. Barco. "On the capability of QoE optimization based on the adjustment of RLC parameters". *IEEE Communications Magazine*, Under review.

6.3.2 Patents

Patents arising from this thesis

[IX] D. Palacios, R. Barco, O. Kaddoura, P. Delgado and I. Serrano, "Inter-technology circuit-switched fallback (CSFB) metrics", WO 2017063700 A1, 2017.

6.3.3 Conferences and Workshops

Conferences arising from this thesis

[X] D. Palacios, E. J. Khatib, I. Z. Kovács, B. Soret, I. de-la-Bandera and R. Barco. "Dynamic Multipath Connection for Low-Latency Vehicle-to-Everything (V2X) Communications". *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Valencia (Spain), 2018, pp. 1-5.

[XI] D. Palacios, R. Barco and I. Serrano. "Combinación de Sistemas para la Diagnosis de Fallos en Self-Organizing Networks", *XXX Simposium nacional de la Unión Científica Internacional de Radio*, Pamplona (Spain), Sept. 2015.

[XII] D. Palacios, I. de-la-Bandera and R. Barco. "Reducción de Dimensionalidad en Funciones SON mediante Técnicas de Extracción de Atributos", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

[XIII] I. de-la-Bandera, D. Palacios and R. Barco, "Asignación Automática de Portadoras en un Escenario 5G", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

Conferences related to this thesis

[XIV] J. Burgueño, I. de-la-Bandera, D. Palacios and R. Barco, "Modelado de TCP en un Entorno Celular con Dual Connectivity", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

[XV] J. Mendoza, A. Herrera, D. Palacios, I. de-la-Bandera and R. Barco, "Optimización de la QoE de un Servicio de Video Streaming en un Entorno Celular", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

6.3.4 Related projects

This thesis was funded by the project P12-TIC-2905 "Gestión integral avanzada de funciones SON (Self-Organizing Networks) para redes móviles futuras", from the Junta de Andalucía.

This thesis has also contributed to the following projects:

- International projects:

  - ONE5G: E2E-aware Optimizations and advancements for the Network Edge of 5G New Radio, funded under H2020-ICT-2016-2, project number: 760809.
  - MONROE: Measuring Mobile Broadband Networks in Europe, funded under: H2020-ICT-11-2014, project number: 644399.

- National projects:
  - 8.06/5.59.3722, contract with Optimi-Ericsson, with support from the Junta de Andalucía (Agencia IDEA, Consejería de Ciencia, Innovación y Empresa) and ERFD.
  - TEC2015-69982-R, Métodos de planificación y optimización de la calidad de experiencia en redes B4G, funded by the Spanish Ministry of Economy and Competitiveness.

6.3.5 Stays

David Palacios carried out a stay as a visiting researcher in the Wireless Communication Networks section in the University of Aalborg and Nokia-Bell Labs (both in Aalborg, Denmark) between July and October 2016 to study wireless heterogeneous networks with mixed-criticality traffic under the supervision of Beatriz Soret and István Kovács. The work in Section 5.2 mainly gathers the results of this stay.

Appendix A

Summary (Spanish)

A.1 Introducción

A.1.1 Antecedentes y justificación

En los últimos años, las comunicaciones móviles, que han evolucionado a través de cinco generaciones en apenas cuatro décadas, han atraído enormemente la atención tanto de la industria como de la comunidad investigadora. El motivo de este crecimiento vertiginoso reside en el hecho de que se han convertido en una parte esencial de la vida cotidiana actual.

Durante sus primeros veinte años, las comunicaciones móviles se orientaron hacia la transmisión de voz. A día de hoy, sin embargo, soportan una gran cantidad de servicios, entre los cuales cabe encontrar la navegación web, la difusión de vídeo o los videojuegos. Junto con estos servicios móviles de banda ancha (*mobile broadband*, MBB), el abanico de las comunicaciones móviles se ha abierto para cubrir otros escenarios, casos de uso y requisitos. Este el caso de las comunicaciones masivas tipo máquina (*massive machine-type communications*, mMTC), o las comunicaciones ultra fiables de baja latencia (*ultra-reliable low-latency communications*, URLLC) [1]. Como resultado, actualmente, el 95% de la población mundial vive en un área cubierta por una red móvil [2].

La primera generación de las comunicaciones móviles se desarrolló de forma local en diferentes regiones, lo que dio lugar a una variedad de implementaciones incompatibles entre sí. A su vez, esta tecnología tenía un carácter puramente analógico, lo que originó una falta de privacidad, fiabilidad y eficiencia en la comunicación. Con la segunda generación de las comunicaciones móviles llegó la digitalización y la estandarización a lo largo de varios países. Lo primero introdujo la fiabilidad y la eficiencia en las redes móviles; lo segundo, una red inalámbrica ubicua, permitiendo una comunicación ininterrumpida a lo largo de los países participantes en el estándar. La red móvil de segunda generación más extendida es el *Global System for Mobile communications* (GSM), que fue estandarizado por el *European Telecommunication Standards Institute* (ETSI),

y que permitió la trasmisión de voz y datos a través de una red celular basada en la conmutación de circuitos. Más adelante, el auge de las redes cableadas de conmutación de paquetes (como Internet) llevó al desarrollo de su contrapartida inalámbrica: una red celular doble, capaz de proporcionar tanto un servicio orientado a la conmutación de circuitos (como lo era la voz), como un servicio orientado a la transmisión de datos mediante la conmutación de paquetes. Es decir, una red que incluía tanto el estándar GSM, principalmente para soportar los servicios de voz, como el *General Packet Radio Service* (GPRS), para soportar los servicios de transmisión de datos. Algunos años más tarde, las demandas de los usuarios de nuevos servicios y mayor rendimiento de la red dieron lugar al desarrollo de la tercera generación (3G) de las comunicaciones móviles. 3G ofrecía una mayor capacidad que la generación anterior, principalmente debido a las mejoras realizadas en el interfaz radio, como el acceso múltiple por división en código (*code division multiple access*, CDMA). La principal tecnología de tercera generación para las comunicaciones móviles fue el *Universal Mobile Telecommunications System* (UMTS), que fue estandarizado por el *Third Generation Partnership Project* (3GPP). Impulsada por el crecimiento del tráfico basado en el protocolo de Internet (*Internet protocol*, IP), se desarrolló una red celular de conmutación de paquetes completamente basada en este este protocolo: la red *Long-Term Evolution* (LTE) [3]. Con LTE, la transmisión de paquetes IP no sólo se llevaba a cabo extremo a extremo (*end-to-end*, E2E), sino también entre los propios elementos de la red. Esto permitió simplificar enormemente la arquitectura de la red, a la vez que mejorar su rendimiento. Como consecuencia del surgimiento de nuevos tipos de comunicación (e.g., mMTC y URLLC) y del aumento de las demandas de rendimiento para los servicios tradicionales (como MBB, que ha dado lugar al servicio móvil mejorado de banda ancha, o *enhanced mobile broadband*, eMBB) se ha desarrollado recientemente el estándar para la quinta generación de las comunicaciones móviles *Fifth-Generation New Radio* (5G NR) [4]. La segunda fase del proceso de estandarización de esta tecnología celular acaba de comenzar y se espera que incluya el desarrollo de herramientas y mecanismos para la gestión y la optimización de las redes 5G NR definidas durante la primera fase.

El sistema móvil actual no está constituido por una única red celular aislado, sino por un gran conjunto de redes de diferentes generaciones. En concreto, a día de hoy coexisten redes de segunda, tercera y cuarta generación en despliegues comerciales. Las particularidades de cada una de estas tecnologías, junto con su influencia mutua, dificultan y encarecen las tareas de gestión, aumentando tanto el gasto en infraestructura como el gasto de operación de la red (*capital expenditure*, CAPEX, y *operational expenditure*, OPEX). Para evitar esto, la alianza *Next-Generation Mobile Networks* (NGMN) propuso el concepto de redes auto-organizadas (*self-organizing networks*, SON) en 2008 [5, 6], consistentes en mecanismos para la automatización de algunas de las tareas de gestión en las comunicaciones celulares. Poco tiempo después, el 3GPP incluyó el concepto de SON como un elemento clave para la gestión de las redes LTE [7], siendo aún más relevante para las futuras redes 5G NR [8].

Las funciones de SON se agrupan en tres categorías diferentes para la automatización de la gestión de la red: autoconfiguración [9], autooptimización [10] y autocuración [11]. La primera categoría hace referencia al grupo de funcionalidades dedicada a automatizar el despliegue de nuevas redes o nuevos elementos de red. Ejemplos de estas funciones serían mecanismos de *plug-and-play* o algoritmos de auto-planificación [12]. Por otro lado, las funciones de autooptimización

buscan maximizar el rendimiento de la red, que puede ser subóptimo debido a múltiples problemas variantes en el tiempo. Ejemplo de éstos serían problemas internos a la red, como un desbalance de carga, o problemas externos a la red, como un alto nivel de interferencia. In todos estos casos, es necesario reajustar los parámetros de configuración de la red para llevar a la red a un punto óptimo de funcionamiento. La cantidad de funcionalidades de la red de las cuales se puede extraer una métrica de rendimiento hace que aparezcan a su vez una gran variedad de casos de uso para la autooptimización. Algunos ejemplos de éstos serían la optimización de capacidad y cobertura (*capacity and coverage optimization*, CCO) o el balance de carga mediante movilidad (*mobility load balancing*, MLB) [14, 15]. Por último, el objetivo de las funciones de autocuración es evitar la degradación del rendimiento de la red, y por tanto, de la calidad de experiencia (*quality of experience*, QoE) percibida por los usuarios. Para ello, la autocuración utiliza cuatro funciones [16]: detección de fallos en la red [17–19], diagnosis de fallos (también conocida como análisis de la causa raíz, o *root cause analysis*, RCA) [20–34], compensación de fallos [35, 36] y recuperación tras un fallo.

La creciente complejidad de la gestión de las redes celulares y el intento de los operadores de redes móviles (*mobile network operators*, MNO) de reducir sus gastos mientras proporcionan una mejor QoE ha hecho que SON se convierta en un tema de investigación interesante. En los últimos diez años han tenido lugar múltiples proyectos de investigación en el ámbito de las SON. Algunos de estos son CELTIC Gandalf [37], E3 [38], SOCRATES [39], SELF-NET [40], UniverSelf [41], SEMAFOUR [42] y COMMUNE [43]. La mayoría de ellos se han centrado en la autoconfiguración y la autooptimización. A pesar de que la autocuración ha atraído menos atención en forma de proyectos internacionales, algunos proyectos nacionales, en los que la autocuración desempeña una labor fundamental, han proporcionado una amplia variedad de resultados de investigación [19, 24–36].

Al contrario que las funciones de autoconfiguración, que tienen lugar durante la fase de despliegue de la red, las funciones de autooptimización y autocuración se ejecutan durante su fase de operación; es decir, durante la mayor parte del tiempo de vida de la red celular. Por este motivo, ambos grupos de funciones ocupan un lugar particularmente relevante dentro de las SON. Esto, junto con la complejidad de las redes celulares actuales y venideras, dadas sus novedosas funcionalidades, casos de uso de uso y categorías de servicio, hace que los esquemas actuales de autooptimización y autocuración sean insuficientes, necesitando un impulso en su desarrollo. Respecto a la autocuración, por ejemplo, trabajos recientes sobre RCA (como [24–34]) no contemplan la combinación de múltiples modelos de diagnosis provenientes de diferentes fuentes (como diferentes expertos en la solución de fallos en la red), lo que llevaría a una mejora notable en la precisión de la diagnosis. Respecto a la autooptimización, nuevas funcionalidades como los esquemas basados en multiconectividad (*multi-connectivity*, MC) previstos para 5G NR hacen que los mecanismos basados en la gestión de un único enlace (como [13–15, 44, 45]) necesiten ser revisados para sacar el máximo partido de las nuevas posibilidades de optimización.

Por otro lado, para llegar a conocer el estado de la red (si éste es subóptimo o si ha llegado incluso a degradarse), las funciones de autooptimización y autocuración utilizan indicadores de rendimiento, que en el ámbito de la gestión de redes celulares se llaman indicadores clave de rendimiento (*key performance indicators*, KPI). Estos indicadores cuantifican el rendimiento de

los procesos y funcionalidades de la red, siendo monitorizados y almacenados mediante el sistema de operaciones y soporte (*operations and support system*, OSS) de la red celular. En consecuencia, y con el objetivo de desempeñar las tareas de gestión de la forma más eficiente, los MNOs y los proveedores de equipos de comunicaciones móviles han hecho grandes esfuerzos por definir una cantidad y variedad suficientemente amplia de KPIs. Sin embargo, una funcionalidad o estado de la red frecuentemente sólo se ve representado por una pequeña cantidad de KPIs, siendo todos los demás una fuente innecesaria de información, y en último término, de ruido. La gran cantidad de KPIs normalmente monitorizados en las redes, así como el gran número de elementos de red bajo evaluación, no sólo conlleva un problema de almacenamiento en las bases de datos del OSS, sino que a menudo contribuyen al sobreajuste de los algoritmos que implementan las funciones SON. Es por esto que para desarrollar un sistema SON eficiente, es necesario realizar en primera instancia una selección de los KPIs a utilizar. Tradicionalmente, esta tarea ha sido llevada a cabo por expertos en la solución de fallos en la red [16–34]. Sin embargo, debido a la gran cantidad de tiempo necesaria para evaluar la relevancia de cada KPI en la red, así como las posibles interrelaciones entre éstos, cada experto en la solución de fallos ha tendido a usar de forma continuada un mismo conjunto de KPIs; el conjunto de KPIs que, según él, mejor permite evaluar estado subyacente de la red. No obstante, en la práctica, el número y variedad de KPIs seleccionados de esta forma a menudo difiere de aquellos que llevarían a un rendimiento y tiempo de procesado óptimo de los algoritmos SON.

A su vez, y motivado por el avance en la capacidad de cálculo de los computadores actuales, el aprendizaje automático (*machine learning*, ML), cuyas principales aplicaciones son la clasificación y predicción de patrones, ha experimentado un progreso sin precedentes en los últimos años. A día de hoy, el ML se ha extendido a una gran variedad de áreas, como la visión por ordenador [46,47], el análisis econométrico [48], o los procesos de fabricación [49]. En este punto, la gestión de las redes celulares no es una excepción, habiéndose dado ya los primeros pasos hacia la integración de técnicas de ML de alto rendimiento en este área. Este es el caso de [24], en el que se entrena una red neuronal avanzada para agrupar medidas provenientes de la red para tareas de diagnosis, o [34], en el que se construye un árbol de decisión sobre métricas de rendimiento reportadas por los usuarios para identificar celdas con algún problema. No obstante, la gestión automática de la red aún puede beneficiarse notablemente de los avances más recientes en ML. Es por ello, que esta tesis tiene como objetivo dar un paso adelante hacia la completa automatización y optimización de las funcionalidades SON mediante el desarrollo y la integración de nuevas técnicas de ML a las SON.

A.1.2 Objetivos

El principal objetivo de esta tesis es la mejora de la gestión de las redes celulares a través del desarrollo e integración de nuevas herramientas de ML. Para ello, esta tesis se centra en mejorar los dos grupos de funciones más relevantes dentro de SON: la autocuración y la autooptimización.

La Figura 1.1 muestra las tareas de autocuración y autooptimización junto con el OSS, encargado de monitorizar de forma continua la red celular y de almacenar las observaciones de la red por medio de KPIs. Estas observaciones pueden estar acompañadas de una etiqueta, que representa el estado de la red bajo el cual se tomó la medida, o carecer de ella, en cuyo

caso sólo se almacenaría la información relativa a los indicadores de rendimiento monitorizados. Estas bases de datos constituyen la base del conocimiento para las tareas de gestión automática, que pueden ser supervisadas o no supervisadas, dependiendo de si usan o no datos etiquetados. Esta tesis se centra en la mejora del sistema mostrado en la Figura 1.1 mediante el desarrollo y la aplicación de herramientas de ML en tres puntos: las tareas de autocuración, las tareas de autooptimización y la forma en la que la información de rendimiento se monitoriza y procesa en el OSS, para su posterior uso en estas funciones SON.

En particular, las líneas de investigación de esta tesis se pueden resumir en los siguientes objetivos (ver Figura 1.1):

- **Objetivo 1:** *Diseño de una herramienta para la mejora de la diagnosis automática.* En la práctica, seleccionar la técnica de diagnosis automática a usar en una red real supone una tarea compleja, dadas las ventajas e inconvenientes que unas y otras plantean. Además, una vez se ha decidido la técnica a utilizar, es necesario construir un modelo de diagnosis, bien a partir de la conocimiento de los expertos en resolución de problemas o a partir de bases de datos con históricos de casos. A menudo, hay varios expertos en la resolución de problemas en la red y varias bases de datos de casos, lo que lleva a construir diferentes modelos de diagnosis con distintas precisiones en la diagnosis.

  En lugar de seleccionar una técnica dada y luego especificar un modelo, este objetivo persigue desarrollar un marco de trabajo para combinar múltiples técnicas y/o modelos, con el objetivo de superar las limitaciones de las técnicas/modelos individuales y, por lo tanto, incrementar la precisión de la diagnosis. En el dominio de las redes celulares, las combinaciones híbridas de sistemas de diagnosis (i.e., la combinación de sistemas independientes de diferente naturaleza) no han sido estudiadas aún. El enfoque adoptado consiste en formular el problema de la diagnosis desde una perspectiva más general, asumiendo que se trata de un problema de clasificación de un conjunto de casos que muestran algún tipo de patrón en un número no siempre conocido de clases o estados de la red.

- **Objetivo 2:** *Desarrollo e integración de técnicas de reducción de dimensionalidad para tareas de autocuración.* El segundo objetivo de esta tesis es desarrollar una variedad de herramientas que proporcionen un conjunto reducido de KPIs que permita reducir las necesidades de monitorización y almacenamiento, a la vez que se mejora el rendimiento de las funciones SON sin la intervención humana. Para ello, se diseñarán y evaluarán diferentes técnicas de reducción de dimensionalidad para su aplicación en el campo de las tareas de autocuración.

- **Objetivo 3:** *Desarrollo de algoritmos para la mejora del rendimiento del tráfico eMBB y de baja latencia en un entorno vehicular en 5G por medio de la gestión de múltiples enlaces.* A día de hoy, las comunicaciones celulares se enfrentan a una etapa de crecimiento vertiginoso, en un intento de de abordar de forma conjunta un conjunto de servicios, casos de uso y requisitos que, hasta ahora, sólo podían ser cubiertos por una variedad de tecnologías inalámbricas [50]. El uso de diferentes y posiblemente simultáneas conexiones entre el equipo de usuario (*user equipment*, UE) y uno o más nodos de la red surge como una de las posibles soluciones para abordar esta disparidad de requisitos y servicios. El tercer objetivo

de esta tesis es sacar partido de este hecho y mejorar alguna de las métricas de rendimiento asociadas a eMBB y al tráfico de baja latencia en 5G. En particular, este objetivo está dividido en dos líneas de investigación, una para cada una de estos tipos de tráfico:

- **Objetivo 3.1**. Desarrollar un algoritmo para gestionar la asignación de portadoras (*component carriers*, CC) proporcionadas por varios nodos 5G, usando el concepto de MC. Centrado en el ámbito de eMBB, el objetivo es incrementar el *throughput* del UE haciendo una asignación adecuada de CCs.

- **Objetivo 3.2**. Desarrollar un algoritmo para satisfacer el tráfico de baja latencia en un entorno de comunicaciones *vehículo a todo* (*vehicle-to-everything*, V2X). En este ámbito, esta línea de investigación persigue el desarrollo de un mecanismo para seleccionar de forma dinámica el interfaz a utilizar por los UEs para enviar mensajes de baja latencia, basado en la evaluación de información de rendimiento obtenida por cada uno de ellos.

A.2 Combinación de múltiples sistemas de diagnosis para RCA mejorado

En el capítulo 3 se propone un método para la combinación de múltiples sistemas de diagnosis automática, con el objetivo de desarrollar una herramienta compuesta para RCA con una precisión en la diagnosis superior a la de sus componentes. Para ello, en primer lugar, se formula el problema de la diagnosis en redes celulares bajo una perspectiva más amplia, la de la clasificación de patrones en un entorno de aprendizaje automático. Después, mediante un mecanismo de dos fases se lleva a cabo la combinación de los sistemas de diagnosis base. Mediante la primera fase, o fase de construcción, se crea un modelo estadístico del comportamiento de cada uno de los sistemas de diagnosis individuales, cuantificando, para cada uno de ellos, la probabilidad de diagnosis de un estado de la red, condicionado a la observación de varios KPIs. Para ello, cada sistema de diagnosis base opera de forma individual; después, los diagnósticos resultantes se usan para la estimación de las PDFs de estos diagnósticos en base a los KPIs a partir de los cuales han sido inferidos. En la segunda fase, o fase de combinación, se integran los modelos estadísticos de los múltiples sistemas de diagnosis mediante una combinación bayesiana, usando diferentes aproximaciones algebraicas para ello. La generalidad del mecanismo de combinación propuesto permite la integración de sistemas de diagnosis de diferente naturaleza, así como la combinación de múltiples modelos de diagnosis, dada una misma técnica de diagnosis de partida.

En este capítulo se evalúan las prestaciones del método propuesto utilizando tanto datos recabados de una herramienta de simulación de una red de acceso LTE, como casos recogidos de una red LTE real.

A.3 Reducción de dimensionalidad aplicada a funciones de autocuración

En el capítulo 4 se estudian los beneficios de aplicar técnicas de reducción de dimensionalidad previos al uso de RCA dentro de las SON. Estos beneficios serían la reducción de las necesidades de almacenamiento en las bases de datos de la red (disminución del CAPEX) y la disminución de la complejidad de los modelos de diagnosis resultantes, así como del tiempo empleado por los expertos en resolución de problemas en la red en la selección manual de estos indicadores (disminución en el OPEX).

En este capítulo, en primer lugar, se propone una técnica no supervisada de selección de KPIs, con el objetivo de encontrar los indicadores de rendimiento que mejor permiten identificar varios estados de la red dado un conjunto de observaciones no etiquetados de la misma, siendo éste el caso más común en la redes comerciales actuales. A continuación, con el objetivo de obtener una mejor precisión en la diagnosis que la obtenida usando unos KPIs seleccionados de forma no supervisada, se propone una técnica supervisada para su selección. Esta técnica utiliza información acerca del estado de la red bajo el que las muestras fueron tomadas, en forma de etiqueta. Los resultados utilizando datos recabados a partir de redes reales muestran la capacidad de las técnicas propuestas para reducir en más de un 90% las necesidades de almacenamiento de la red y simultáneamente reducir la tasa de error de diagnosis (*diagnosis error rate*, DER) en aproximadamente un 50%. Asimismo, estas técnicas muestran su capacidad a la hora de relevar al experto en resolución de problemas en la red en la selección de los KPIs más relevantes, al eliminar el error derivado del sesgo humano.

Más adelante, se estudia el uso de varias familias de técnicas de extracción de atributos como método para la generación de un conjunto reducido de KPIs sintéticos con alta carga de información útil. Los resultados muestran la efectividad de la propuesta; especialmente, mediante el uso de técnicas lineales de extracción de atributos.

Por último, mediante el uso conjunto de técnicas de selección y extracción de atributos, se propone un marco de trabajo generalizado para la reducción de dimensionalidad para RCA. Este marco, además de permitir una compresión aún mayor de la información útil de rendimiento de la red en un número más reducido de indicadores que usando únicamente técnicas de selección o extracción, facilita la integración de diferentes fuentes de información de rendimiento para RCA, como KPIs recogidos por los elementos de la red de acceso, información recogida por los propios UEs o información de contexto. Los resultados muestran la capacidad del marco de trabajo propuesto de reducir la DER hasta un 0% usando datos provenientes de una red real, así como los beneficios de integrar múltiples fuentes de información de rendimiento.

A.4 Autooptimización para 5G NR

El capítulo 5 de esta tesis aborda la autooptimización en redes móviles de quinta generación. En particular, en este capítulo se estudia la gestión de múltiples enlaces para la mejora del *throughput* para servicios eMBB en un entorno con MC y la disminución de la latencia en comunicaciones

vehiculares. En ambos casos, las soluciones propuestas se aplican a nivel de usuario, permitiendo así una optimización del rendimiento extremo a extremo (*end-to-end*, E2E). En el primero de los casos, para maximizar el *throughput* de un UE, se propone un algoritmo para identificar qué nodos de la red de acceso pasan a ser nodos secundarios (*secundary node*, SN) y qué CCs de éstos y del nodo principal (*master node*, MN) se asignan a este UE. Para ello, para cada CC recibida por un UE se evalúa la calidad de la señal recibida y su nivel de carga, así como la carga de los nodos de red implicados en su despliegue. Los resultados obtenidos a partir de una simulación de una distribución no homogénea de carga en el escenario muestran la capacidad del método propuesto de aumentar el *throughput* de pico hasta un 50% en el caso de usar 5 CCs respecto a una asociación UE-CC únicamente basada en la potencia recibida, así como una disminución en la carga de los enlaces entre nodos de un 60% respecto al caso peor.

En el caso de la reducción de latencia en un entorno vehicular, y dados dos posibles interfaces para el envío de mensajes con diferentes niveles de criticidad, se ha propuesto un algoritmo para la selección dinámica del interfaz con menos latencia, usando para ello un árbol de decisión junto la definición de funciones de coste para cada interfaz. Los resultados de la simulación del escenario y algoritmo propuestos muestran la capacidad del último de reducir la latencia de los mensajes etiquetados como críticos, provocando un impacto despreciable en la latencia de los etiquetados como no críticos, respecto al uso determinista de cualquiera de los dos interfaces implicados.

A.5 Conclusiones

El capítulo 6 resume las principales contribuciones de la tesis y las posibles líneas futuras. Además, incluye una lista de las publicaciones relacionadas.

A.5.1 Contribuciones

Esta tesis se centra en la mejora de las tareas de autocuración y autooptimización dentro de las SON, siguiendo un enfoque de ML. En particular, en cuanto a la autocuración, se incluyen nuevas propuestas para un RCA más preciso, mejorando tanto las técnicas de diagnosis en sí, como la forma en la que se procesa la información de rendimiento utilizada por éstas. Respecto a la autooptimización, se proponen mecanismos para la mejora del rendimiento de la red haciendo uso de una gestión multienlace. Las contribuciones de esta tesis se detallan en las siguientes líneas:

a) Autocuración:

- **Desarrollo de nuevas técnicas de diagnosis para un RCA mejorado**. La literatura existente sobre RCA en redes de comunicaciones móviles propone algunos sistemas de diagnosis automática basados en diferentes técnicas de inteligencia artificial, donde cada uno tiene sus pros y sus contras. Cuando un operador decide incluir una herramienta para RCA en su red, la primera decisión es qué técnica de diagnosis automática usar. Una vez que esta técnica ha sido seleccionada, se ha

de elegir un modelo de diagnosis. En lugar de escoger una única técnica y modelo, en esta tesis se ha propuesto un método para combinar cualquier número y tipo de diferentes sistemas de diagnosis independientes, de forma que el sistema de diagnosis compuesto resultante proporcione un mejor rendimiento que los sistemas de diagnosis componentes. En particular, el método propuesto permite tanto combinar sistemas de diagnosis basados en diferentes técnicas de inteligencia artificial, como combinar diferentes modelos de diagnosis (i.e., modelos basados en la misma técnica, pero cuyos parámetros son diferentes, al haber sido construidos por diferentes expertos u obtenidos a partir de diferentes conjuntos de datos de entrenamiento). Para ello, se lleva a cabo una combinación algebraica del modelo de comportamiento estadístico de los sistemas de diagnosis independientes. Los resultados muestran que el rendimiento del método propuesto es significativamente mejor en términos de DER comparado con sus componentes independientes, usando tanto casos provenientes de simulación como casos provenientes de una red LTE real, siempre que la precisión de los sistemas de diagnosis independientes sea similar y que sus errores de diagnosis sean incorrelados. Además, el método propuesto se basa en conceptos que no están vinculados con una tecnología de comunicación móvil particular, sino en teoría de clasificación de patrones, permitiendo que sea aplicado bien en redes celulares maduras y asentadas, como LTE, o en tecnologías venideras, como 5G NR.

- **Uso de técnicas de reducción de dimensionalidad para la completa automatización de RCA y la reducción del OPEX y del CAPEX**. Las SON tienen como objetivo la automatización de las tareas de gestión de las redes celulares. Sin embargo, algunas tareas como la selección de los KPIs más apropiados para RCA aún las llevan a cabo expertos, lo que supone una de sus tareas más costosas en tiempo, además de constituir una causa para el rendimiento subóptimo de las funciones de RCA debido al sesgo humano. En esta tesis, se han estudiado y usado técnicas de reducción de dimensionalidad, habilitando la completa automatización de RCA en el ámbito de la autocuración. En particular, las principales contribuciones en este tema son las siguientes:

  - El desarrollo de una técnica *no supervisada* para la selección de los KPIs más útiles para RCA, consistente en una etapa de agrupación de datos seguida de un procedimiento supervisado para la selección supervisada de atributos. Su naturaleza no supervisada permite al método determinar los KPIs más apropiados para discernir entre un conjunto de estados de red no conocidos a priori. Como resultado, su principal ventaja es que permite usar bases de datos de datos no etiquetados para la selección de KPIs, siendo ésta la situación más común en redes móviles recientemente desplegadas. Los resultados muestran que la selección llevada a cabo por el método propuesto supera la selección realizada tanto por un experto en resolución de problemas en redes móviles como las realizadas por varias técnicas de selección de atributos no supervisadas del estado del arte, permitiendo una reducción drástica del volumen y complejidad de las bases de datos de la red y de las técnicas de RCA respectivamente, sin intervención humana.

– El desarrollo de una técnica *supervisada* para la selección de KPIs para RCA. A pesar de que las técnicas supervisadas necesitan información adicional para operar en forma de etiquetas, proporcionan un mejor rendimiento que las técnicas no supervisadas para la selección de atributos en el campo del RCA. Esto las hace especialmente adecuadas para su uso redes celulares estables, que suelen disponer de bases de datos históricas de muestras de red pasadas. La técnica supervisada propuesta para la selección de KPIs usa la diferencia estadística de un KPI bajo la presencia de diferentes estados de red como figura de mérito, medida como la región sin solape de las funciones de densidad de probabilidad (*probability density function*, PDF) de este KPI cuando está condicionado de forma independiente a estos estados de la red. Es decir, cuanto mayor sea la diferencia del comportamiento estadístico de un KPI dados dos estados subyacentes de la red, mayor es la utilidad de este KPI para distinguir entre estos estados. Además, para modelar de forma precisa el comportamiento estadístico de un KPI, se utilizan técnicas no paramétricas para la estimación de su PDF. Los resultados muestran que el método propuesto mejora la selección llevaba a cabo por otras técnicas supervisadas y no supervisadas del estado de arte, así como la selección de un experto en la resolución de problemas en la red, mediante el uso de datos provenientes de una red real, demostrando su validez para automatizar por completo las tareas de diagnosis cuando es usado de forma conjunta con una herramienta para RCA.

– La evaluación de diferentes técnicas para la extracción de atributos, con el objetivo de ser integradas como una etapa intermedia entre la monitorización de los KPIs de la red y su uso como entrada de herramientas de RCA. A costa de perder el significado de los KPIs resultantes, las técnicas de extracción de atributos permiten condensador la información de rendimiento relevante en un conjunto más reducido de KPIs sintéticos que las técnicas de selección de atributos. Los resultados, obtenidos sobre un conjunto de datos recogidos de una red celular real, muestran los beneficios de este enfoque en términos de ahorro en almacenamiento y en una posterior mejora de la función de RCA usada. Los beneficios han mostrado ser especialmente relevantes con el uso de técnicas lineales de extracción de atributos, dada la dependencia principalmente lineal de los KPIs monitorizados.

– Aprovechando el estudio sobre las técnicas de reducción de dimensionalidad llevado a cabo hasta este punto en la tesis, se ha propuesto un marco de trabajo para las SON de nueva generación (*next-generation self-organizing networks*, NG-SON) en el ámbito de la autocuración, usando tanto técnicas de selección de atributos como técnicas de extracción de atributos para mejorar la precisión del RCA. Este marco de trabajo permite utilizar tanto datos etiquetados como no etiquetados, así como integrar la información de rendimiento de diferentes fuentes, como medidas recogidas por los UEs, KPIs de la red o información de contexto. Este marco de trabajo ayuda e incluso libera a los expertos en resolución de problemas en la red de tener que tratar con tal cantidad de datos diferentes (disminuyendo el OPEX) y permite un reducción las necesidades de almacenamiento de la red,

así como la complejidad final de los mecanismos de autocuración (disminuyendo el CAPEX). Los resultados muestran que el método propuesto puede gestionar de forma efectiva un entorno con un gran número de dimensiones, provenientes de diferentes fuentes de información, automatizando en último término las tareas normalmente realizadas por expertos en la resolución de problemas de la red a la vez que optimizando el rendimiento de la herramientas de RCA.

- **Definición de métricas de rendimiento basadas en trazas de llamadas inter-tecnología para una detección y RCA mejoradas**. A pesar de que tradicionalmente, la fuente de información de rendimiento más ampliamente usada para la autocuración es la proveniente de los KPIs de la red, las trazas de llamadas (información de señalización, recogida de estas llamadas) proporciona una fuente de informacion mucho más detallada. Si bien esta información no siempre está disponible, a menudo proporciona información de rendimiento específica de cada usuario, así como marcas temporales y datos de ubicación, lo que permite emplear técnicas de detección y de RCA mucho más precisas, así como llevar a acabo acciones más específicas para resolver posibles problemas en la red. En esta tesis se define un método para evaluar el rendimiento de un mecanismo inter-tecnología, el *circuit-switched fallback* (CSFB), utilizando para ello las trazas de llamada. Con este objetivo, se analizan las bases de datos de trazas de llamadas de las tecnologías radio de origen y destino, llevando a cabo un estudio E2E, a diferencia de los métodos del estado del arte, que sólo llevan a cabo una evaluación en uno de los extremos. Los resultados, tras utilizar datos de trazas de llamadas de redes LTE y UMTS, muestran este método como una herramienta valiosa para detectar situaciones problemáticas que permanecerían ocultas al ser evaluadas mediante métodos del estado del arte.

b) Autooptimización:

- **Gestión de CCs en un entorno con MC para optimizar el rendimiento de la red**. Para abordar la disparidad de requisitos entre las diferentes categorías de servicio abordadas por 5G NR, en la actualidad se proponen varias soluciones en el ámbito de gestión de recursos radio (*radio resource management*, RRM), entre las cuales se encuentra la MC. La MC ofrece varias ventajas: desde una fiabilidad mejorada debido a la diversidad de los múltiples enlaces, a mejoras de *throughput* debido al uso adicional de recursos radio a partir de las estaciones base (*base station*, BS) implicadas. Con la MC, sin embargo, surgen dos retos: la selección de las BSs con las que el UE debería asociarse y la selección de las CC de estas BSs a asignar al UE. Mientras que los trabajos del estado del arte abordan estas dos cuestiones de forma separada, en este trabajo, estos retos se llevan a cabo de forma conjunta, lo que conlleva una mejora en el rendimiento del UE y una mejora en la utilización de los recursos de la red. Para ello se ha propuesto el uso de un mecanismo RRM basado en un controlador de lógica difusa (*fuzzy-logic controller*, FLC). La solución propuesta, que permite la definición de políticas del operador sobre la red debido a su naturaleza basada en reglas, acepta una variedad de fuentes de información como el contexto de usuario. A su vez, esto permite desarrollar un enfoque centrado en el usuario, en línea con la

tendencia actual en las tareas de gestión de 5G. Las simulaciones en un entorno eMBB han mostrado que, en un escenario con carga desbalanceada, el mecanismo propuesto permite incrementar el *throughput* de usuario a la vez que se mejora la eficiencia de la red mediante un balance de carga inter-CC e inter-BS, aplicado en un entorno de MC, cuando se compara con los esquemas de asociación de BS y asignación de CC basados en la potencia recibida.

- **Redirección de tráfico para comunicaciones V2X de baja latencia**. Se prevé que las futuras redes celulares den servicio no sólo a comunicaciones de gran ancho de banda, sino también a comunicaciones de muy baja latencia, y entre éstas, las comunicaciones V2X ocupan un lugar especialmente relevante. Este tipo de comunicaciones se caracterizan por el intercambio tanto de mensajes de baja latencia, orientados a mensajes de control y notificaciones de alta prioridad, como a mensajes tolerantes al retardo, dedicados a servicios de entretenimiento e información general. En esta tesis, se ha propuesto un mecanismo para la redirección dinámica de tráfico sobre los interfaces Uu y PC5 (bajo el estándar del 3GPP) de las comunicaciones V2X, para asegurar una baja latencia para los mensajes críticos, incluso bajo condiciones de alta carga. Como prueba de concepto, se ha llevado a cabo una simulación usando macroceldas LTE y puntos de acceso IEEE (*Institute of Electrical and Electronics Engineers*) 802.11n para proporcionar los interfaces Uu y una aproximación al PC5, respectivamente. La solución propuesta ha demostrado redirigir de forma efectiva los paquetes de datos, atendiendo a su nivel de criticidad, proporcionando un enlace con mínimo retardo para mensajes críticos y un comportamiento *best-effort* para los mensajes tolerantes al retardo (mensajes no críticos).

Además, y como parte de su trabajo de investigación, David Palacios ha coordinado el despliegue de una red LTE de investigación, en forma de banco de pruebas, en la Universidad de Málaga, teniendo ésta un papel fundamental en dos de los proyectos internacionales en los cuales la Universidad de Málaga está implicada. Se puede encontrar una descripción de esta red en el Apéndice C.

A.5.2 Publicaciones y proyectos

Las siguientes secciones presentan las publicaciones relacionadas con esta tesis.

Revistas

Publicaciones derivadas de esta tesis

[I] D. Palacios, E. J. Khatib y R. Barco. "Combination of Multiple Diagnosis Systems in Self-Healing Networks". *Expert Systems with Applications*, vol. 64, pp. 56-68, 2016.

[II] D. Palacios y R. Barco. "Unsupervised Technique for Automatic Selection of Performance Indicators in Self-Organizing Networks". *IEEE Communications Letters*, vol. 21, no. 10, pp. 2198-2201, Oct. 2017.

[III] D. Palacios, I. de-la-Bandera, A. Gómez-Andrades, L. Flores y R. Barco. "Automatic Feature Selection Technique for Next-Generation Self-Organizing Networks". *IEEE Communications Letters*, vol. 22, no. 6, pp. 1272-1275, June 2018.

[IV] D. Palacios, S. Fortes, I. de-la-Bandera y R. Barco. "Self-Healing Framework for Next-Generation Networks through Dimensionality Reduction". *IEEE Communications Magazine*, vol. 56, no. 7, pp. 170-176, July 2018.

[V] D. Palacios, I. de-la-Bandera y R. Barco. "Multi-Node Component Carrier Management for Multi-Connectivity in 5G New Radio". *IEEE Communications Magazine*. Bajo revisión.

Publicaciones relacionadas con esta tesis

[VI] N. Mahmood, D. Laselva, D. Palacios, E. Mustafa, F. Miltiadis, D. M. Kim y I. de-la-Bandera. "Multi-Channel Access Solutions for 5G New Radio". *IEEE Communications Magazine*, bajo revisión.

[VII] S. Fortes, D. Palacios, I. Serrano y R. Barco. "Applying Social Event Data for the Management of Cellular Networks". *IEEE Communications Magazine*, vol. 56, no. 11, pp. 36-43, Nov. 2018.

[VIII] J. Mendoza, D. Palacios, I. de-la-Bandera, A. Herrera y R. Barco. "On the capability of QoE optimization based on the adjustment of RLC parameters". *IEEE Communications Magazine*, bajo revisión.

Patentes

Patentes derivadas de esta tesis

[IX] D. Palacios, R. Barco, O. Kaddoura, P. Delgado e I. Serrano, "Inter-technology circuit-switched fallback (CSFB) metrics", WO 2017063700 A1, 2017.

Conferencias

Conferencias derivadas de esta tesis

[X] D. Palacios, E. J. Khatib, I. Z. Kovács, B. Soret, I. de-la-Bandera y R. Barco. "Dynamic Multipath Connection for Low-Latency Vehicle-to-Everything (V2X) Communications". *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Valencia (Spain), 2018, pp. 1-5.

[XI] D. Palacios, R. Barco e I. Serrano. "Combinación de Sistemas para la Diagnosis de Fallos en Self-Organizing Networks", *XXX Simposium nacional de la Unión Científica Internacional de Radio*, Pamplona (Spain), Sept. 2015.

[XII] D. Palacios, I. de-la-Bandera y R. Barco. "Reducción de Dimensionalidad en Funciones SON mediante Técnicas de Extracción de Atributos", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

[XIII] I. de-la-Bandera, D. Palacios y R. Barco, "Asignación Automática de Portadoras en un Escenario 5G", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

Conferencias relacionadas con esta tesis

[XIV] J. Burgueño, I. de-la-Bandera, D. Palacios y R. Barco, "Modelado de TCP en un Entorno Celular con Dual Connectivity", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

[XV] J. Mendoza, A. Herrera, D. Palacios, I. de-la-Bandera y R. Barco, "Optimización de la QoE de un Servicio de Video Streaming en un Entorno Celular", *XXXIII Simposium nacional de la Unión Científica Internacional de Radio*, Granada (Spain), Sept. 2018.

A.5.3 Proyectos relacionados

Esta tesis ha sido financiada bajo el proyecto P12-TIC-2905 "Gestión integral avanzada de funciones SON (Self-Organizing Networks) para redes móviles futuras", de la Junta de Andalucía.

Esta tesis también ha contribuido a los siguientes proyectos:

- Proyectos internacionales:
  - ONE5G: E2E-aware Optimizations and advancements for the Network Edge of 5G New Radio, financiada bajo H2020-ICT-2016-2, número de proyecto: 760809.
  - MONROE: Measuring Mobile Broadband Networks in Europe, financiada bajo: H2020-ICT-11-2014, número de proyecto: 644399.
- Proyectos nacionales:
  - 8.06/5.59.3722, contrato con Optimi-Ericsson, con la ayuda de la Junta de Andalucía (Agencia IDEA, Consejería de Ciencia, Innovación y Empresa) y el Fondo Europeo de Desarrollo Regional (FEDER).
  - TEC2015-69982-R, Métodos de planificación y optimización de la calidad de experiencia en redes B4G, financiado por el Ministerio de Economía y Competitividad.

A.5.4 Estancias

David Palacios ha llevado a cabo una estancia como investigador visitante en la sección de Redes de Comunicación Inalámbricas en la Universidad de Aalborg y Nokia-Bell Labs (ambos en Aalborg, Dinamarca) entre julio y octubre de 2016 para el estudio de redes inalámbricas heterogéneas con tráfico con diferentes de criticidad, bajo la supervisión de Beatriz Soret e István Kovács. El trabajo aquí desarrollado está recogido en la Sección 5.2.

Appendix B

Inter-technology circuit-switched fallback metrics

As an intermediate step between RCA and multi-connectivity studies in the self-healing and self-optimization fields, respectively, a multi-RAT self-healing research line has been addressed. In this case, the performance of the mechanism to redirect packet-switched (PS) calls, originated in an all-IP network, to the circuit-switched (CS) subnetwork of a legacy RAT has been studied. This mechanism is called CSFB.

In this appendix, a set of performance metrics and an accurate, scalable and computationally low cost call tracing method based on call trace data are defined to evaluate the E2E performance of the CSFB mechanism, which, prior to the full deployment of the voice over LTE (VoLTE) functionality, has been the procedure by which voice calls are performed in LTE networks.

This appendix is organized as follows. First, Section B.1 outlines the related work. Next, Section B.2 describes the problem formulation. Then, in Section B.3, the proposed method for call tracing and the definition of the performance metrics are described. In Section B.4, a proof of concept is carried out with data gathered from a live co-located LTE/UMTS pair of networks. And finally, Section B.5 summarizes the conclusions.

The content of this appendix gathers the patent application *Inter-technology circuit-switched fallback (CSFB) metrics*, with application number WO2017063700A1.

B.1 Related work

The fact of a network relying on another is not new, neither are the efforts on quantifying the effectiveness of this support. One example of this came with the inter-system or inter-RAT handovers between UMTS and GSM, studied during several years before LTE came in. In [125],

some procedures were proposed to identify 3G-2G IRAT HOs for both CS and PS services regarding the signaling contained in the call traces data.

Concerning the inter-system mobility from and to LTE, some other performance indicators have been described. In [126], several mobility-related KPIs were defined regarding the call trace data. In [127], a similar group of KPIs were defined taking event counters (described in [128]) from the network as the input for the calculations.

However, despite these procedures work inherently between two technologies, the data used to compute these indicators is gathered only from the source technology. In these cases, the reception in the source technology of a message of acknowledgment from the target technology upon the incoming call notification is enough to consider the technology hop is successful, at least in the side of the source technology. This is, the state-of-the-art performance indicators only ensure the call has been successfully released towards a legacy technology, not that the call has been successfully initiated in that technology. In order to be sure that the call performing a technology hop (such as CSFB) successfully establishes a voice path, it should be followed up in the target technology and its event flow should be checked.

B.2 Problem formulation

The fast growth and expansion of novel mobile technologies has led the communication networks to become even more complex compounds of multi-technology networks, where the newer ones overlap the already existing networks. In this context, big efforts have been made to build a more reliable network in which the legacy RATs, like GSM or UMTS support the yet immature new technologies during the development of all their features. Such is the case of voice traffic in LTE, which over several years has been mostly supported by the CS side of the legacy technologies, GSM and UMTS, by means of two mechanisms: CSFB, [129], and single-radio voice call continuity (SRVCC), [130]. The first procedure is used when the LTE user is in an area where the VoLTE service has not been deployed yet and a voice call is intended to be started. At that moment, during the call setup, this is redirected towards the CS subsystem of a legacy technology. By now, although VoLTE is gaining momentum worldwide, its deployment is limited to main cities. For these reasons, CSFB arises as a mid-term solution to hold voice traffic in the worldwide deployment of VoLTE.

B.3 Proposed method

The method and metrics herein are based on the analysis of call trace data. These data are first collected by the eNBs in LTE and by the radio network controllers (RNCs) or the base station controllers (BSCs) in the case of the legacy technologies. Then, they are collected in files in their respective OSSs and next, these data files are gathered, parsed and processed by an external tool, generating two call databases: one for the source technology (LTE), and one for the target (UMTS or GSM). Each database contains: 1), a set of calls with call related information (e.g., a call identifier, its starting and ending times and some information about the user's or equipment's identity) and 2), the list of signaling events exchanged between the UE and the rest of the RAN

in the control plane for each call. These databases act as the input for the proposed method, which has two outputs: the E2E CSFB metrics and a set of call identifiers taken pairwise. This is, the pair of call identifiers corresponding to the same call in its both sides, obtained once both sides have been matched.

The proposed method consists of three procedures. They are used to classify and match the calls from both databases regarding their features and the contents of their call events. The first of these algorithms is Algorithm B.1: LTE calls filter. It identifies the LTE calls with an available user or UE identifier (ID) that start within the registration area of the target legacy technology. These identifiers may be either the international mobile subscriber identity (IMSI) or the international mobile station equipment identity (IMEI) and are used afterwards for call matching purposes. The registration area of either the source or the target technology is the space region that comprises the cells whose call traces have been registered.

Algorithm B.1: LTE calls filter

for *LTE call* $C_{LTE} \in$ *LTE call database* **do**
  if {[(C_{LTE}.*Target cell ID is known*) ***and*** (C_{LTE}.*Target cell ID* $\in$ *UMTS/GSM registration area*) ***and*** (*neither* C_{LTE}.*IMSI nor* C_{LTE}.*IMEI are null*)] ***or*** [(C_{LTE}.*Originating cell* $\in$ *overlapping area*) ***and*** (*neither* C_{LTE}.*IMSI nor* C_{LTE}.*IMEI are null*)]} ***and*** {C_{LTE} *has a CSFB-triggering event*} **then**
    Add C_{LTE} to CSFB ATTEMPTS;
    if C_{LTE} *has a release-notifying event* **then**
      Add C_{LTE} to SUCCESSFUL RELEASE FROM LTE
    else
      Add C_{LTE} to FAILED CSFB IN LTE

During a CSFB preparation, the target ID is determined by the source eNB and is sent to the MME by means of a *Handover Required* message. In practice, however, this field or even the whole message itself often lack due to mobile operator data traces storing policies or just because of a data traces registration failure. In both cases, the result is the same, the lack of the target cell ID. Therefore, with the aim of maximizing the number of identified call attempts in the source technology having a “call end" message in the target technology, an area constraint must be set on the source technology registration area. This is shown in Figure B.1. In this figure, the LTE registration area is shown on the left; the UMTS/GSM registration area, on the right, and their intersection is shadowed in grey in the center. However, the call attempts that take place near the edge of the UMTS/GSM registration area may still be redirected outside this region, losing the trace of the call in the target side. To prevent this, the target area of interest is reduced by means of a factor k, resulting in a smaller LTE and UMTS/GSM overlapping area. As it can be seen, this edge effect is progressively removed as this factor increases.

In practice, these areas of interest have been defined as rectangle-shaped and are determined by the latitude and longitude of the farthest UMTS/GSM cells for which call trace data have been registered. Their bounds are defined as:

$$Lon|Lat_{min} = Lon|Lat_{min_t} + k \cdot (Lon|Lat_{max_t} - Lon|Lat_{min_t}) \tag{B.1a}$$

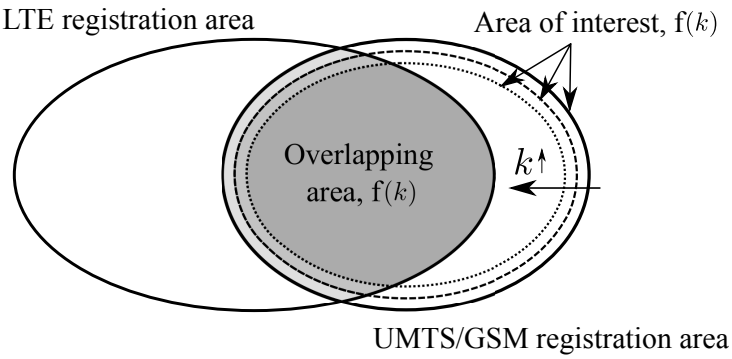


Figure B.1: Scheme of the source and target registration areas and their overlapping region, controlled by the parameter k.

$$Lon|Lat_{max} = Lon|Lat_{max_t} - k \cdot (Lon|Lat_{max_t} - Lon|Lat_{min_t}) \tag{B.1b}$$

where $Lon|Lat_{min|max}$ stands for the minimum or maximum longitude or latitude of the area of interest; $Lon|Lat_{min|max_t}$ stands for the minimum or maximum longitude or latitude of the cells in the target technology registration area and k is defined between 0 and 0.5. The overlapping areas in step 3 in Algorithm B.1 depend on this definition of the areas of interest.

Once the LTE call starts have been filtered by their location, two more checks are performed. First, Algorithm B.1 looks for a CSFB-triggering event in the call trace data. If it finds it, the call is added to the set CSFB ATTEMPTS. Then, Algorithm B.1 looks for an event that notifies a successful release from LTE among the calls in CSFB ATTEMPTS. If it finds it, the call is added to the set SUCCESSFUL RELEASE FROM LTE; if it does not, the call is added to the set FAILED CSFB IN LTE.

At this point, both sides of each call can be matched. This is done by Algorithm B.2. This algorithm takes every call from the UMTS/GSM call database and from SUCCESSFUL RELEASE FROM LTE and checks whether any of their user or equipment identifiers (IMSI and IMEI, respectively) matches within a temporal window. This time check evaluates whether the call in the target technology started between the starting and ending time of the call in the source technology. The starting and ending times of the call are defined as the timestamps for its first and last registered event. In this time window, a threshold Δ is added to the end of the call in the source technology, so that the call registration delays in the target technology can be taken into account. If these checks are fulfilled, the UMTS/GSM call is added to the set MATCHED CALLS; it they are not, the UMTS/GSM call is added to UNMATCHED CALLS.

The next step is to identify the CSFB failures and successes in the UMTS/GSM environment. A CSFB is considered as E2E successful, and therefore, added to the set SUCCESSFUL CSFB, once the network setup for the voice link is established. This is done by Algorithm B.3, which will look for the CS call control events: alert, connect, call progress or disconnect (which are exchanged in both UMTS and GSM) sent or received by the UE, depending on the case [131]. The reasons why these messages have been chosen are:

- **Alert:** This message is sent by the network to the calling UE in the mobile-originated side and in the opposite direction in the mobile-terminated side. It is used to notify that the ongoing call has reached the stage where the destination UE begins to ring. At this point, the CN in the mobile-terminated side has accepted the incoming call and a voice E-RAB must have been set up.
- **Call progress**: This message is sent from the network to the UE to indicate the progress of the call when interworking with other networks. For example, when the call leaves the public land mobile network (PLMN)/integrated services digital network (ISDN) environment.
- **Connect**: This message may be either sent in the CN-UE direction or in the opposite, usually after an *alert* message, and indicates the call acceptance by the called user. However, this event may also take place without a previous *alert* message if the call is automatically accepted. This situation usually takes place when the user calls some service of automatic attendance.
- **Disconnect**: This message is taken into account in the CN-UE direction to consider the cases in which the called user does not accept the incoming call despite being successfully setup in both mobile-originated or mobile-terminated sides. To avoid considering disconnections related to the lack of network resources or protocol errors, only disconnections with normal causes must be considered. These are *user busy*, *no user responding*, etc. The disconnection causes can be normally found in the contents of the disconnect-notifying events.

Algorithm B.2: Call matching

for *UMTS/GSM call* $C_{3G/2G} \in$ *UMTS/GSM call database* **do**
  for *LTE call* $C_{LTE} \in$ *SUCCESSFUL RELEASE FROM LTE* **do**
    if $(C_{LTE}.IMSI = C_{3G/2G}.IMSI$ ***or*** $C_{LTE}.IMEI = C_{3G/2G}.IMEI)$ ***and*** $(C_{LTE}.Start \leq C_{3G/2G}.Start \leq C_{LTE}.End + \Delta)$ **then**
      Add $C_{3G/2G}$ to MATCHED CALLS
    else
      Add $C_{3G/2G}$ to UNMATCHED CALLS

Algorithm B.3: Call assessment

for *UMTS/GSM call* $C_{3G/2G} \in$ *MATCHED CALLS* **do**
  if *any of the events from* $C_{3G/2G}$ *contains an* {*alert, connect, call progress or disconnect*} *call control message* **then**
    Add $C_{3G/2G}$ to SUCCESSFUL CSFB
  else
    Add $C_{3G/2G}$ to FAILED CSFB IN TARGET

If none of these cases takes place, the call is labeled as FAILED CSFB IN TARGET. Once

the calls have been classified into sets, e.g. FAILED CSFB IN LTE, some performance metrics can be defined:

$$CFR_{LTE} = \frac{|FAILED\ CSFB\ IN\ LTE|}{|CSFB\ ATTEMPTS|} \tag{B.2}$$

$$CFR_t = \frac{|FAILED\ CSFB\ IN\ TARGET|}{|CSFB\ ATTEMPTS|} \tag{B.3}$$

$$MR = \frac{|MATCHED\ CALLS|}{|SUCC.\ RELEASE\ FROM\ LTE|} \tag{B.4}$$

$$CSR = \frac{|SUCCESSFUL\ CSFB|}{|CSFB\ ATTEMPTS|} \tag{B.5}$$

In equations (B.2) to (B.5), $|A|$ stands for the cardinality of the set A (i.e., the number of elements in set A). CFR_{LTE} and CFR_t represent the CSFB failure rate in the source and target technology, respectively; MR is the call matching rate and CSR denotes the CSFB success rate. Equation (B.6) can afterwards be derived from these equations and Algorithms B.1 to B.3.

$$CSR + CFR_t + CFR_{LTE} + (1 - MR)(1 - CFR_{LTE}) = 1 \tag{B.6}$$

B.4 Proof of concept

The proposed method has been tested in a live multi-RAT network, comprising two in time and space co-located LTE and UMTS networks. The considered LTE network is composed of 19879 cells, while the UMTS network, for which only a RNC has been considered, comprises 4515 cells. In this case, the LTE call database included no information about the target ID, neither for the regular IRAT HOs, nor for the CSFB. This emphasizes again the necessity of setting an inter-technology overlapping area to avoid considering those CSFB attempts that came from LTE cells which are far from the registered legacy sectors and which will contribute to increase the number of unmatched CSFB calls.

In this proof of concept, the signaling messages taken as the CSFB trigger and release-notifying events are *NAS Extended Service Request* and *S1AP UE Context Release Complete*, respectively. The first is sent in the UE-MME direction in both the mobile-originated and the mobile-terminated scenarios (in the latter, after the UE received a paging request and a CS service notification) and it is used to notify the MME its intention to initiate a CSFB. The *UE Context Release Complete* event, on the other hand, is sent by the eNB to the MME to confirm the release of the UE-associated S1-logical connection over the S1 interface.

Figure B.2 represents equation (B.6) and shows the results of applying the method described in Section B.3 and computing Eq. (B.2), (B.3) and (B.5). UR has been computed as $(1 - MR)(1 - CFR_{LTE})$, which is proportional to an unmatched call rate. The LTE-UMTS call databases collect three fifteen-minute periods of time: T1 to T3. In this proof of concept, k has

been set to 0.15, 0.3 and 0.4 and Δ has been set to 0 and 1.5 seconds. This figure also shows the size of the areas of interest for each value of k the number of CSFB attempts (#Att) for each T and k to give an idea on the considered areas and sizes of the sample, respectively. As it can be seen, as the parameter for edge cell effect removal, k, is increased for a fixed Δ, the rate of unmatched calls tends to decrease. This occurs at the expense of considering a lower number of CSFB attempts: only those call attempts within the ever-decreasing inter-technology overlapping area. Likewise, if the value of k is held in a time period, the number of matched calls is always higher for $\Delta = 1.5$ s than for $\Delta = 0$. It can be seen how the cases with the maximum k and Δ present a MR of 100%. It is in these cases where the E2E CSFB assessment can be fully performed. This is, the cases in which every CSFB attempt can be classified into successful and failed, and within the latter, in which technology the failure took place. In these cases, two bars have been plotted for each T: a striped "bs" bar for the baseline methods and a filled "pr" bar for the proposed method. As it can be seen, the "bs" and "pr" bars output the same results in T3, since the failures took place in the source technology. However, these bars output different results in T1 and T2. In both cases, there are failures in the target technology that went unnoticed from the point of view of the source technology. Thus, the methods in [126–128] would provide a false $CSR = 100\%$ in the cases T1 and T2 with $MR = 100\%$, whereas the proposed method would detect the failed CSFB. This fact highlights the importance of tracking the calls from one RAT to the other in an IRAT HO. Compared to the state-of-the-art, traditional methods for assessing IRAT performance would ignore failures in legacy technologies caused by, for example, a faulty location area update (LAU)/routing area update (RAU) if the mobile switching center (MSC) is changed or a failing CS call establishment procedure. Furthermore, by using the proposed approach, these in-target-technology failed calls could be afterwards troubleshot by inspecting their event flows in the corresponding database.

B.5 Conclusion

A set of metrics to assess the E2E CSFB performance has been presented and applied on a live co-located LTE/UMTS network. With that aim, a method for effectively tracking the calls from the source to the target technology based on the call traces analysis has also been proposed. These contributions allow the user not only to quantify the performance of the CSFB inter-technology transitions as a coarse first glance, but to in-depth troubleshoot the failed calls in either the source or the target technologies regarding the event flows from both sides of the call, and therefore, to take concise and appropriate actions to solve the arising failures.

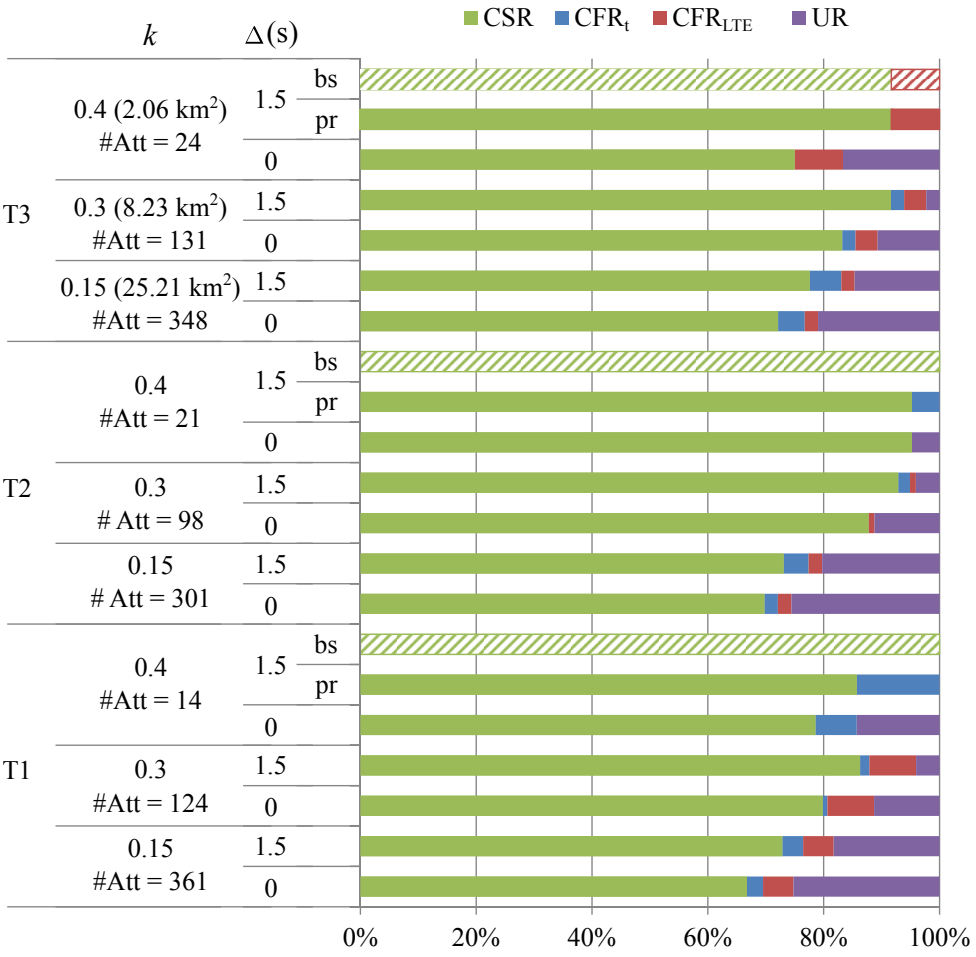


Figure B.2: Computed metrics: CSR, CFR$_t$, CFR$_{LTE}$ and UR.

Appendix C

Description and deployment of UMAHetNet

This appendix includes the description of the indoor LTE network deployed at UMA, UMAHetNet; a key element in two of the most recent international projects of UMA (namely, ONE5G and MONROE), and in whose deployment David Palacios acted as the coordinator.

This appendix is organized as follows. Section C.1 describes the motivation of UMA to add UMAHetNet as one of its main research tools. In Section C.2, the main components of this network are described. Section C.3 outlines the layout of the network along the facilities of UMA. And finally, Section C.4 summarizes the research results provided by UMAHetNet up to the present day.

C.1 Overall context

UMA, and in particular, the Communications Engineering Group ("Grupo Ingeniería de Comunicaciones", GIC, TIC-102) has a long expertise in the development of ML algorithms for SON. Examples of these may be found in [13–16, 19, 24–36]. These studies, however, have been traditionally tested by means of simulation tools [75, 120], or have been assessed in constrained trials in commercial networks. The main reason for this has been the usual reluctance of MNOs to modify their network configurations, specially in self-optimization tasks, which could temporarily degrade the network (and even more, the users' perceived) performance, thus eventually leading to customers' complaints.

Given this, a tool to extensively test and validate the results of this research in a realistic environment appeared as a pressing need. As a result, research funds were recently destined to acquire a fully reconfigurable indoor cellular infrastructure: the UMAHetNet.

C.2 General scheme

The UMAHetNet is a full indoor LTE network, compliant with 3GPP Release 9 and it is based on a solution for private networks, from Huawei. UMAHetNet contains all the functionalities and network elements for packet routing, mobility handling and QoS managing. In UMAHetNet, all the CN elements (MME, S-GW, P-GW, home subscriber server —HSS— and policy and charging rules function —PCRF) are grouped into a single compact equipment, namely, the evolved core network solution (eCNS), as shown in Figure C.1.

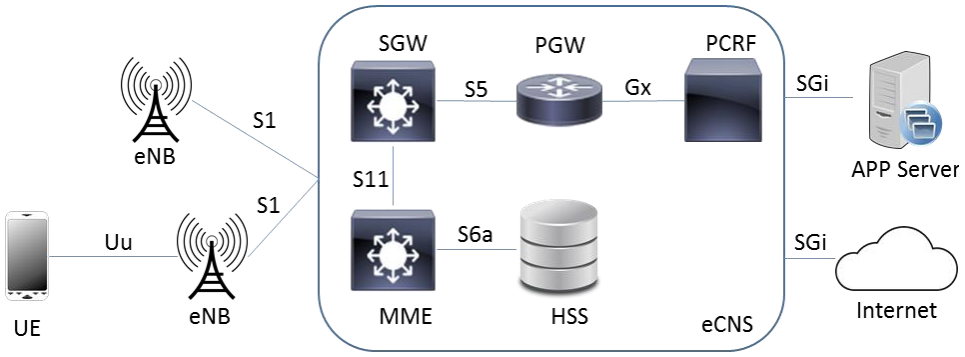


Figure C.1: General scheme of the LTE network deployed at the University of Málaga.

Together with the eCNS, a tool for the management of the whole network is provided: the iManager U2000. This manager assumes the role of the OSS, and as such, is in charge of collecting network performance information and delivering configuration settings to the network elements.

The RAN is composed of 12 picocells, which include both an LTE and a WiFi interface. The infrastructure is completed with a global WiFi controller, which, together with the iManager U2000 are expected to provide LTE/WiFi mobility. The UEs are 12 LTE/WiFi-capable cell phones and two phone-based drive test tools, to measure both RAN performance metrics and service-specific user-centric quality indicators.

Both the picocells (Figure C.2a) and the eCNS (Figure C.2b) are fully configurable. The picocells can be managed either directly through a local management terminal client or remotely, by means of an ad-hoc application to manage the U2000 tool. This allows the user to fully customize all the network parameters, as well as to monitor the network status using built-in and user-defined KPIs. Furthermore, this tool allows the user to integrate his own SON applications and algorithms, from self-configuration techniques to any self-optimizing and self-healing application.

C.3 Picocell deployment

The 12 picocells have been deployed throughout two buildings and three floors in the "Escuela Técnica de Ingeniería de Telecomunicación" (ETSIT) of UMA. Figures C.3 to C.5 show the location of the picocells, from the first to the third floor, respectively. In these figures, the

(a) Example of a ceiling-mounted picocell.

(b) eCNS600, front view.

Figure C.2: UMAHetNet equipment.

current position of the picocells is presented as a red circle. Blue triangles represent ready-to-plug points. That is, possible locations where to move a picocell, to have a new network layout. This fact allows a great flexibility, enabling high interference scenarios as well as high mobility along corridors and between floors.

C.4 Current research results

Until now, the UMAHetNet has been extensively used in the EU projects ONE5G and MONROE, which are under the H2020 framework. Specifically, the following works have been carried out using this tool: [132] and [133], under the scope of the ONE5G project, and [134] under the scope of the MONROE project. In [132], the E2E throughput is used to drive an MLB-based optimization. To that end, throughput measurements from the application layer (gathered at the UEs) are used to steer the HO margins. In [133], service-specific quality indicators (rather than usual RAN KPIs) are used to configure network configuration parameters. Specifically, the MOS of a file transfer protocol (FTP) service is assessed by adjusting the network bandwidth. Finally, in [134], the MOS of an FTP service is also optimized by tuning different configuration parameters of the RLC layer. Besides these, a number of works have been further submitted to be peer-reviewed.

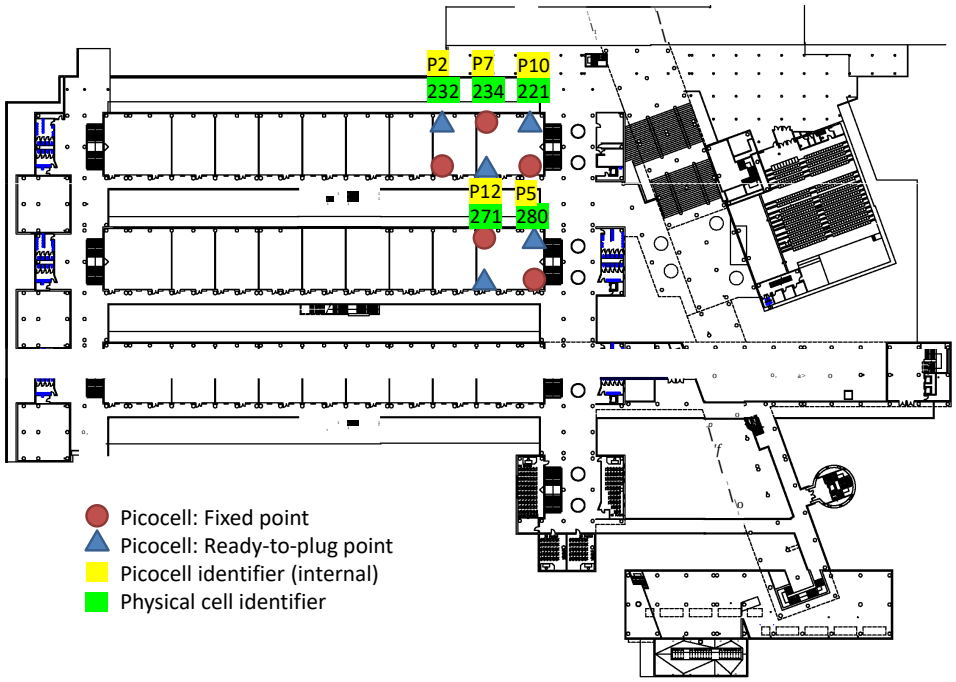


Figure C.3: Distribution of picocells in the ETSIT, first floor.

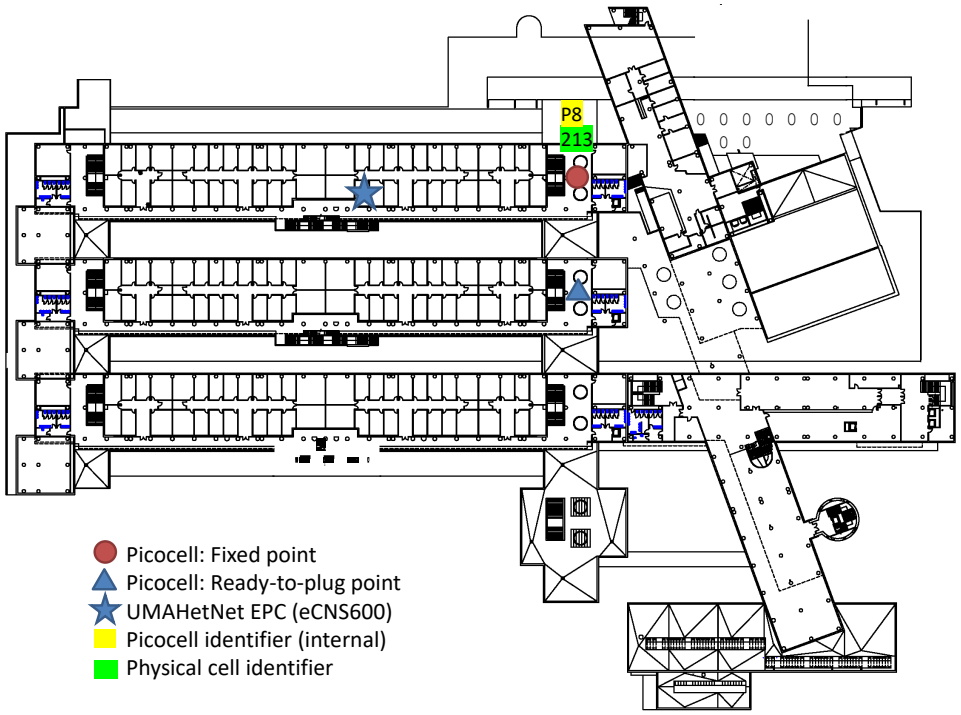


Figure C.4: Distribution of picocells in the ETSIT, second floor.

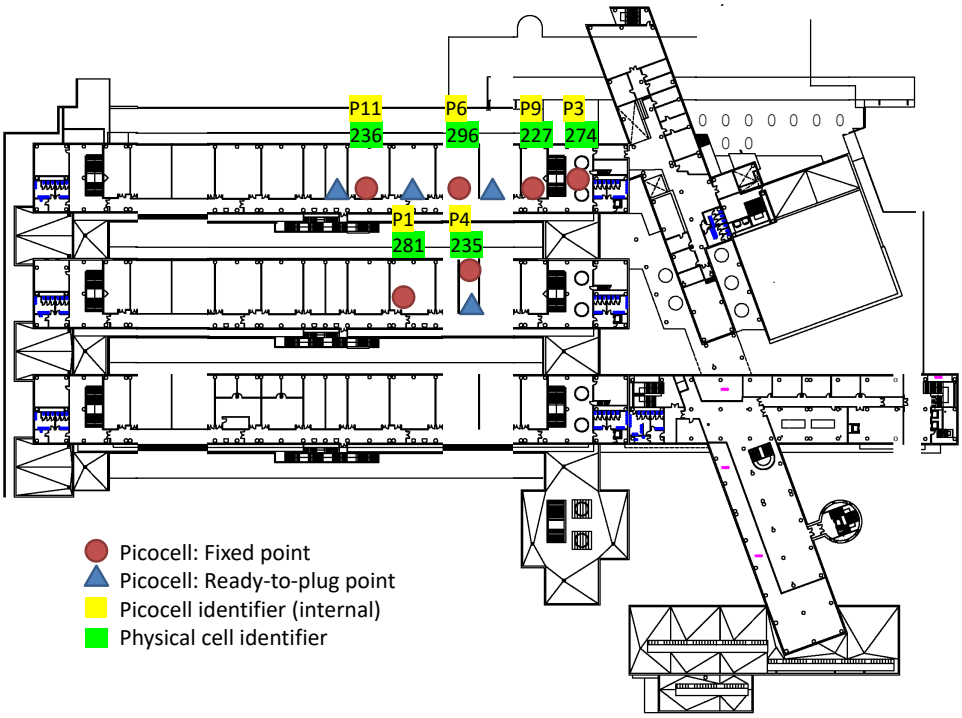


Figure C.5: Distribution of picocells in ETSIT, third floor.

Bibliography

[1] Recommendation ITU-R M.2083-0: IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond, 2015.

[2] Information and Communication Technology (ICT) Statistics, International Telecommunication Union (ITU). ICT Facts and Figures, 2016.

[3] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2; Rel-15, V15.2.0 (2018-07). TS 36.300.

[4] 3GPP. NR; Overall description; Stage-2; Rel-15, V15.2.0 (2018-06). TS 38.300.

[5] NGMN, Recommendation on SON and O&M Requirements, 2008.

[6] NGMN, Use Cases related to Self-Organising Network, Overall Description, 2008.

[7] 3GPP. Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements; Rel-15, V15.0.0 (2018-06). TS 32.500.

[8] A. Imran and A. Zoha. Challenges in 5G: how to empower SON with Big Data for enabling 5G. *IEEE Network*, 28(6):27–33, Nov 2014.

[9] 3GPP. Telecommunication management; Self-configuration of network elements; Concepts and requirements; Rel-15, V15.0.0 (2018-06). TS 32.501.

[10] 3GPP. Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions; Rel-9, V9.3.1 (2011-04). TR 36.902.

[11] 3GPP. Telecommunication management; Self-Organizing Networks (SON); Self-healing concepts and requirements; Rel-14, V14.0.0 (2017-04). TS 32.541.

[12] M. Toril, S. Luna-Ramírez, and V. Wille. Automatic replanning of tracking areas in cellular networks. *IEEE Transactions on Vehicular Technology*, 62(5):2005–2013, Jun 2013.

[13] V. Buenestado, M. Toril, S. Luna-Ramírez, J. M. Ruiz-Avilés, and A. Mendo. Self-tuning of Remote Electrical Tilts Based on Call Traces for Coverage and Capacity Optimization in LTE. *IEEE Transactions on Vehicular Technology*, 66(5):4315–4326, May 2017.

[14] P. Muñoz, R. Barco, J. M. Ruiz-Avilés, I. de la Bandera, and A. Aguilar. Fuzzy Rule-Based Reinforcement Learning for Load Balancing Techniques in Enterprise LTE Femtocells. *IEEE Transactions on Vehicular Technology*, 62(5):1962–1973, Jun 2013.

[15] P. Muñoz, R. Barco, J. M. Ruiz-Avilés, I. de la Bandera, and A. Aguilar. Fuzzy rule-based reinforcement learning for load balancing techniques in enterprise lte femtocells. *IEEE Transactions on Vehicular Technology*, 62(5):1962–1973, Jun 2013.

[16] R. Barco, P. Lázaro, and P. Muñoz. A unified framework for Self-Healing in wireless networks. *IEEE Communications Magazine*, 50(12):134–142, December 2012.

[17] G.F. Ciocarlie, U. Lindqvist, S. Nováczki, and H. Sanneck. Detecting anomalies in cellular networks using an ensemble method. In *Proceedings of the 9th International Conference on Network and Service Management (CNSM 2013)*, pages 171–174, Oct 2013.

[18] G. Ciocarlie, U. Lindqvist, K. Nitz, S. Nováczki, and H. Sanneck. DCAD: Dynamic cell anomaly detection for operational cellular networks. In *2014 IEEE Network Operations and Management Symposium (NOMS)*, pages 1–2, May 2014.

[19] I. de-la Bandera, R. Barco, P. Muñoz, and I. Serrano. Cell outage detection based on handover statistics. *IEEE Communications Letters*, 19(7):1189–1192, July 2015.

[20] Raquel Barco, Luis Díez, Volker Wille, and Pedro Lázaro. Automatic diagnosis of mobile communication networks under imprecise parameters. *Expert Systems with Applications*, 36(1):489–500, 2009.

[21] Raquel Barco, Pedro Lázaro, Volker Wille, Luis Díez, and Sagar Patel. Knowledge acquisition for diagnosis model in wireless networks. *Expert Systems with Applications*, 36(3):4745–4752, April 2009.

[22] P. Szilágyi and S. Nováczki. An automatic detection and diagnosis framework for mobile communication systems. *IEEE Transactions on Network and Service Management*, 9(2):184–197, June 2012.

[23] S. Nováczki. An improved anomaly detection and diagnosis framework for mobile network operators. In *2013 9th International Conference on the Design of Reliable Communication Networks (DRCN)*, pages 234–241, March 2013.

[24] A. Gómez-Andrades, P. Muñoz, I. Serrano, and R. Barco. Automatic root cause analysis for LTE networks based on unsupervised techniques. *IEEE Transactions on Vehicular Technology*, 65(4):2369–2386, April 2016.

[25] A. Gómez-Andrades, R. Barco, I. Serrano, P. Delgado, P. Caro-Oliver, and P. Muñoz. Automatic root cause analysis based on traces for LTE Self-Organizing Networks. *IEEE Wireless Communications*, 23(3):20–28, June 2016.

[26] A. Gómez-Andrades, P. Muñoz, E. J. Khatib, I. de-la Bandera, I. Serrano, and R. Barco. Methodology for the design and evaluation of self-healing LTE networks. *IEEE Transactions on Vehicular Technology*, 65(8):6468–6486, Aug 2016.

[27] 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*, 16(6):1587–1600, June 2017.

[28] Emil J. Khatib, Raquel Barco, Ana Gómez-Andrades, and Inmaculada Serrano. Diagnosis based on genetic fuzzy algorithms for LTE Self-Healing. *IEEE Transactions on Vehicular Technology*, 2015.

[29] Emil J. Khatib, Raquel Barco, Ana Gómez-Andrades, Pablo Muñoz, and Inmaculada Serrano. Data mining for fuzzy diagnosis systems in LTE networks. *Expert Systems with*

Applications, 42(21):7549–7559, 2015.

[30] E. J. Khatib, R. Barco, P. Muñoz, I. D. La Bandera, and I. Serrano. Self-healing in mobile networks with big data. *IEEE Communications Magazine*, 54(1):114–120, January 2016.

[31] Sergio Fortes, Raquel Barco, Alejandro Aguilar-García, and Pablo Muñoz. Contextualized indicators for online failure diagnosis in cellular networks. *Computer Networks*, 82(Supplement C):96–113, 2015.

[32] S. Fortes, A. Aguilar Garcia, J. A. Fernandez-Luque, A. Garrido, and R. Barco. Context-aware self-healing: User equipment as the main source of information for small-cell indoor networks. *IEEE Vehicular Technology Magazine*, 11(1):76–85, March 2016.

[33] P. Muñoz, I. de la Bandera, E. J. Khatib, A. Gómez-Andrades, I. Serrano, and R. Barco. Root cause analysis based on temporal analysis of metrics toward self-organizing 5G networks. *IEEE Transactions on Vehicular Technology*, 66(3):2811–2824, March 2017.

[34] P. Muñoz, R. Barco, E. Cruz, A. Gómez-Andrades, E. J. Khatib, and N. Faour. A method for identifying faulty cells using a classification tree-based UE diagnosis in LTE. *EURASIP Journal on Wireless Communications and Networking*, 2017(1):130, Jul 2017.

[35] I. de la Bandera, P. Muñoz, I. Serrano, and R. Barco. Improving cell outage management through data analysis. *IEEE Wireless Communications*, 24(4):113–119, Aug 2017.

[36] I. de la Bandera, P. Muñoz, I. Serrano, and R. Barco. Adaptive cell outage compensation in self-organizing networks. *IEEE Transactions on Vehicular Technology*, 67(6):5231–5244, June 2018.

[37] Z. Altman et al. The Celtic Gandalf framework. In *Proc. of IEEE Mediterranean Electrotechnical Conference (MELECON)*, 2006.

[38] ICT-2007-216248 E3. Project Presentation Report. Technical Report Deliverable D0.2, Version 1.0, May, 2008.

[39] INFSO-ICT-216284 SOCRATES. Use Cases for Self-Organising Networks. Technical Report Deliverable D2.1, Version 1.0, March, 2008.

[40] INFSO-ICT-224344 Self-NET. System Deployment Scenarios and Use Cases for Cognitive Management of Future Internet Elements. Technical Report Deliverable D1.1, Version 1.0, October, 2008.

[41] FP7-257513 UniverSelf. Self-diagnosis and self-healing for IMS VoIP and VPN services. Technical Report Case study - Part I, Version 1.0, September, 2012.

[42] INFSO-ICT-316384 SEMAFOUR. Self-Management for Unified Heterogeneous Radio Access Networks. Technical Report Deliverable D6.1, Version 1.0, October, 2012.

[43] CP08-004 COMMUNE. Outline Requirements for COMMUNE. Technical Report Deliverable D2.1, Version 1.0, April, 2012.

[44] P. Muñoz, R. Barco, and I. de la Bandera. On the potential of handover parameter optimization for self-organizing networks. *IEEE Transactions on Vehicular Technology*, 62(5):1895–1905, Jun 2013.

[45] Pablo Muñoz, R. Barco, D. Laselva, and P. Mogensen. Mobility-based strategies for traffic steering in heterogeneous networks. *IEEE Communications Magazine*, 51(5):54–62, May 2013.

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[47] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[48] Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[49] Jinjiang Wang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 2018.

[50] 3GPP. Study on scenarios and requirements for next generation access technologies, Rel-14, V14.3.0 (2017-08). TR 38.913.

[51] 3GPP. General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access; Rel-15, V15.4.0 (2018-06). TS 23.401.

[52] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation; Rel-15, V15.2.0 (2018-07). TS 36.211.

[53] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification; Rel-15, V15.2.0 (2018-07). TS 36.321.

[54] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification; Rel-15, V15.1.0 (2018-07). TS 36.322.

[55] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification; Rel-15, V15.0.0 (2018-07). TS 36.323.

[56] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA) Radio Resource Control (RRC); Protocol Specification, Rel-15, V15.2.2, (2018-07). TS 36.331.

[57] C. Rosa, K. Pedersen, H. Wang, P. H. Michaelsen, S. Barbera, E. Malkamaki, T. Henttonen, and B. Sebire. Dual connectivity for LTE small cell evolution: functionality and performance aspects. *IEEE Communications Magazine*, 54(6):137–143, June 2016.

[58] 3GPP. NR; Multi-connectivity; Overall description; Stage-2; Rel-15, V15.2.0 (2018-06). TR 37.340.

[59] J. Rao and S. Vrzic. Packet Duplication for URLLC in 5G: Architectural Enhancements and Performance Analysis. *IEEE Network*, 32(2):32–40, March 2018.

[60] S. Hamalainen, H. Sanneck, and C. Sartori. *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency.* Wiley, 2011.

[61] K. Hamied J. Ramiro. *Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE.* Wiley, 2011.

[62] A. Aguilar-Garcia, S. Fortes, A. Fernandez Duran, and R. Barco. Context-aware self-optimization: Evolution based on the use case of load balancing in small-cell networks. *IEEE Vehicular Technology Magazine*, 11(1):86–95, March 2016.

[63] P. Muñoz, R. Barco, I. Serrano, and A. Gómez-Andrades. Correlation-based time-series analysis for cell degradation detection in SON. *IEEE Communications Letters*, 20(2):396–

399, Feb. 2016.

[64] Hong Liu, Guangju Chen, Guoming Song, and TaiLin Han. Analog circuit fault diagnosis using bagging ensemble method with cross-validation. In *International Conference on Mechatronics and Automation, 2009. ICMA 2009*, pages 4430–4434, Aug 2009.

[65] Hong-Bin Shen and Kuo-Chen Chou. Ensemble classifier for protein fold pattern recognition. *Bioinformatics*, 22(14):1717–1722, 2006.

[66] B.V. Dasarathy and Belur V. Sheela. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5):708–713, May 1979.

[67] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87, March 1991.

[68] S.E. Yuksel, J.N. Wilson, and P.D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, Aug 2012.

[69] Leo Breiman. Bagging predictors. In *Machine Learning*, pages 123–140, 1996.

[70] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, August 1997.

[71] L.I. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, Feb 2002.

[72] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, GeoffreyJ. McLachlan, Angus Ng, Bing Liu, PhilipS. Yu, Zhi-Hua Zhou, Michael Steinbach, DavidJ. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[73] R. Barco, P. Lázaro, L. Díez, and V. Wille. Continuous versus Discrete Model in Autodiagnosis Systems for Wireless Networks. *IEEE Transactions on Mobile Computing*, 7(6):673–681, June 2008.

[74] J. Kittler, M. Hatef, R. P W Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar 1998.

[75] Pablo Muñoz, Isabel de la Bandera, Fernando Ruíz, Salvador Luna-Ramírez, Raquel Barco, Matías Toril, Pedro Lázaro, and J. Rodríguez. Computationally-efficient design of a dynamic system-level LTE simulator. *Intl. Journal of Electronics and Telecommunications*, 57(3):347–358, April 2011.

[76] C. Mehlführer, M. Wrulich, J. Colom Ikuno, D. Bosanska, and M. Rupp. Simulating the Long Term Evolution Physical Layer. In *Proc. of 17th European Signal Processing Conference (EUSIPCO)*, 2009.

[77] 3GPP. OFDM-HSDPA System level simulator calibration (R1-040500). 3GPP TSG-RAN WG1 37, 3^{rd} Generation Partnership Project (3GPP) May 2004.

[78] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, Jul 1997.

[79] J. Dai, L. Lieu, and D. Rocke. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, 5(1), Feb 2006.

[80] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5):971–989, September 2016.

[81] J. J. Saucedo-Dorantes, M. Delgado-Prieto, R. A. Osornio-Rios, and R. de Jesus Romero-Troncoso. Multifault diagnosis method applied to an electric machine based on high-dimensional feature reduction. *IEEE Transactions on Industry Applications*, 53(3):3086–3097, May 2017.

[82] Zhang Hongli, Lu Gang, T. Qassrawi Mahmoud, Zhang Yu, and Yu Xiangzhan. Feature selection for optimizing traffic classification. *Computer Communications*, 35(12):1457 – 1471, 2012.

[83] M. Shafiq, X. Yu, A. A. Laghari, and D. Wang. Effective feature selection for 5G IM applications traffic classification. 2017(ID 6805056), May 2017.

[84] J. Yang, Z. Ma, C. Dong, and G. Cheng. An empirical investigation into CDMA network traffic classification based on feature selection. In *The 15th International Symposium on Wireless Personal Multimedia Communications*, pages 448–452, Sept 2012.

[85] Marton Kajó and Szabolcs Nováczki. A genetic feature selection algorithm for anomaly classification in mobile networks. In *19th International ICIN conference - Innovations in Clouds, Internet and Networks*, Mar. 2016.

[86] Nadir Omer Fadl Elssied, Othman Ibrahim, and Ahmed Hamza Osman. A novel feature selection based on one-way ANOVA F-test for e-mail spam classification. *Research Journal of Applied Sciences*, 7(3):625 – 638, Jan. 2014.

[87] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 507–514, Cambridge, MA, USA, 2005. MIT Press.

[88] Wei Yang, Kuanquan Wang, and Wangmeng Zuo. Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, 7(1):161–168, 2012.

[89] Frank Bauer and Mark A. Lukas. Comparingparameter choice methods for regularization of ill-posed problems. *Mathematics and Computers in Simulation*, 81(9):1795 – 1841, 2011.

[90] Emil J. Khatib, Ana Gómez-Andrades, Inmaculada Serrano, and Raquel Barco. Modelling LTE solved troubleshooting cases. *Journal of Network and Systems Management*, Feb 2017.

[91] Geoffrey McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley series in probability and statistics. Aug. 2004.

[92] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 333–342, 2010.

[93] R. A. Stine and J. F. Heyse. Non-parametric estimates of overlap. *Statistics in medicine*, 20(2):215–36, 2001.

[94] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[95] R. P. W. Duin. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, C-25(11):1175–1179, Nov 1976.

[96] Z. Wang, Y. Zhang, Z. Chen, H. Yang, Y. Sun, J. Kang, Y. Yang, and X. Liang. Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 755–758, July 2016.

[97] Thomas Rückstieß, Christian Osendorfer, and Patrick van der Smagt. *Sequential Feature Selection for Classification*, pages 132–141. Springer Berlin, 2011.

[98] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

[99] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York, 2nd edition, 2002.

[100] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In Wulfram Gerstner, Alain Germond, Martin Hasler, and Jean-Daniel Nicoud, editors, *Artificial Neural Networks — ICANN'97*, pages 583–588, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.

[101] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287 – 314, 1994. Higher Order Statistics.

[102] A. Gramfort and G. Varoquaux. FastICA on 2D point clouds. *www.scikit-learn.org/stable/auto_examples/decomposition/plot_ica_vs_pca.html*, Sept. 2018.

[103] F. Pedregosa. Swiss roll reduction with LLE. *www.scikit-learn.org/stable/auto_examples/manifold/plot_swissroll.html*, Sept. 2018.

[104] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[105] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

[106] 3GPP. Universal Mobile Telecommunications System (UMTS); LTE; Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Measurement Collection for Minimization of Drive Tests (MDT); Overall description; Stage 2; Rel-14, V14.0.0 (2017-03). TS 37.320.

[107] Y. Wang, K. I. Pedersen, T. B. Sorensen, and P. E. Mogensen. Carrier load balancing and packet scheduling for multi-carrier systems. *IEEE Transactions on Wireless Communications*, 9(5):1780–1789, May 2010.

[108] F. Liu, W. Xiang, Y. Zhang, K. Zheng, and H. Zhao. A novel qoe-based carrier scheduling scheme in lte-advanced networks with multi-service. In *2012 IEEE Vehicular Technology Conference (VTC Fall)*, pages 1–5, Sept 2012.

[109] Z. Chen, G. Cui, C. Zhai, W. Wang, Y. Zhang, and X. Li. Component carrier selection based on user mobility for lte-advanced systems. In *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*, pages 1–5, Sept 2013.

[110] M. A. Lema, M. Garcia-Lozano, S. Ruiz, and D. G. Gonzalez. Improved component carrier selection considering mpr information for lte-a uplink systems. In *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 2191–2196, Sept 2013.

[111] Y. Qi and H. Wang. Qos-aware cell association based on traffic prediction in heterogeneous

cellular networks. *IET Communications*, 11(18):2775–2782, 2017.

[112] Y. Sun, G. Feng, S. Qin, and S. Sun. Cell association with user behavior awareness in heterogeneous cellular networks. *IEEE Transactions on Vehicular Technology*, 67(5):4589–4601, May 2018.

[113] N. Dreyer, A. Moller, Z. H. Mir, F. Filali, and T. Kurner. A data traffic steering algorithm for IEEE 802.11p/LTE hybrid vehicular networks. In *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, pages 1–6, Sept 2016.

[114] M. Ben Brahim, Z. Hameed Mir, W. Znaidi, F. Filali, and N. Hamdi. QoS-aware video transmission over hybrid wireless network for connected vehicles. *IEEE Access*, 5:8313–8323, 2017.

[115] Z. H. Mir and F. Filali. Applications, requirements, and design guidelines for multi-tiered vehicular network architecture. In *2018 Wireless Days (WD)*, pages 15–20, April 2018.

[116] A. Gharsallah, N. Omheni, K. Ghanmi, F. Zarai, and M. Neji. A seamless mobility mechanism for V2V communications. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 1063–1069, Oct 2017.

[117] 3GPP. TS 23.303, Proximity-based services (ProSe), Rel-15, V15.0.0. Technical report, 2017.

[118] 3GPP. TR 36.885, Study on LTE-based V2X Services, Rel-14, V14.0.0. Technical report, 2016.

[119] George F. Riley and Thomas R. Henderson. *The ns-3 Network Simulator*, pages 15–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[120] Nicola Baldo, Marco Miozzo, Manuel Requena-Esteso, and Jaume Nin-Guerrero. An open source product-oriented LTE network simulator based on ns-3. In *Proceedings of the 14th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '11, pages 293–298, New York, NY, USA, 2011. ACM.

[121] 3GPP. TR 22.885, Study on LTE support for Vehicle-to-Everything (V2X) services, Rel-14, V14.0.0. Technical report, 2017.

[122] 3GPP. TR 22.886, Study on enhancement of 3GPP support for 5G V2X services, Rel-15, V15.1.0. Technical report, 2017.

[123] P. Muñoz, R. Barco, and S. Fortes. Conflict resolution between load balancing and handover optimization in LTE networks. *IEEE Communications Letters*, 18(10):1795–1798, Oct 2014.

[124] J. Moysen, M. García-Lozano, L. Giupponi, and S. Ruiz. Conflict resolution in mobile networks: A self-coordination framework based on non-dominated solutions and machine learning for data analytics [application notes]. *IEEE Computational Intelligence Magazine*, 13(2):52–64, May 2018.

[125] R. Kreher. *UMTS performance measurement: A practical guide to KPIs for the UTRAN environment.* 2006.

[126] K. Gaenger R. Kreher. *LTE Signaling: Troubleshooting and Optimization.* 2010.

[127] 3GPP. Telecommunication management; Key Performance Indicators (KPI) for the Evolved Packet Core (EPC); Definitions; Rel-15, V.15.0.0. TS 32.455, 2018.

[128] 3GPP. Telecommunication management; Performance Management (PM); Performance measurements Evolved Packet Core (EPC) network; Rel-15, V.15.0.0. TS 32.426, 2018.

[129] 3GPP. Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2; Rel-15, V.15.0.0. TS 23.272, 2017.

[130] 3GPP. Single Radio Voice Call Continuity (SRVCC); Stage 2; Rel-15, V15.2.0. TS 23.216, 2018.

[131] 3GPP. Mobile radio interface Layer 3 specification; Core network protocols; Stage 3; Rel-15, V15.4.0. TS 24.008, 2018.

[132] C. Gijón, S. Luna-Ramírez, and M. Toril. Un nuevo criterio basado en calidad de experiencia para el balance de carga en redes LTE. In *XXXIII Simposium Nacional de la Unión Científica Internacional de Radio (URSI)*, Granada, Spain, Sept. 2018.

[133] A. Herrera, S. Fortes, E. Baena, and R. Barco. KPI-to-KQI metrics mapping. In *ONE5G Demonstration at the European Conference on Networks and Communications (EuCNC)*, Ljublana, Slovenia, Jun. 2018.

[134] J. Mendoza, E. Baena, I. de-la Bandera, D. Palacios, S. Fortes, and R. Barco. MONROE project. *eSON: Network Self-optimization based on End-to-end Measurments*. D2: Final Report, Jul. 2018.