Network Optimisation using Big Data Analytics

Mohammed Safa Hadi

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

School of Electronic and Electrical Engineering

January 2020

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 3 is based on the work from:

M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Big Data Analytics for Wireless and Wired Network Design: A Survey," Computer Networks, 2018.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani, Dr Ahmed Lawey, and my co-supervisor Dr. Taisir Elgorashi.

Chapter 4 is based on part of the work from:

M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Patient-Centric Cellular Networks Optimization using Big Data Analytics," IEEE Access, vol. 7, pp. 49279-49296, 2019.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani, Dr Ahmed Lawey, and my co-supervisor Dr. Taisir Elgorashi.

Chapter 5 is based on part of the work from:

M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Patient-Centric Cellular Networks Optimization using Big Data Analytics," IEEE Access, vol. 7, pp. 49279-49296, 2019.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani, Dr Ahmed Lawey, and my co-supervisor Dr. Taisir Elgorashi.

Chapter 6 is based on the work from:

M. Hadi, A. Lawey, T. El-Gorashi, and J. Elmirghani, "Using Machine Learning and Big Data Analytics to Prioritize Outpatients in HetNets," in IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2019, pp. 726-731

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani, Dr Ahmed Lawey, and my co-supervisor Dr. Taisir Elgorashi.

Chapter 7 is based on the work from:

M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Patient-centric HetNets Powered by Machine Learning and Big Data Analytics for 6G Networks," IEEE Access, vol. 8, pp. 85639-85655, 2020.

This paper has been published jointly with my PhD supervisor prof. Jaafar Elmirghani, Dr Ahmed Lawey, and my co-supervisor Dr. Taisir Elgorashi.

Dedication

My parents provided me with their monumental and unconditional love, continuous and unparalleled support throughout all stages of my life, they deserve the greatest credit. Therefore, I dedicate this thesis to them.

Acknowledgments

This thesis has one author but many people to thank for – family, supervisors, friends, colleagues, and staff.

First and foremost, I would like to thank my supervisor; Professor Jaafar Elmirghani for his mentorship, constant encouragement, and for his trendsetting suggestions that kept my research progressing in the right direction throughout my PhD journey.

I would like to acknowledge my co-supervisor; Dr Taisir Elgorashi for her support and useful discussion during my PhD study.

I wish to express my deep gratitude to my friends Dr Ahmed Lawey, Dr Ahmed Al-Quzweeni, Dr Haider Al-Shammari, and Dr Mohamed Musa for all the constructive discussions, immense encouragement, and support. Without them, I would not have been able to achieve so much.

I would like to thank my beloved friend Samer Majeed for their support, and for bringing joy to my life.

I would like to thank my family for the unconditional love and for standing by me throughout my PhD journey.

Last but not the least, I wish to express my sincere gratitude to my colleagues and the wonderful academic staff in the School of Electronic and Electrical Engineering in the University of Leeds.

Abstract

Interdisciplinary research is fuelling a paradigm shift to endow technology-based services with a personalised dimension. The main contributors for such innovatory change are the surge in data production rate, the proliferation of data generators in the form of IoT and other network-connected devices, the incorporation of innovative data technologies like Artificial Intelligence, Machine Learning and Big Data Analytics, and the advancements in computing powers that are getting closer to dethroning Moor's law and deliver more processing per unit time. Moreover, there is an ever-increasing demand for smart and fast-responsive applications such as predictive analytics, business analysis and digital marketing. In this thesis, patient-centric cellular network optimisation is investigated as a promising paradigm that can contribute to the personalisation of present and future cellular networks with the aim of saving people's lives where every second counts. This calls for transforming current cellular networks from merely being blind tubes that convey data, into a conscious, cognitive, and self-optimizing entity that adapts intelligently according to the users' needs.

The work carried out in this thesis started by comprehensively exploring the role of using big data analytics in network design. Subsequently, we considered incorporating the concepts of priority, e-healthcare, Big Data Analytics, and resource allocation in a single system. The system's goal is to use big data harvested from out-patient electronic health records and body-connected medical Internet of Things sensors to be processed and analysed in a big data analytics engine to predict the likelihood of a stroke. This prediction is then used to ensure that the out-patients are assigned optimal physical resource blocks that provide good signal to interference and noise ratio (SINR) dictated by the severity of their medical state. Hence, granting channels of high spectral efficiency to the out-patients, empowering them to transmit their critical data to the designated medical facility with minimal delay.

The use of several Machine Learning algorithms residing within the big data analytics engine is investigated, namely, a naïve Bayesian classifier, a decision tree classifier, and a logistic regression classifier. Further, the incorporation of the aforementioned classifiers in an ensemble system running as a soft voting classifier is examined and the performance of all classifiers is compared. The combinatorial optimisation problem of maximising the system's overall SINR while prioritising the OPs in terms of radio resource assignment is solved using Mixed Integer Linear Programming and a heuristic. The use of two resource allocation approaches, namely, a Weighted Sum Rate Maximisation approach and a Proportional Fairness approach is considered and compared in terms of fairness and the attained SINRs. The proposed system was extended from a single-tier (homogenous) LTE-A network, to multi-tier Heterogeneous Networks employing spectrum partitioning strategy, and finally to a multi-tier Heterogeneous Network with no interference-mitigation strategies employed. Thus, enabling a further study of the system's performance over different networks and interference strategies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABS	Almost Blank Subframe
AC	Agglomerative Clustering
ANN	Artificial Neural Networks
ANN	Artificial Neural Networks
API	uses application programming interface
ARLAS	Apervi’s Real-time Log Analytics Solution
ASBL	asynchronous sparse Bayesian learning
AWGN	Additive White Gaussian Noise
BDA	Big Data Analytics
BDD	Big Data Driven
B-dIDS	Big-distributed Intrusion Detection System
BMI	body mass index
BP	Blood Pressure
BS	Base Station
CDN	Content Delivery Network
CMPs	Chip Multiprocessors
CoS	Cost of Service
CPU	Central Processing Unit
C-RAN	Cloud-Based Radio Access Network
C-RAN	a cloud-based radio access network
CSPs	Communication Service Provides
CT	Computerized Tomography

CVD	Cardiovascular Disease
DBSCAN	density- based spatial clustering of applications with noise
DDPG	Deep Deterministic Policy Gradient
DoC	Datacentre-on-Chip
DQN	Deep Q Network
DT	Decision Tree
eICIC	enhanced Inter-Cell Interference Coordination
eMBB	enhanced Mobile Broadband
eNB	evolved Node B
ENN	Elman neural network
ESC	European Society of Cardiology
ESH	European Society of Hypertension
FEC	Forward Error Correction
FN	False-Negative
FNR	False-Negative Rate
FP	False Positive
FPRB	Free Physical Resource Block
GBM	Generalized Boosted Model
GIS	geographic information system
GPS	geographic positioning system
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HetNet	Heterogeneous Network
HTTP	Hyper Text Transfer Protocol
ICU	Intensive Care Unit

ILP	Integer Linear Programming
IMM	Interacting Multiple Model
IMM	interacting multiple model
IoT	Internet-Of-Things
IP	Internet Protocol
iRODS	integrated Rule-Oriented Data
KNN	K-Nearest Neighbour
KPI	Key Performance Indicator
LAA	LTE-licence assisted access
LAC	Location Area Code
LBA	low-complexity beam allocation
LR	Logistic Regression
LTE-A	Long Term Evolution
LTE-U	LTE Unlicensed
MAC	media access control
Mb	Mega Bit
MBS	Macro Base Station
MeNB	Macro cells evolved Node B
MILP	Mixed Integer Linear Programming
MIMO	multi-input multi-output
MKF	Multi-Kalman Filter
ML	Machine Learning
mMTC	massive machine-type communication
MNO	Mobile Network Operators
NB	Naïve Bayesian

NE	Nash Equilibrium
NMF	Non-Negative Matrix Factorisation
NOMA	non-orthogonal multiple access
NP	Network Parameter
NPV	Negative Predictive Value
NR	without reservation
NSA	National Security Agency
NVM	Non-volatile memory
O&M	Operation and Maintenance
OBO	Operational and Business Objectives
OD	Origin-Destination
OFDMA	Orthogonal Frequency Division Multiple Access
OP	Out-Patients
PBS	Pico Base Station
PCA	Principal Component Analysis
PDI	Probability Distribution Identification
PDS	post-decision state
PDS-ERT	post-decision state based experience replay and transfer
PE	Provider Edge
PF	Proportional Fairness
PI	Performance Indicator
PPV	Positive Predictive Value
PRB	Physical Resource Block
QKT	Q-learning algorithm with knowledge transfer
QL	Q-Learning

QoE	Quality of Experience
QoS	Quality Of Service
RAN	Radio Access Network
RDBMS	Relational Database Management System
RF	random forests
ROP	Resource Output Period
RSA	Routing and Spectrum Allocation
SARSA	State-Action-Reward-State-Action
SBPP	Shared Backup Path Protection
SD	Standard Deviation
SDN	Software Defined Network
SDN	Software Defined Networks
SeNB	Small cell eNBs
SINR	Signal to Interference Plus Noise
SU	Secondary Users
SVM	Support Vector Machines
SWPs	Spectrum Windows Planes
TCO	Total Cost of Ownership
TIED	Tunnel Endpoint Identifier
TN	True Negative
TP	True Positive
TTI	Transmission Time Interval
UE	User Equipment
URI	Uniform Resource Identifier
uRLLC	ultra-Reliable Low Latency Communications

VNT	Virtual Network Topology
WCDMA	Wideband Code Division Multiple Access
WEKA	Waikato Environment for Knowledge Analysis
WSN	Wireless Sensor Network
WSRMax	Weighted Sum Rate Maximisation
YARN	Yet Another Resource Negotiator
YALE	Yet Another Learning Environment

Chapter 1
Introduction

1.1 Background

Prior to the emergence of big data, decisions were made relying on data samples. Consequently, the decisions were semi-optimum. Those ill-informed decisions spanned over different areas from marketing to law enforcement, sports, and healthcare. The powerful capability of big data analytics (BDA) in analysing massive amounts of data and inferring knowledge from it has brought about better predictions paving the way for better decisions.

Healthcare is a vital subject due to its role in people's lives. The continuous increase in the world population and other factors, like insufficient healthcare budgets, has resulted in crowded hospitals, over-worked medical staff, and extended queuing times for the patients. Given the global nature of the problem, researchers are developing new approaches to improve the level of care delivered by healthcare providers while ensuring a reduction in all previously mentioned points. BDA can be used to ensure medical service is reaching those most in need, in a timely manner. Brain strokes are one of the rising health issues and though they might cause significant disabilities to the patient, immediate treatment can effectively increase recovery chances. According to statistics from England, Wales and Northern Ireland for 2016-2017, one-third of stroke patients arrived at the hospital unaware of the date and time their symptoms began. The severity of this matter is even starker when knowing that the average waiting time for a patient from the start of symptoms until hospital admission is 7.5 hours, with an additional 55 minutes for door-to-needle time (the time between arriving at an emergency department and having an anaesthetic administered). Adding to all that, the patient is loses 1.9 million neurons each minute until the treatment begins. Thus, a proactive and timely diagnosis is vital.

BDA and machine learning (ML) methods can be optimally utilized to process disparate data such as patient's electronic health record (EHR), diet, genetic data and their daily routine, and produce a quick and accurate diagnosis can be time-consuming and require a certain level of expertise to be carried out by medical personnel. Thus, saving lives, improving the level of care, and lowering costs. It worth mentioning that BDA is reportedly being used to diagnose and predict future complications in patients. Acquiring this diagnosis beforehand gave the medical professionals a head start to address these complications.

In the healthcare sector, there are many sources of big data, for example; medical IoT sensors, wearable sensors, and smartphone medical applications. What the above-mentioned data generators have in common is their reliance on network connectivity. Maintaining this connectivity and ensuring its quality is a dilemma that many researchers tried to solve optimally. In this work, the OP's big data can play a double role. In addition to diagnosis, it can guide the network operator to the OPs with the most pressing needs, Hence, radio resources can be allocated to them. We believe that ensuring high-quality connectivity between the patient-linked peripherals and their healthcare provider is an important step towards highly personalized e-healthcare services and applications.

A wireless connection is preferred over a wired one for what it has to offer in terms of mobility. Consequently, cellular and Wi-Fi are the most popular connectivity technologies. The level of freedom (mobility-wise) varies between wireless technologies, for example, Wi-Fi may provide an adequate data rate, nevertheless, it forces an Out-Patient (OP) that needs to keep their medical IoT sensor (e.g. IoT pacemaker) connected, to stay within a relatively small coverage area (i.e., indoors mainly). Utilizing the already-existing cellular networks can provide much-needed freedom to that OP. However, cellular connections can experience channel fading and path loss where the connection can become unreliable or cannot be established due to a very low signal to interference plus noise ratio (SINR). A slow fading channel may indicate that the signal level is inadequate at the instance(s) when an OP's critical data must be conveyed urgently to the healthcare provider.

Big data is portrayed in as a next-generation tool that can be used to find an optimal trade-off problem between resource sharing, allocation, and optimisation in wireless networks. Nevertheless, optimizing cellular networks in a user-centric style is still underexplored. In this work, we contend that maintaining a high-quality connection between the OP's medical IoT and the medical provider is a step towards transforming conventional cellular networks into a cognitively personalized e-healthcare-centric service. Building self-adaptive, intelligent, and self-aware network is an operator's high-level objective. Therefore, BDA can endow the network the capability of learning from experience and improving its performance. Thus, BDA can transform the network from being reactive to predictive. we introduce for the first time OP-conscious approaches optimizing the uplink side of a multi-cell Orthogonal Frequency Division Multiple Access (OFDMA) network. In these models, the objective function prioritises the OPs by maximising their SINR received at the Base Station (BS) while keeping the goal of maximising the network's overall SINR.

The network that serves the OPs can either be a dedicated or a non-dedicated network. We chose to optimise a non-dedicated cellular network for several reasons. Firstly, a non-dedicated network can be deployed at a fraction of the cost of a dedicated one and such a network requires much lower commissioning time to be operational. Secondly, the proposed approach can help provide the same level of service to other users while improving the OPs' SINRs. Thirdly, using an established operational network can facilitate the adoption of the proposed approach and the idea of providing such service can be appealing to operators and regulators as it is for the benefit of patients. Fourthly, a dedicated network can limit the mobility of the OPs to within the network's coverage, while using the proposed approach can provide nation-wide (if not more) freedom, especially if it was standardized and regulated.

The models comprise an assignment scheme powered by BDA where OPs are assigned Physical Resource Blocks (PRBs) with powers proportional to their current medical situation. Fairness was incorporated to minimise the negative impact of such assignment on other users. It worth noting that topics that discusses patient monitoring, radio resource allocation, prioritisation, fairness, and ensemble-aided disease risk prediction are popular in the literature across several disciplines. However, proposing a cellular network optimisation framework that incorporates all the above is, to the extent of our knowledge, unique.

1.1.1 Using Big Data Analytics for Cellular Networks Resource Allocation

The topic of utilising BDA in network design was thoroughly discussed in a previous work where we observed that the highest number of papers in this area are in the wireless field [1]. Significant effort is dedicated currently to endowing wireless cellular networks with the ability to seamlessly *prioritise* users and serve them accordingly. Previous work in this area includes the work in [2] who proposed the use of configuration, alarm, and log files and processing the mentioned data using a big data processing environment, thus identifying the behaviour of both the user and the network. The goal is to solve the problem of radio resource allocation to users in the Radio Access Network (RAN) in a manner that ensures minimal delay between resource request and assignment. Another idea was presented by the authors of [3] to manage the network resources in Heterogeneous Networks (HetNets). This was achieved through the utilisation of sentimental and behavioural analysis of data collected from social networks, along with communication network data. The latter was exploited to predict sudden increases in the usage of the mobile network. The aim was to achieve minimal service disruption by servicing the right place at the right time.

1.1.2 Using Big Data Analytics in Healthcare

Several approaches have attempted to address the riddle of employing BDA to accomplish the task of OP monitoring. A system that has a real-time response when an emergency case arises was proposed by the authors in [4]. The system is capable of processing data collected from millions of Wireless Body Area Network (WBAN) sensors. The authors of [5] investigated the challenges associated with designing and implementing big data services that utilise data harvested from medical sensors as well as other IoT applications. They also considered the requirement of processing this data in real-time. Another approach to help patients with Parkinson's disease was proposed by the authors of [6]. The system monitors the loss of flexibility as it is a sign of disease progression. This is done by analysing big data collected from the body and 3D sensors, such as the Microsoft Kinect sensor system. The disease development and treatment effectiveness can both be observed by the patients as well as their healthcare providers in real-time. A survey conducted by the authors in [7] summarised different approaches to detect heart disease at an early stage. The common theme among those approaches is that they are all based on data mining, machine learning (ML), and BDA techniques.

1.1.3 Missing Piece of the Jigsaw

All the approaches mentioned in the previous subsection assumed networks with ideal connectivity. However, in a real-world scenario, opposing elements like channel fading and noise need to be taken into consideration. Our approach exploits BDA for the purpose of optimising the RAN side of a Long Term Evolution-Advanced (LTE-A) network to serve a specific category of people, in this case, the OPs. Our approach ensures service availability to OPs, especially at times when they are in desperate need for it. We argue that by analysing the OPs' big data we can predict the ones that are at high risk of having a stroke. It should be noted that strokes are the medical condition studied in this thesis, however our network optimisation frame is general and can be used to cater for other types of patients, with single or multiple long term conditions so long as they can be prioritised depending on the severity of their conditions (using machine learning for example). Consequently, OPs will be prioritised over normal users and the network's attention (in terms of the quality of the assigned resources) can be shifted towards them. In the US, about **795 thousand people** suffer a stroke annually [8]. This is equivalent to **1.5 stroke incidents per minute** on average which is significant and frequent. In England, Northern Ireland and Wales, a third of stroke patients went to the hospital during 2016-2017 not knowing what time their symptoms commenced [9]. The problem is serious given an average time from the start of the symptoms till admission to a hospital **of 7.5 hours, with another 55 minutes door-to-needle time** (duration between arrival at the emergency department and administering an anaesthetic) and the fact that a stroke **patient is loses on average 1.9 million neurons each minute** before treatment commence [9]. The use of our proposed system can have a tremendous impact on minimising this time since patients are prioritised and given reliable resources. Moreover, the increase in the SINR will result in an increase in the spectral efficiency hence fewer resources are required to transmit the same amount of data [10]. The proposed system can also help in providing reliable connectivity to medical IoT devices when transmitting the patient's vital signs to the healthcare provider. In addition, it can help with early detection of symptoms and facilitate early emergency admittance to the hospital to help save patients' lives. If other forms of ill health are included, the proposed system will be called upon even more frequently. It should be noted that the delay component from the collection of outpatient's current state till the processing of data in the cloud is negligible in comparison to the 7.5 hours and 55 minutes figures quoted earlier, hence, it is not considered in this thesis.

In terms of the need to respond fast to the channel variation and the changes in patients' needs, we would like to note that the MILP is used only to establish the optimal solution, while the simple heuristic developed is used to provide the fast response needed (at the cost of sub-optimal, but good performance).

The wireless channel might change in a fast way, nevertheless, for optimisation purposes, the coherence time of the wireless network in a slow-fading channel is assumed to be longer than the duration of one transmission time interval (TTI) as observed in the literature [11-14]. Thus, the channel state remains essentially constant for the duration of one TTI. Despite the time constraints, the use of MILP to find the optimal resource allocation is for reference only. MILP is a popular tool for optimising many real-time problems, including the uplink and downlink of cellular networks. Many examples of such use cases can be found in the literature. The authors in [15] used MILP (and a heuristic) to jointly minimise network power consumption and transmission delay in an LTE network. Fairness of dynamic channel allocation was investigated by [16]. The authors in [17] used MILP to minimise the number of femtocells in an enterprise environment while guaranteeing a minimum threshold SINR. The authors in [18] proposed a MILP model and a near-optimal metaheuristic to maximise the SINR subject to user power and subcarrier assignment constraints in the uplink of an OFDMA network. The authors of [19] proposed a MILP-based optimisation framework to study the optimal performance of the uplink in HetNets. Several admission control policies for uplink WiMAX networks were proposed by the authors in [20]. The authors employed MILP and a heuristic for that purpose.

At the patient's end, the authors in [21] emphasised that home-measured blood pressure has stronger predictive power than conventional blood pressure measurements. Additionally, the authors concluded that while there is no specific threshold (within the range of 1-14) for the number of measurements, they suggested as many as 14 or more measurements per day can enhance the prediction of a stroke. Taking the worst-case scenario by doubling this number (i.e., 28 measurements/day), the proposed system still only performs measurements and predictions every 50 minutes which is more than sufficient.

Lastly, we would like to draw attention to the fact that what we have integrated with our proposed approach the ability to access OP's vital signs, classify their medical state, and optimise the network in light of this state while taking into consideration other (healthy) users.

1.2 Research Objectives

The primary research objectives of this thesis can be summarised as follows:

1- To develop a framework that uses BDA to endow cellular networks with the ability to prioritise users (i.e., OPs) and serve them accordingly, while keeping the impact on other network users to a minimum.

2- To quantify the likelihood of a stroke in an OP as a risk factor using BDA methods and transform this likelihood into a priority granted to the OPs during the radio resource assignment stage.

3- To develop an approach to maximise the OPs' SINRs along with the total system SINR by maximising the individual sum-rates of the users' SINRs.

4- To investigate the wireless network response to OP prioritisation in an LTE-A network comprised of Macro BSs.

5- To inspect the performance of the prioritisation approaches in a HetNet environment where inter-tier interference is present.

6- To examine the system response over time using different probabilities of stroke.

1.3 Original Contributions

The main contributions of this thesis are as follows:

1- Surveyed the role BDA can play in wireless and wired network design. As a result, we made the following contributions: (i) helping academic researchers save much effort by understanding the state-of-the-art and identifying the opportunities, as well as the challenges facing the use of BDA in network design; (ii) in addition to academic approaches, we surveyed network equipment manufacturing companies highlighting network solutions based on BDA; (iii) we also identified the common areas of interest among these solutions, and thus the conducted survey can benefit both academic and industrial-oriented readers.

2- Developed MILP models to prioritise the OPs in terms of radio resource allocation in an LTE-A network. As a result, we made the following contributions: (i) the introduction of an interdisciplinary approach to optimise the uplink of a LTE-A network while prioritising cellular-connected-OPs using BDA and MILP optimisation to grant the OPs suitable PRBs according to their current health condition; (ii) the development of method to determine the likelihood of a stroke using a naïve Bayesian classifier and real patient big data sets; (iii) we developed, using MILP, two approaches to maximise the OPs' SINRs, namely, the weighted sum-rate maximisation (WSRMax) approach and the (proportional fairness) PF approach and compared them in terms of the fairness achieved between the users and the increase in the SINR.

3- Developed a MILP model to prioritise the OPs in terms of radio resource allocation in uplink HetNets where inter-cell interference is mitigated by employing a spectrum partitioning strategy and thus made the following contributions: (i) investigated the system response over seven different current states resulting in different priority levels granted to the OPs. A *current state* refers to a feature vector of several values acquired by medical and IoT sensors (e.g., total cholesterol and blood pressure) that we run through the classifier to determine stroke probability; (ii) examined the system response in HetNets with activated spectrum partitioning strategy in terms of fairness and the percentage of maximised OPs' SINRs over 300 instances representing different network realisations.

4- Developed a MILP model to prioritise the OPs in terms of radio resource allocation in uplink HetNets: (i) extending the aforementioned work to include a larger dataset, incorporating the decision tree (DT), the logistic regression (LR), and the naïve Bayesian (NB) classifiers in an ensemble system where a voting classifier resides; (ii) rigorously scrutinising the classifiers' performance by conducting various tests of accuracy, recall, specificity, false-positive rate, false-negative rate, negative prediction rate, precision, and F1 score. Furthermore, reporting the cross-validation test scores for all datasets; (iii) extending the aforementioned work to study the effects of inter-cell and intra-cell interference in HetNets and added a reliability-aware aspect to the PF approach; (iv) testing the fairness among users, and conducting the required sensitivity analysis over 300 instances.

1.4 Related Publications

This work resulted in the following journal and conference papers that have been published:

1- M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Big Data Analytics for Wireless and Wired Network Design: A Survey," Computer Networks, 2018.

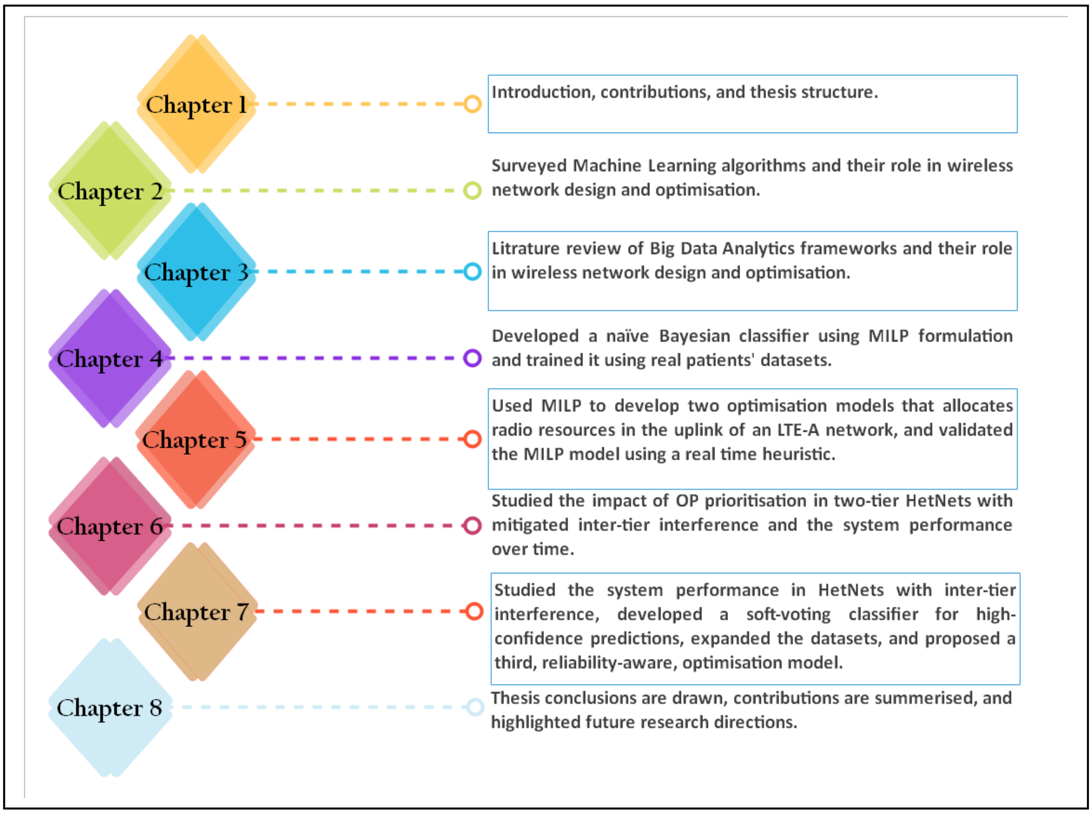
2- M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Patient-Centric Cellular Networks Optimization using Big Data Analytics," IEEE Access, vol. 7, pp. 49279-49296, 2019.

3- M. Hadi, A. Lawey, T. El-Gorashi, and J. Elmirghani, "Using Machine Learning and Big Data Analytics to Prioritize Outpatients in HetNets," in IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2019, pp. 726-731

4- M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Patient-centric HetNets Powered by Machine Learning and Big Data Analytics for 6G Networks," IEEE Access, vol. 8, pp. 85639-85655, 2020.

1.5 Thesis Structure

The relationship between the chapters is presented as follows; a literature review is presented in Chapter 2 where we illustrated the various types of ML algorithms and discussed their role in the subject of radio resource allocation. In Chapter 3, we presented a literature review showing how the power of prediction provided by various BDA frameworks is employed in wireless network design and optimisation. We laid the foundation of our proposed work in Chapter 4 where we developed a naïve Bayesian classifier using MILP formulation and we trained it using a dataset of 30 entries. The work of this chapter served as the core of our BDA engine for Chapter 5 and Chapter 6. In Chapter 5, we considered using MILP to develop two optimisation models for the allocation of radio resources in the uplink of an LTE-A network. The OPs were allocated PRBs with powers relative to the seriousness of the OPs' medical state and we tackled the concept of fairness during that allocation. Further, we developed a heuristic using MATLAB to validate the MILP models and provide a semi-optimal but faster result. Additionally, the system's computational complexity was calculated. We expanded our work in Chapter 6 to study the impact of OP prioritisation in two-tier HetNets with mitigated inter-tier interference using spectrum partitioning strategy. Moreover, we considering several measurements (current states) to reflect a change in the OPs' current health condition. Thus, observing the system response over time. Our work was further extended in Chapter 7 to include HetNets with existing inter-tier interference. Further, aiming for high-confidence predictions, we developed a soft-voting classifier where the predictions of three ML algorithms, namely, a naïve Bayesian classifier, a decision tree classifier, and a logistic regression classifier are combined to produce high-confidence predictions. Furthermore, we scrutinised the soft-voting classifier (i.e., our BDA engine) through several performance metrics and cross validation tests and using an expanded dataset of 200 entries. Finally, A third optimisation model with the concept of reliability is introduced and compared to against the other two models. The thesis concludes with Chapter 8 where conclusions are drawn, the major contributions of this work are summarised, and we highlight the proposed future directions. A summary for the above structure is illustrated in Figure 1-1.



Introduction, contributions, and thesis structure.

Surveyed Machine Learning algorithms and their role in wireless network design and optimisation.

Litrature review of Big Data Analytics frameworks and their role in wireless network design and optimisation.

Developed a naïve Bayesian classifier using MILP formulation and trained it using real patients' datasets.

Used MILP to develop two optimisation models that allocates radio resources in the uplink of an LTE-A network, and validated the MILP model using a real time heuristic.

Studied the impact of OP prioritisation in two-tier HetNets with mitigated inter-tier interference and the system performance over time.

Studied the system performance in HetNets with inter-tier interference, developed a soft-voting classifier for high-confidence predictions, expanded the datasets, and proposed a third, reliability-aware, optimisation model.

Thesis conclusions are drawn, contributions are summerised, and highlighted future research directions.

Figure 1-1: Thesis structure

Chapter 2
Role of Machine Learning in Wireless Network Optimisation

2.1 Introduction

More intelligence is required to overcome the current limitations imposed upon cellular networks. Given its ability to predict a future incident and maximising a reward by learning a certain policy from historical data, whether this data was labelled or not, Machine Learning is, for this reason, a crucial element to enable networks full automation in areas like radio resource optimisation, network management, cache optimisation, backhaul optimisation, capacity and user mobility patterns discovery, coverage optimisation, and spectrum learning in cognitive radio. In this chapter, we conduct a brief literature review of the field of Machine Learning. In this review, we describe the important concepts in this vital field, the well-known algorithms, and illustrate its growing role in communication networks design.

2.2 Machine Learning Types

Learning, by definition, is the process of getting better results carrying out a certain task through practice. Thus, we can have a simple broad definition for Machine Learning (ML) as the process where a computer program improves its performance in upcoming tasks through experience gained from observed data [22]. The main field of ML is divided into three subfields as shown in Figure 2-1. The main branches are supervised learning, unsupervised learning and reinforcement learning. Although ML is not a newly devised field, however its current popularity is due to the large amount of generated data and the high computational power supplied by available now in modern-day computers.

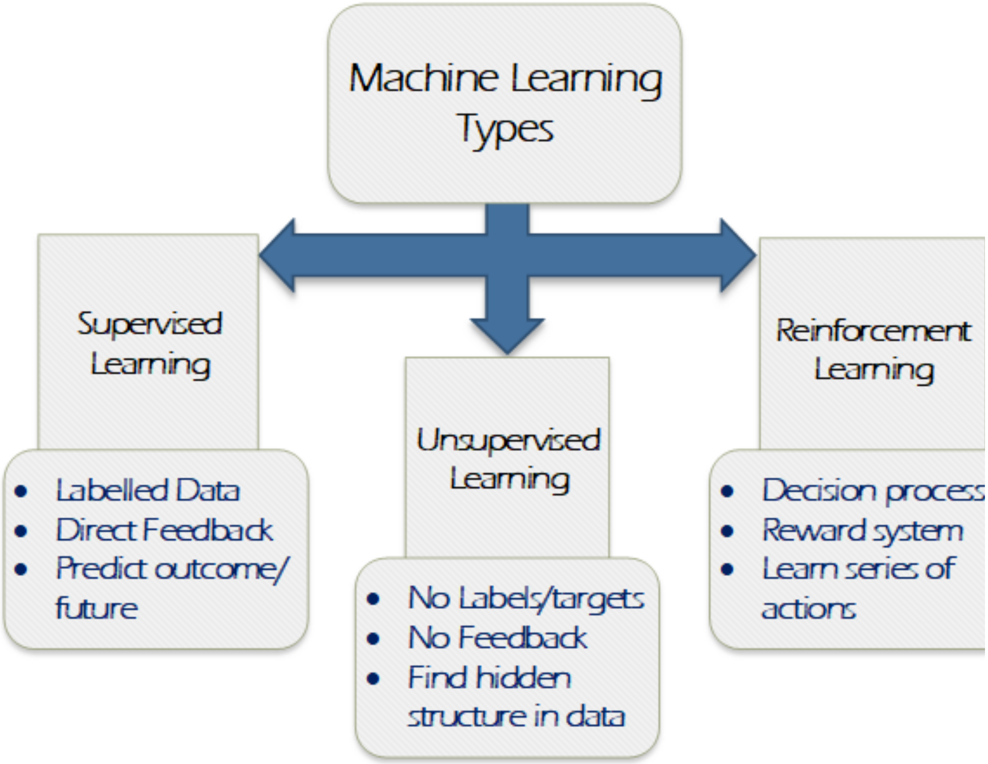


Figure 2-1: Machine Learning Types

2.2.1 Supervised Learning

This type of machine learning relies on the existence of a labelled dataset. This means that for every ***feature*** (or independent) variable, there is a ***class*** (dependent or target) variables. In this type of learning, the right outcomes are already known during the model training and they are provided by a *supervisor*. The learning target is to learn a mapping from the input feature variables to the output class variables. This trained model is then used to predict the outcome of future (unseen) data. The term ***supervised*** refers to the dataset where the class (target) variables are already known. This type of ML can be further divided into two subcategories; ***classification***, where the model uses previous observations (training dataset) to predict the *categorical* class of a new vector of feature variables (i.e. instance), and ***regression***, where the class (outcome) is a number (i.e., continuous value) [23, 24]. There are a number of algorithms classified as supervised learning algorithms, the most popular are K-nearest Neighbours (K-NN), Naïve Bayes (NB), Support Vector

Machines (SVM), Logistic Regression (LR), Linear Regression, DT, and Artificial Neural Networks (ANN).

2.2.2 Unsupervised Learning

In this type of learning, there is no supervisor. Hence, the outcome (class) variable are unknown and the data are unlabelled or of unknown structure. The task in unsupervised learning is to extract meaningful information by exploring the structure of the training data without supervision (i.e., unknown outcome variable with no reward function). This type of learning can be further divided into two subcategories; ***clustering***, which enables the organisation of data objects in the form of meaningful subgroups or *clusters.* It should be noted that objects within the cluster share some similarities and are more different to the objects belonging to other clusters (groups). The other subcategory is ***dimensionality reduction***, which takes a dataset of high-dimensionality (i.e., high number of feature variables) and finds a way to remove noise from data (remove unimportant feature variables). This learning subgroup is useful to enhance the algorithm's prediction performance by lowering the storage and computational requirements of ML algorithms. Algorithms falling into this type of learning include [23, 24]; K-means, hidden Markov Model, Principal Component Analysis (PCA), density-based spatial clustering of applications with noise (DBSCAN), Non-Negative Matrix Factorisation (NMF), Agglomerative Clustering (AC).

2.2.3 Reinforcement Learning

The goal in this type of learning is to develop a system (agent) that learns to maximise a reward given by a reward function measuring the agent's interaction with the environment via trial and error. This type of ML is somewhat linked to supervised learning. However, it differs from it in that the feedback received by the agent is not a class label or a certain class value, but rather it is a reward (e.g., win or lose) [23]. Algorithms of this type of learning include; Q-Learning, State-Action-Reward-State-Action (SARSA), Deep Q Network (DQN), and Deep Deterministic Policy Gradient (DDPG).

A visual representation of the types of ML, their learning subcategories, and common applications is illustrated in Figure 2-2.

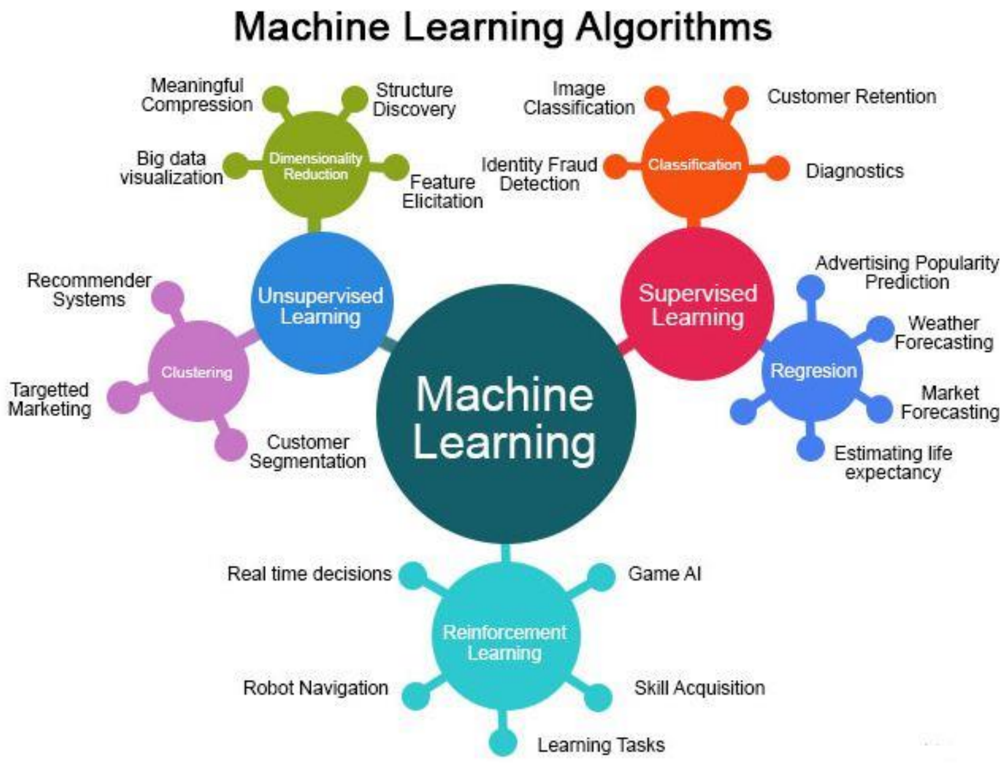


Figure 2-2: ML algorithms, subcategories, and applications **[25]**

2.3 Learning Tools

ML algorithms can be virtually programmed using the majority of programming languages. In this section, we illustrate the most common tools that can be used to build and run ML algorithms. We can classify ML tools into two categories; programming-based tools and Graphical User Interface (GUI)-based tools

2.3.1 Programming-Based Tools

2.3.1.1 R

This open source tool is a combination of both a language and an environment. Created by Ross Ihaka and Robert Gentleman, R can be described as a dedicated tool for statistical purposes that helps data scientists create codes with less programming requirements [26].

2.3.1.2 Python [27]

Python is described as the *lingua franca* for a major number of data science applications as it joins ease of use of domain-specific scripting language (e.g., R and

MATLAB) with the power of being a general-purpose programming language. Python's powerful libraries support data scientists with both general and specific-purpose libraries. Furthermore, it can be integrated with existing systems. Python benefits from scikit-learn, which is a widely used in both academic and industrial communities and a constantly-developed and improved open source ML library with a very active user community.

2.3.1.3 TensorFlow [23]

Developed by Google Brain team of engineers and researchers, TensorFlow is a scalable multi-platform programming interface. The fast training of ML models using this tool is due to the utilisation of both the Central Processing Unit (CPUs) and Graphics Processing Units (GPUs). This programming interface uses application programming interfaces (APIs) to support a number of programming languages mainly C++ and Python.

2.3.2 Graphical User Interface-Based Tool

In addition to programming-based tools, there are a number of GUI-based tools that provide ease-of-use to the state-of-the-art ML algorithms. A number of the common GUI-based ML tools are presented in the following sections.

2.3.2.1 Weka [28]

First introduced in 1992, the Waikato Environment for Knowledge Analysis (WEKA) is regarded as a unified workbench running ML algorithms with wide acceptance from academia and the industry. WEKA was not only as a ML toolbox, but also as a framework aimed for researchers to develop new algorithms without being restricted by a supporting infrastructure for data manipulation and scheme evaluation.

2.3.2.2 Orange [29]

Introduced in the late 1990s, Orange is an open source software and one of the oldest ML and data mining tools. It can use both Python scripting and visual programming through its GUI interface, offering a learning flexibility to both experienced and unexperienced users.

2.3.2.3 RapidMiner

First introduced in 2001 at the University of Dortmund under the name YALE (Yet Another Learning Environment). The name was changed more than once before becoming known as RapidMiner. RapidMiner is a widely-spread open source tool for ML, statistical methods, and data mining [30]. It is worth noting that Gartner placed this software for the sixth year in the leader quadrant of its Magic Quadrant for Data Science & Machine Learning Platforms [31].

2.4 ML as a tool in Network Design

Given the predictive abilities of ML algorithms, the availability of computing resources, and data, recent years have witnessed a rise in the number of papers employing ML to perform inference from historical data or to propose a certain strategies that reduce a penalty. Consequently, there are a number of review papers in the literature that strive to cover the state-of-the-art approaches. The authors of [32] surveyed a number of review papers in the ML subfield of deep learning. Although it was limited to this subfield, it gave a clear view on the vastness of the deep learning field. Further, the authors in [33] presented a comprehensive survey of a number of ML-powered, self-organised cellular networks. They classified the surveyed papers according to the optimisation objectives. However, the survey was limited to 4G cellular networks. Surveying the learning problems in the topic of cognitive radio was addressed by the authors in [34]. Nevertheless, now it can be considered dated due to new developments and proposed solutions. A brief review by the authors in [35] surveyed the rudimentary concepts of ML and proposed their utilisation in a number of applications working in 5G networks. However, their work was limited in terms of the number of surveyed papers. In this chapter, we will focus on the literature where ML is employed as part of radio resource optimisation systems in a cellular network. Thus, committing to the most relevant part given the work in this thesis.

2.5 ML in Radio Resource Allocation

With the ever-increasing demand for bandwidth, the proliferation of wireless-connected devices, and the growing interference from other mobile users, resource allocation is becoming a more challenging problem. In this section, we present the role of ML in solving the resource allocation problem in wireless networks.

2.5.1 Supervised-Learning-based Approaches

2.5.1.1 Network Planning

The use of ML techniques in conjunction with optimisation methods in the wireless field is gaining a momentum. The authors in [36] used support vector machines (SVM) and genetic algorithms to develop a network planning tool. The metric they wanted to minimise is the number of physical resource blocks (PRB) per mega bit (Mb), PRB/Mb, which will allow serving users with the minimum amount of resources possible while maintaining the QoS. The authors reported that improving the metric (PRB/Mb) they used, caused the system to provide resources effectively in a way to ensure all outage users are recovered. The genetic algorithm served all the users at the 20th generation and was able to increase the resource efficiency as it evolved.

The same authors proposed in [37] the use of different ML algorithms (KNN, NN, SVM, and DT). Additionally, ensemble methods (Bagging and AdaBoost) are used for enhancing the learners' accuracy of prediction. The goal is to propose a network planning tool capable of predicting a specific QoS metric that associates the interest of the users with that of the operators (i.e., PRB/Mb). The proposed prediction assists in future dense deployments in wireless networks. Thus, radio measurements are employed to develop correlative statistical models predicting the QoS to improve QoS-based network planning.

2.5.1.2 Using historical decisions to reduce optimisation time

The authors in [38] proposed a solution to minimise the time consumed by traditional optimisation methods. The proposed solution employs a cloud-based ML framework to extract similarities from a huge number of historical scenarios. The optimal or near-optimal solutions for these scenarios are then searched offline and stored to function as a dataset for the supervised learning algorithm to work on. A feature vector comprised of measured data for a newly arriving scenario is then compared to the training data and the adopted solution is the one with the most amount of similarities. The authors used the proposed solution to allocate beams to users in a massive multi-input multi-output (MIMO) system using the KNN algorithm to compare to other methods which include Exhaustive Search and the low-complexity beam allocation (LBA) method. The proposed method was able to reach a solution in less time than the exhaustive reach. Further, it outperformed the LBA in terms of the average sum-rate as the size of the training set grew larger than 1000 instances.

2.5.1.3 Self-adaptive flexible transmission time interval

Taking the opportunity of standardisation of new numerology technologies and 5G new radio (NR), the authors in [39] proposed a self-adaptive flexible transmission time interval (TTI) scheduling strategy aimed at satisfying the service requirements in a scenario where both ultra-Reliable Low Latency Communications (uRLLC) and enhanced Mobile Broadband (eMBB) coexist. The proposed scheduling strategy is implemented using Random Forest based Ensemble where the TTI length is chosen for each service according to channel conditions and BS features. The proposed system was compared to the existing ML methods, namely, SVM, NN, and random forests (RF) where it showed better accuracy. The results reported a reduced packet loss and delay for the uRLLC services as the eMBB requirements are guaranteed.

2.5.1.4 5G uplink grant-free transmission

Signalling overhead caused by handshaking-based scheduling is one of challenges facing massive machine-type communication (mMTC). Grant-free access enables devices in the wireless network to transmit without waiting for the BS to grant them radio resources. Active user identification and channel estimation are required in grant-free uplink transmission. This is due to the fact that the receiver in a grant-free uplink transmission is oblivious to the channel information and the active user identification. The authors in [40] proposed to use asynchronous sparse Bayesian learning (ASBL) and SVM algorithms for channel estimation and active user identification/classification, respectively.

Performance evaluation was carried out using link-level simulation, the performance of both the channel estimation and active user identification was compared to other compressed sensing-based methods where it showed that the proposed receiver has a better detection performance and suitability for uplink grant-free asynchronous non-orthogonal multiple access (NOMA) transmission.

2.5.1.5 Traffic and flow control in LTE-A

To alleviate the allocation process, control the transmission of recently served application, and reduce the overall load, a cross-layer communication approach is implemented between the media access control (MAC) and the application layers in the downlink scheduling procedure of an LTE-A network. A solution presented by the authors in [41] constitutes an added stage where traffic is classified before a conventional scheduling procedure takes part. The classified traffic is either sent to the scheduling stage or rejected where it is prevented from transmission over a period of time. A KNN-based supervised machine learning algorithm is used towards that end where traffic is classified according to average bit rate and delay features. The simulation results show effective resource allocation for real time applications measured in terms of fairness, packet lose and delay. However, the proposed approach did not investigate the throughput.

2.5.1.6 Bandwidth reservation to reduce termination

Tackling the problem of connection drop during handoff, the authors in [42] proposed a self-adaptive bandwidth reservation scheme to reduce the termination probability. The proposed scheme employed a support vector machine algorithm to compute the amount of reserved bandwidth at the target cell. The SVM algorithm utilises existing data at the BS to predict the moving direction of a mobile terminal. The simulation results report a reduction in the call dropping probability and call blocking probability. The bandwidth utilisation, on the other hand, witnessed better performance under heavy traffic loads by other schemes, namely, the schemes without (NO) reservation (NR).

2.5.1.7 Optimising spectrum allocation, route selection, and peer discovery in vehicular networks

Content distribution in cooperative vehicular networks was investigated by the authors in [43]. The objective is to optimise spectrum allocation, route selection, and peer discovery from a delay perspective. To this end, they proposed to utilise big data generated by a geographic positioning system (GPS) and a geographic information system (GIS) to predict the vehicle trajectories using a combination of interacting multiple model (IMM) estimation with multi-Kalman filter (MKF). The optimisation part was formulated as a coalition formation game and was compared against two heuristic schemes; the non-cooperative content distribution scheme and the random group formation based content distribution scheme and the simulation results showed that the proposed approach achieved better performance.

2.5.2 Unsupervised-learning-based approaches

2.5.2.1 Traffic prediction in base stations

Aiming to predict data traffic volume at BSs, the authors in [44] proposed to use k-means algorithm to cluster the BSs into groups of geographically adjacent BSs with correlated traffic flows. Subsequently, the time series traffic data is pre-processed by decomposing them into high-frequency and low-frequency components using a wavelet decomposition method. Finally, and after reconstructing both frequency parts to time series components by wavelet reconstruction, an Elman neural network (ENN) is used on each of the time series components to predict the traffic flow.

2.5.2.2 User Clustering for downlink beams

The authors in [45] proposed to use K-means algorithm to group users in downlink 3G cellular systems adaptive cell Sectorisation. Thus, users are grouped according to their spatial characteristics into clusters using ML and used as a reference to shape antenna beams to enable the minimisation of specific features (e.g., interference, power).

2.5.2.3 Supporting cooperative spectrum sensing

The authors in [46] used ML techniques to propose cooperative spectrum sensing algorithms for cognitive radio networks. The goal is to determine (i.e., classify) the channel availability (i.e. class) by classifying an energy vector (i.e., feature vector) comprised of the energy levels reported by all secondary users. Supervised (i.e. SVM and KNN) and unsupervised algorithms (i.e. K-means and Gaussian Mixture Model) classification techniques are used for that purpose. The authors used the average training time, receiver operating characteristic performance, and sample delay classification to quantify the classifiers' performance for comparison purposes. Further, the authors compared their proposed schemes to the Fisher Linear Discriminant method. The results showed that SVM achieved the highest detection performance with the K-means following very closely in terms of the receiver operating characteristic (ROC) performance. The weighted KNN required the least amount of training time. Therefore, it is very suitable for channel spectrum sensing as it requires an on-the-fly update for its training vectors.

2.5.3 Reinforcement-learning-based approaches

2.5.3.1 Dynamic resource allocation in LTE-U networks

Solving the problem of LTE unlicensed (LTE-U) and Wi-Fi coexistence in the unlicensed spectrum was the focus of the authors in [47]. They proposed a scheme where blank subframes are dynamically allocated using a Q-learning algorithm. The number of subframes within a frame is kept but the subframe length is reduced signifying less transmission time and guaranteeing that a percentage of the subframes are blank subframes. The authors proposed sharing the transmission-related information so that the LTE-U decides when to allocate blank subframes, and when to allocate dynamically adjusted blank subframe numbers proportional to the Wi-Fi traffic size. The results showed that the proposed approach improved the overall system spectrum utilisation.

2.5.3.2 Spectrum Monitoring for cloud-based RAN

A reinforcement learning method was employed by the authors in [48] to propose an approach for faster dynamic spectrum allocation decisions in a cloud-based RAN (C-RAN). The proposed system uses regression analysis to operate on big data collected by a monitoring system at Sofia Airport to predict spectrum occupancy and usage activity in a predefined frequency band. The authors introduced a frequency-time resource indicator to act as a measure for spectrum usage. The authors reported that the prediction accuracy of their system is proportional the amount of collected data and outlined the "accuracy vs latency" trade-off problem solvable through the use of cloud-based generic processing architecture.

2.5.3.3 Increasing throughput and fairness for users in HetNets

The work in [49] considered semi- and uncoordinated deployment of small cells and proposed combining Q-learning with mobile users' geographical locations. To improve the dynamic allocation of radio resources, a game theoretical dimension is added by attributing roles relative to the interference at the BSs with the objective of enabling cells to cooperate even when indirectly communicating to each other. The results reported that combining user locations and Q-learning resulted in an increase in cell throughput while maintaining an acceptable user throughput. A further improvement in terms of system performance and fairness among users with an increase in the average cell throughput can be attained when incorporating the game theoretical approach.

2.5.3.4 Energy-efficient resource management in HetNets

A HetNet architecture was presented by the authors in [50]. Combining radio frequency (providing wide coverage area) and visible light communication (providing high data rate) to guarantee different QoS requirements. The joint uplink/downlink energy-efficient resource management decision making problem was formulated as a Markov decision process. The objective was to maximise the network energy efficiency while ensuring that the QoS requirements are met for Industrial-IoT or IoT devices. The proposed architecture is to function in an industrial IoT network setting where Ultra-Reliable Low-Latency performance is required. A reinforcement learning method was proposed by the authors to attain an optimal policy for resource management, named post-decision state (PDS) based experience replay and transfer (PDS-ERT). The simulation results showed that better performance can be attained through the proposed approach comparing it to Deep PDS and Q-learning algorithm with knowledge transfer (QKT)-learning algorithms.

2.5.3.5 Distributed resource allocation in asynchronous networks

The problem of allocating resources in downlink LTE-licence assisted access (LAA) network is tackled in [51]. Assuming limited channel state information (CSI) exchange, the objective is to maximise the proportional fairness of the users summed rate. Using a reinforcement-based approach through a fully connected neural network where random seeds are employed, the learning target is the seed with the highest in-cell proportional fairness. The simulation results reported that the in-cell proportional fairness contributed to the maximisation of the overall proportional fairness. Comparing the proposed algorithm to another (fairness allocation scheme) showed that the proposed approach attained 6.8% higher geometric mean.

2.5.3.6 Spectrum auction in cognitive radio networks

Using spectrum sensing to detect available frequency bands, the authors in [52] proposed to employ Q-learning-based bidding algorithm for spectrum auction by the secondary users to allocate them the available bands. The algorithm enables secondary users to learn from the competitors so that they can automatically place better bids for the available frequency bands. Secondary users who win multiple bids can utilise multiple bands per time slot. Hence, they can send their data using multiple frequency bands simultaneously in one time slot. The results show that the proposed approach managed to allocate the frequency bands efficiently, automatically and in a fair manner.

2.5.3.7 Circumstance independent policy for resource allocation

It is worth stressing the fact that network circumstances (e.g., number of users and QoS requirements) are generally key for reinforcement learning policy structure. Thus the policies are circumstance-dependant and this can hinder the policy implementation in practical systems. The authors in [53] proposed the use of a circumstance-independent policy for resource allocation in wireless networks to function on different network circumstances and developed a deep reinforcement learning algorithm to learn it. The proposed approach can be applied in practical systems over different circumstances. The proposed policy was compared against the circumstance-depended policy where it attained close performance for each circumstance.

2.5.3.8 Spectrum sharing and spatial reuse

To address the problem of underutilised spectrum in the millimetre-wave band, ultra-dense networks, the authors in [54] proposed a generalised temporal-spatial spectrum sharing scheme, establishing a dynamic spectrum sharing model where the same channel is utilised by several shared links at the same slot. A non-cooperative game between devices is formulated as the spectrum utilisation problem, and is proven to be an ordinary potential game. Thus, a Nash Equilibrium (NE) is guaranteed. A novel decentralised Q-learning is used to help the secondary users learn the environment and adapt to achieve NE with partial feedback information and by depending on action-reward history. The action, and reward of each secondary user are channel selection and channel capacity, respectively. The new Q-learning algorithm defines the actions over Q values instead of the legacy state-action pair. Thus, each action correlates with a Q value updated as the weighted sum of the current Q-value and the instant reward whilst the Q values of the other actions remain unchanged. The results of evaluating the proposed approach against other schemes showed a faster and more stable convergence. Further, an improvement to network throughput is witnessed promoting the increase of 5G-connected devices.

2.5.4 Novel approaches

2.5.4.1 Delay-Aware Brain-Centric Radio Resource Optimisation

The authors in [55] developed a framework to manage resources in wireless networks while considering the delay perception in the human brain. Based on the brain features, a probabilistic model is developed using a probability distribution identification (PDI) learning method to predict the delay perceived by the human users and quantify the reliability of this prediction. They defined a closed-form expression that identifies the relationship linking wireless physical layer metrics and system reliability. Using the aforementioned relationship and a developed learning method named PDI consisting of two, supervised and unsupervised learning parts, the authors proposed a Lyapunov-based brain-aware optimisation approach to allocate human users with radio resources. The results show that the brain-aware approach yielded up to 78% power savings when compared to another system that considered the QoS metrics exclusively.

2.6 Chapter Summary

This chapter provided a review of ML types and algorithms. Further, it shed light on the role played by ML in the design of current and future cellular networks. Given the chance to summarise the literature, we highlight the fact that the majority of the paper reviewed fall within the supervised and reinforcement learning types. Unsupervised learning had a smaller share in the literature due to the topic in hand (i.e., radio resource allocation). Thus, the use of labelled data and the employment of an agent to discover the environment are the most common features in the optimisation process. Nonetheless, we highlighted several unsupervised learning use cases. An extension to this survey can include all other wireless network design and optimisation aspects that incorporates ML algorithms.

Chapter 3

Literature Review – Using Big Data Analytics in Network Design

3.1 Background

Networks generate traffic in rapid, large, and diverse ways, which leads to an estimate of 2.5 exabytes created per day [56]. There are many contributors to the increasing size of the data. For instance, scientific experiments can generate lots of data, such as CERN's Large Hadron Collider (LHC) that generates over 40 petabytes each year [57]. Social media also has its share, with over 1 billion users, spending an average 2.5 hours daily, liking, tweeting, posting, and sharing their interests on Facebook and Twitter [58]. It is without a doubt that using this activity-generated data can affect many aspects, such as intelligence, e-commerce, biomedical, and data communication network design. However, harnessing the powers of this data is not an easy task. To accommodate the data explosion, data centres are being built with massive storage and processing capabilities, an example of which is the National Security Agency (NSA) Utah data centre that can store up to 1 yottabyte of data [59], and with a processing power that exceeds 100 petaflops [60]. Due to the increased needs to scale-up databases to data volumes that exceeded processing and/or storage capabilities of simple computer systems, systems that ran on computer clusters started to emerge. Perhaps the first milestone took place in June 1986 when Teradata [61] used the first parallel database system (hardware and software), with one terabyte storage capacity, in Kmart data warehouse to have all their business data saved and available for relational queries and business analysis [62, 63]. Other examples include the Gamma system of the University of Wisconsin [64] and the GRACE system of the University of Tokyo [65].

In light of the above, the term “Big Data” emerged, and it can be defined as high-volume, high-velocity, and high-variety data that provides substantial opportunities for cost-effective decision-making and enhanced insight through advanced processing which extracts information and knowledge from data [66]. Another way to define big data is by saying it is the amount of data that is beyond traditional technology capabilities to store, manage, and process in an efficient and easy way [67]. Big data is already being employed by digital-born companies like Google and Amazon to help these companies with data-driven decisions [68]. It also helps in the development of smart cities and campuses [69], as well as in other fields like agriculture, healthcare, finance [70], and transportation [71].

3.2 Big Data Characteristics

Big data is better defined through its characteristics, which are:

Volume: This is a representation of the data size [72].

Variety: Generating data from a variety of sources results in a range of data types. These data types can be structured (e.g. e-mails), semi-structured (e.g. log files data from a webpage); and unstructured (e.g. customer feedback), and hybrid data [73].

Velocity: Is an indication of the speed of the data when being generated, streamed, and aggregated [74]. It can also refer to the speed at which the data has to be analysed to maintain relevance [72].

Depending on the research area and the problem space, other terms or Vs can be added. For example, is this data of any value? How long can we consider this an accurate and valid data? Since we are conducting a survey, we find it compelling to briefly introduce other Vs as well. Typically, the number of analysed Vs is 3 to 7 in a single study (e.g. 6V+C [75]), where C represents Complexity, however, different papers analyse different sets of Vs and the union (sum) of all the analysed Vs among all surveyed papers is 8V and a C, as shown in Table 3-1.

Value: Is a measure of data usefulness when it comes to decision making [74], or how much added-value is brought by the collected data to the intended process, activity, or predictive analysis/hypothesis [76].

Veracity: Refers to the authenticity and trustworthiness of the collected data against unauthorised access and manipulation [76, 77].

Volatility: An indication of the period in which the data can still be regarded as valid and for how long that data should be kept and stored [78].

Validity: This might appear similar to veracity; however, the difference is that validity deals with data accuracy and correctness regarding the intended usage. Thus, certain data might be valid for an application but invalid for another.

Variability: This refers to the inconsistency of the data. This is due to the high number of distributed autonomous data sources [79]. Other researchers refer to the variability as the consistency of the data over time [77].

Complexity: A measure of the degree of interdependence and inter-connectedness in big data [75]. Such that, a system may witness a (substantial, low, or no) effect due to a very small change(s) that ripples across the system [74]. Also, complexity can be considered in terms of relationship, correlation and connectivity of data. It can further manifest in terms of multiple data linkages, and hierarchies. Complexity and its mentioned attributes can however help better organise big data. It should be noted that complexity was included among the big data attributes (Vs) in [75] where big data was characterised as having 6V + complexity. This is how we will arrange it in **Table 3-1**.

Table 3-1: Various big data dimensions

No. of Vs	References	Dimensions (Characteristics)								
		Volume	Velocity	Variety	Veracity	Value	Variability	Volatility	Validity	Complexity
3Vs	[80-86]	√	√	√						
4Vs	[59, 87-89]	√	√	√	√					
	[90-94]	√	√	√		√				
5Vs	[58, 66, 76, 95, 96]	√	√	√	√	√				
6Vs	[75, 77, 79, 97]	√	√	√	√	√	√			√
7Vs	[78, 98]	√	√	√	√	√		√	√	

3.3 BDA tools

The process of extracting hidden, valuable patterns, and useful information from big data is called *Big Data Analytics* [99]. This is done through applying advanced analytics techniques on large data sets [83]. Before commencing the analytics process, data sets may comprise certain consistency and redundancy problems affecting their quality. These problems arise due to the diverse sources from which the data originated. *Data pre-processing* techniques are used to address these problems. The techniques include integration, cleansing (or cleaning), and redundancy elimination, and they were discussed by the authors in [94].

BDA can be carried out using a number of frameworks (shown below) that usually require an upgradeable cluster dedicated solely for that purpose [72]. Even if the cluster can be formed using a number of commodity servers [100], however, this still forms an impediment for limited-budget users who want to analyse their data. The solution is presented through the democratisation of computing. This made it possible for any-sized company and business owners to analyse their data using cloud computing platforms for BDA. Consequently, the use of BDA is not limited to enterprise-level companies. Furthermore, business owners do not have to heavily invest in an expensive hardware dedicated to analysing their data [56]. Amazon is one of the companies that provide 'cloud-computed' BDA for its customers. The service is called Amazon EMR (Elastic MapReduce), and it enables users to process their data in the cloud with a considerably lower cost in a pay-as-you-use fashion. The user is able to shrink or expand the size of the computing clusters to control the data volume handled and response time [56, 101]

Dealing with big amounts of data is not an easy task, especially if there is a certain goal in mind since data arrives in a fast manner, it is vital to provide fast collection, sorting, and processing speeds. Apache Hadoop was created by Doug Cutting [102] for this purpose. It was later adopted, developed, and released by Yahoo [103]. Apache Hadoop can be defined as a top-level, java-written, open source framework. It utilises clusters of commodity hardware [104].

Hadoop V1.x (shown in Figure 3-1) consists of two parts: the Hadoop Distributed File System (HDFS) that consists of a storage part, and a data processing and management (MapReduce) part. The master node has two processes, *a Job Tracker* that manages the processing tasks and a *Name Node* that manages the storage tasks [105].

When a Job Tracker takes job requests, it splits the accepted job into tasks and pushes them to the *Task Trackers* located in the slave nodes [106]. The *Name Node* resembles the master part, while the *Data Nodes* represent the slave part [67].

Many projects were developed in a quest to either complement or replace the above parts, and not all projects are hosted by the Apache Software Foundation, which is the reason for the emergence of the term *Hadoop ecosystem* [102].

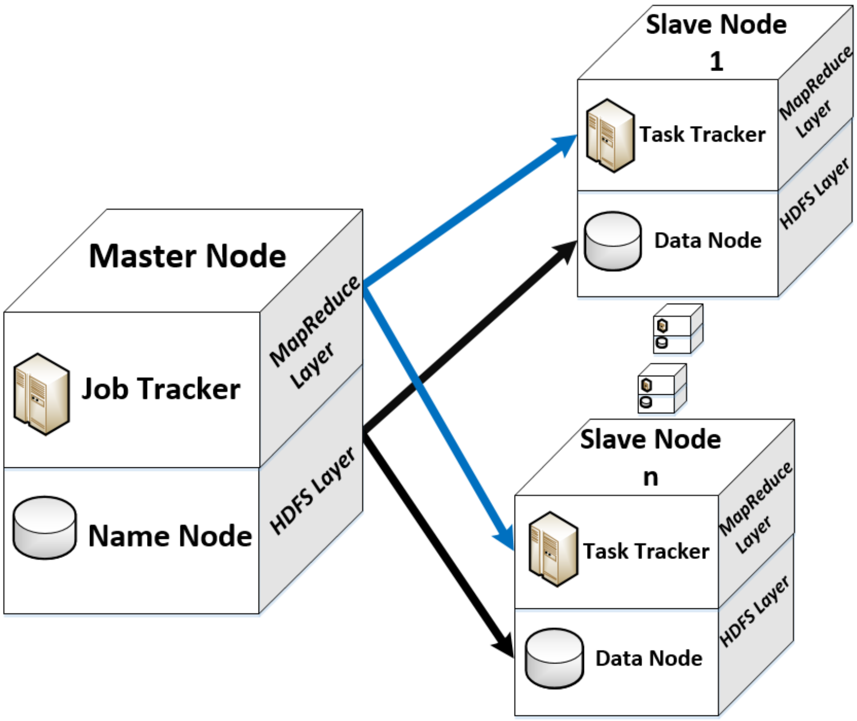


Figure 3-1: Hadoop V1.x architecture

Hadoop V2.x is viewed as a three-layered model. These layers are classified as storage, processing, and management, as shown in Figure 3-2**.** The current Hadoop project has four components (modules), which are MapReduce, the HDFS, Yet Another Resource Negotiator (YARN), and Common utilities [72].

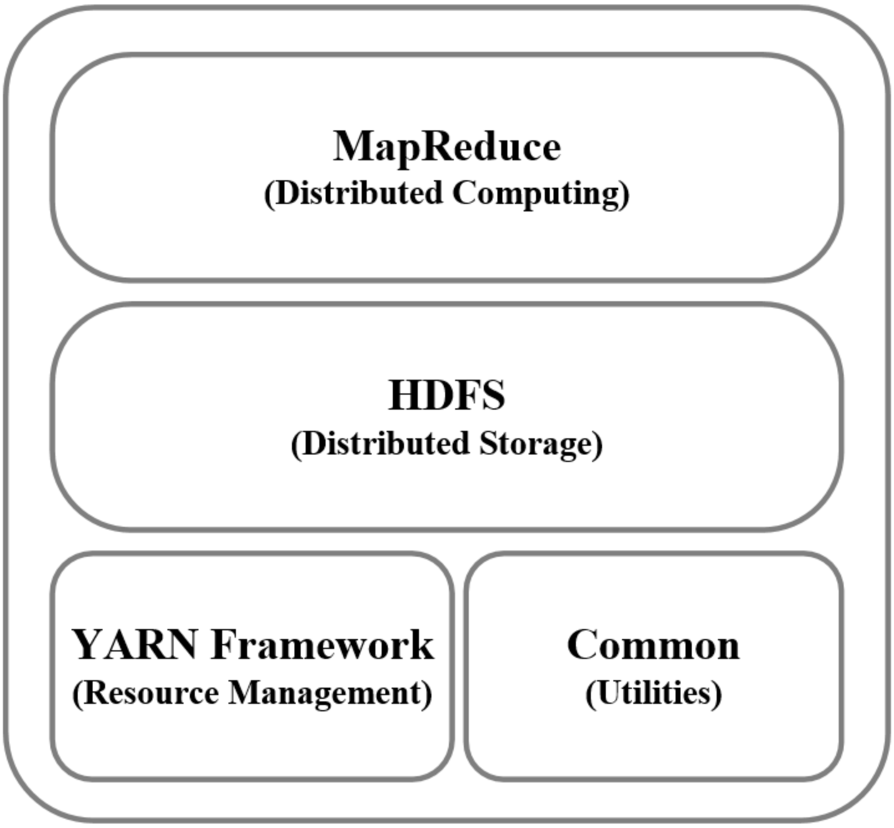


Figure 3-2: Hadoop V2.x architecture

MapReduce: As a programming model, MapReduce is used as a data processing engine and for cluster resource management. With the emergence of Hadoop v2.0, the resource management task became YARN's responsibility [72]. WordCount is an example illustrating how MapReduce works. As the name implies, it calculates the number of times a specific word is repeated within a document. Tuples $\langle w, 1\rangle$ are produced by the map function, where w and 1 represents the word and the times it appeared in the document respectively. The reduce function groups the tuples that share the same word and sums their occurrences to reach the concluding result [61].

HDFS: HDFS represents the storage file-system component in the Hadoop ecosystem. Its main feature is to store huge amounts of data over multiple nodes and stream those data sets to user applications at high bandwidth. Large files are split into smaller 128 MB blocks, with three copies of each block of data to achieve fault tolerance in case of disk failure [72, 107, 108].

YARN: YARN was introduced in Hadoop version 2.0, and it simply took over the tasks of cluster resource management from MapReduce and separated it from the programming model, thus making a more generalised Hadoop capable of selecting programming models, like Spark [109], Storm [110], and Dryad [111, 112].

Common utilities: To operate Hadoop's sub-projects or modules, a set of common utilities or components are needed. Shared libraries support operations like error detection, Java implementation for compression codes, and I/O utilities [72, 113].

Over the last few years, researchers in telecommunication networks started to consider BDA in their design toolbox. Characterised by hundreds of tuneable parameters, wireless network design informed by BDA received most of the attention, however, other types of networks received increasing attention as well.

The vast amount of data that can be collected from the networks, along with the distributed modern high-performance computing platforms, can lead to new cost-effective design space (e.g. reducing total cost of ownership by employing dynamic Virtual Network Topology adaptation) when compared to classical approaches (i.e. static Virtual Network Topologies) [114]. This new paradigm is promising to convert networks from being sightless tubes for data into insightful context-aware networks.

3.4 Case studies of the use of BDA for wireless and wired networks

3.4.1 Detection of Sleeping Cells in 5G SON

A wireless cell may cease to provide service with no alarm triggered at the Operation and Maintenance Centre (OMC) side. Such cells are referred to as sleeping cells in self organising networks (SON). The authors in [115] tackled this problem and presented a case study on the identification of the sleeping cells (SC). The simulation scenario comprised of 27 macro sites each with three sectors. The user equipment (UE) is configured to send radio measurement and cell identification data of the serving and neighbouring cells to the BS, in addition to event-based measurements. The simulation considered two scenarios; reference (a normally-operating network) and SC. The latter was simulated by dropping the antenna gain from 15 dBi (reference scenario) to -50 dBi (SC scenario). The reference scenario provided measurements used by an anomaly detection model based on KNN algorithm to provide a network model with normal behaviour. Multidimensional Scaling (MDS) is used to produce a minimalistic Key Performance Index (KPI) representation. Thus the interrelationship between Performance Indexes (PIs) is reflected and an embedded space is constructed. Consequently, similar measurements (i.e. normal network behaviour) lie within close distances while dissimilar measurements (i.e. anomalous network behaviour) are far-scattered and hence easily identified. The model attained 94 percent detection accuracy with 7 minutes training time.

3.4.2 An Architecture for Fully Automated MNO Reporting System.

Mobile Network Operators (MNOs) collect vast amounts of data from a number of sources as it can offer actionable plans in terms of service optimisation. Visibility and availability of information is vital for MNOs due to its role in decision making. Employing a reporting system is pivotal to transform data to information, knowledge, and lastly to actionable plans. The authors in [116] presented a case study for the potential role of BDA in the development a fully automated reporting system. A Moroccan MNO is to benefit from the alternative architecture. The authors highlighted the shortcomings of the existing automatic reporting system that uses traditional technologies. Moreover, they inferred that using BDA can provide the opportunity to overcome those shortcomings. The authors chose the Apache Flink [116] in their proposed architecture to serve as their BDA framework.

3.4.3 Network Anomaly Detection using NetFlow Data

BDA can support the efforts in the subject of network anomaly and intrusion detection. To that end, the authors in [117] proposed an unsupervised network anomaly detection method powered by Apache Spark cluster in Azure HDInsight. The proposed solution uses a network protocol called NetFlow that collects traffic information that can be utilised for the detection of network anomalies. The procedure starts by dividing the NetFlows data embedded in the raw data stream into 1-minute intervals. NetFlows are then aggregated according to the source IP, and data standardisation is carried out. Afterwards, a k-means algorithm is employed to cluster (according to normal or abnormal traffic behaviour) the aggregated NetFlows. The authors considered a dataset containing 4.75 hours of records captured from CTU University to analyse botnet traffic. The proposed approach attained 96% accuracy and the results were visualised in 3D after employing Principal Component Analysis (PCA) to attain dimension reduction.

3.5 Role of BDA in Cellular Network Design

In this section, we review the research done on the use of BDA for the design of cellular networks. Compared to other network design topics, we observed that the wireless field has received the highest attention, as measured by its share of research papers. These papers can be classified according to the application or area under investigation. Consequently, we have classified those papers into the following:

1- Counter-failure-related: This includes fault tolerance (i.e. detection and correction), prediction, and prevention techniques that use BDA in cellular networks.

2- Network monitoring: This illustrates how BDA can be beneficial as a large-scale tool for data traffic monitoring in cellular networks.

3- Cache-related: Investigates how BDA can be used for content delivery, cache node placement and distribution, location-specific content caching, and proactive caching.

4- Network optimisation: BDA can be involved in several topics including predictive wireless resource allocation, interference avoidance, optimising the network in light of Quality of Experience (QoE), and flexible network planning in light of consumption prediction.

It should be noted that Table 3-2 provides further detailed classification, with the chance to compare the role played by BDA across different network types and applications

3.5.1 Failure Prediction, Detection, Recovery, and Prevention

3.5.1.1 Inter-technology Failed Handover Analysis using BDA

One of the most frustrating encounters happens when a mobile subscriber gets surprised by a sudden call drop. Many of these incidents occur when the user is at the edge of a coverage area and moving towards another, technologically-different area, e.g., moving from a 3G BS to a 2G BS. The common solutions to address such shortcomings are by either conducting drive tests or performing network simulation. However, another solution that leverages the power of big data was proposed by the authors in [118]. The proposed solution uses BDA (Hadoop platform) to analyse the Base Station System Application Part (BSSAP) messages exchanged between the Base Station Subsystem (BSS) and Mobile Switching Centre (MSC) nodes. Location updates (only those involved in the inter-technology handover) are identified and the geographic locations where the 3G-service disconnections occur are identified by relying on the provided target Cell ID.

The results of the above method were then compared with a drive test (which is an expensive and time-consuming approach) results, where coherence between the two results was demonstrated. Another comparison was conducted with the Key Performance Index (KPI)-based approach and the results were in favour of the proposed approach.

3.5.1.2 Signalling Data-based Intelligent LTE Network Optimisation

By utilising the combination of all-around signalling and user and wireless environment data, combined with Self-Organised Network technologies (SON), full-scale automatic network optimisation could be realised.

The authors of [82] developed an intelligent cellular network optimisation platform based on signalling data. This system involves three main stages:

1- *Defining network performance indicators through the extraction of XDR keywords*: The External Data Representation (XDR) contains the key information of the signalling (e.g., the causes of the process failures and signalling types). The status of a complete signalling process can also be identified by the XDR (e.g., the success or failure of signalling establishment and release). A number of performance indicators are defined by relying on this information. Querying these indicators is possible from multiple dimensions and levels (e.g., user, cell, and grid level).

2- *Problem discovery:* Service establishment rate, the handover success rate, and drop rate are among the network signalling-plane statuses that can be reflected by the XDR-based network performance indicators. Network equipment with unsatisfactory performance indicators can be further analysed, and this can be done by conducting a further excavation of the corresponding indicators' original signalling.

3- *Providing best practice solutions*: Identified and solved problems can provide an optimisation experience. As a consequence, a variety of network problems can be verified. For example, when a cell has a low handover success rate, according to the definition of the associated indicators, the reason is suggested to be the low success rate of the handover preparation. The solution would be to adjust the overlapping coverage areas formed between the source and the target cells and the parameters (e.g., the decision threshold offset and the handover initiation).

A recommended solution can be provided when a deteriorating indicator surfaces, and this is simply done by clicking the index query that caused the deterioration

3.5.1.3 Anomaly Detection in Cellular Networks

When a certain problem occurs in the cellular network, the user would usually be the first who feels the service disruption and suffers the impact. An abnormal and disrupted service may be identified by examining the Call Detail Record (CDR) of the users in a specific area. CDR files are generated upon making a call, and include, among other information, the caller and called numbers, the call duration, the caller location, and the cell ID where the call was initiated or received.

A CDR based Anomaly Detection Method (CADM) was proposed by the authors in [119]. CADM was used to detect the anomalous behaviour of user movements in a cellular network. This was done, first, with the CDR data being collected from the network nodes and stored in a mediation department. Then, the second phase starts by distributing the collected CDRs to the relevant departments (e.g., data warehouse, billing, and charging departments). After that, the Hadoop platform is used to detect the anomalies. The discovered anomalies are then fed-back to the mediation department for adequate actions.

The use of BDA was essential in this case. Large datasets that require distributed processing across computer clusters were processed by the Hadoop Platform. The result was an improved system that is able to detect location-based anomalies and improve the cellular system's performance.

3.5.1.4 Self-healing in Cellular Networks

The idea of developing a system that is capable of monitoring itself, detecting the faults, performing diagnoses, issuing a compensation procedure, and conducting a recovery is very appealing. However, the self-healing process has another factor to keep in mind, which is time. The process should be carried out within a reasonable amount of time so it would not degrade the quality of the delivered services.

Three use cases were presented by the authors in [120] for a self-healing process in cellular networks:

1- *Data Reduction:* The Operation and Maintenance (O&M) database can be used for troubleshooting purposes. However, the database size is relatively large as it contains the data related to both normal and degraded intervals, which makes it difficult to process. Separating the intervals to just keep the degraded intervals will help in reducing that size. The authors proposed parallelising this process independently by analysing each BS separately.

They chose the degraded interval detection algorithm of [121] (a degraded interval is the time where the BS behaviour is degraded), and these intervals were detected by comparing the BS's KPIs to a certain threshold. This algorithm was parallelised by implementing it as a *map* function, a field is added to identify each BS, and all the fields are added by a *reduce* function.

2- *Detecting Sleeping Cells:* Cell outage or sleeping cells is a common problem in mobile networks. Users are directed to neighbouring cells instead of the nearest and optimal cell. According to the algorithm described in [122], sleeping cells can be detected through the utilisation of neighbouring BS measurements hence calculating the impact of the sleeping cell outage. The detection process relies on the Resource Output Period (ROP), where each BS produces Configuration Management (CM), Fault Management (FM), and Performance Management (PM) data every 15 minutes. For each BS, incoming handovers from neighbouring BSs are aggregated for the current and previous ROP. If the number of handovers suddenly dropped to zero, and a malfunction is indicated by the cell’s Performance Indicators (PIs), the cell is regarded as a sleeping cell.

The authors in [120] proposed the use of the above-mentioned algorithm under the big data principle. They proposed to divide the terrain into partitions that are the maximum distance between neighbours, where each BS within the partitioned area is sequentially tested by an instance of the algorithm, and this is done by examining the data of its neighbours.

This approach was compared to other methods (e.g., lack of KPIs and availability of KPIs), and most of the simulated outages were detected (5.9% false negatives and 0% false positives). While a lack of KPIs and availability of KPIs methodologies showed a high percentage of false negatives.

3- KPI Correlation-Based Diagnosis: The authors in [120] used a method that utilises most correlated KPIs to identify the problem cause. To simplify the analysis task, the algorithm considers the PIs of both the affected BS and the neighbouring sectors.

MapReduce was used to implement this algorithm in a parallelised manner, the correlation process and the creation of a PIs list arranged by correlation were implemented as map and reduce functions, respectively.

3.5.1.5 Cell-site Equipment Failure Prediction

A sudden outage of services might have serious consequences, and this is why keeping communication equipment, like cell sites, in a good working state is of high importance. The challenge identified by the authors in [123] is to analyse the user’s bandwidth on the cell level. Equipment(s) failure and infrastructure faults can be predicted by analysing the bandwidth trends in a particular cell.

Due to the size and diversity of the collected data, it is essential to use BDA to process it. Thus, the customers' received bandwidth can be acquired over a particular time period (i.e., month or year, etc.). Next the data from diverse data sources are integrated and then analysed to know the bandwidth trends.

3.5.2 Network Monitoring

3.5.2.1 Large-scale Cellular Network Traffic Monitoring and Analysis

Large cellular networks have relatively high data rate links and high requirements to meet. Usually these networks use a high-performance and large capacity server to perform traffic monitoring and analysis.

However, with the continuous expansion in data rates, data volumes, and the requirements for detailed analysis, this approach seems to have a limited scalability. Hence, the authors of [124] proposed a system to undertake that task, utilising the Hadoop MapReduce, HDFS, and HBase (a distributed storage system that manages the storage of structured data and stores them in a key/value pair) as an advanced distributed computing platform. They exploited its capability of dealing with large data volumes while operating on commodity hardware. The proposed system was deployed in the core side of a commercial cellular network, and it was capable of handling 4.2 TB of data per day supplied through 123 Gbps links with low cost and high performance.

3.5.2.2 Mobile Internet Big Data Operator

China Unicom, China's Largest WCDMA 3G mobile operator with 250 million subscribers in 2012, introduced an industry ecosystem. The researchers in [125] highlighted this as a telecom operator-centric ecosystem that is based on a big data platform.

The above-mentioned big data platform is developed for retrieving and analysing data generated by mobile Internet users. With the aim of optimising the storage, enhancing the performance, and accelerating the database transactions, the authors proposed a platform that uses HDFS for distributed storage. The cluster had 188 nodes used to store data, perform statistical data analyses, and act as management nodes. The approximate storage space was 1.9 PB. HBase has the role of a distributed database, with a writing rate that can reach 145k records per second; HBase stores the structured data located on the HDFS.

Compared with the Oracle database, it is noted that the system achieved four times lower insertion rate. The query rate was compared to an Oracle database as well, and the HBase showed better performance when taking into consideration the impact of the records' size.

3.5.3 Cache and Content Delivery

3.5.3.1 Optimised Bandwidth Allocation for Content Delivery

Mobile networks, usually, have a large number of users, and with the increase in Internet-based applications, it has become essential to allocate the required bandwidth that meets the user expectations, as well as to ensure a competitive level of service quality. Cellular networks can provide Internet connectivity to their users at any time; however, video (especially high quality) contents are still slow and relatively expensive. From the BS's point of view, the impact of forwarding the same video content to several users on the same BS is massive. The LTE system addressed this through multicast techniques. However, multicast is still regarded as a big challenge in cellular networks. To overcome the above problem, the authors of [81] proposed a solution that can dynamically allocate bandwidth. The idea is based on sharing the BS's wireless channel by a user cluster that wishes to download the contents. Thus, saving the BS resources, as well as providing a better data rate for the clustered users, and providing an opportunity for the users who did not join the cluster to benefit from the saved resources (bandwidth). It should be noted that the clustered users can receive the contents from the cluster head by using short range communication techniques like Wi-Fi Direct [126] and Device to Device (D2D) communication.

Two conditions must be satisfied before forming a user cluster. First, the users who request the same content are the ones who form the cluster. Second, the users should either be at the time (or will be) within a short range of each other. For that reason, the authors suggested using BDA to identify the users' closeness and to group the users into cluster(s). A cluster head is then selected among the nearby users, and the process is repeated among the BS users until there is either a cluster of users or a free (un-clustered) user(s). The simulation was carried for a single BS network and the results showed faster content delivery and improved throughput at the user level.

3.5.3.2 Improve Cache Node Determination, Allocation, and Distribution Accuracy in Cognitive Radio Networks

In cognitive radio networks, Secondary Users (SU) have to leave the licensed spectrum when their activity starts to affect the QoS level of the licensed users. This move would require the existence of a cache node to compensate for the interrupted data transactions during the SU switch to the unlicensed spectrum. The author of [127] proposed the use of BDA to process the data accumulated over time within the nodes. The goal was to utilise this data to reach a decision on the cache node distribution in a cluster network. The author selected two out of three categories (open and selectively open systems) of cognitive radio networks. For the selectively open systems, the SU selectively shares its information with either some cache nodes, with the cluster head for a particular time interval, or with specific SUs in a cluster. This results in a variable amount of shared data, thus resulting in variable accuracy.

3.5.3.3 Tracking and Caching Popular Data

The number of social network (i.e., Facebook and Twitter) users is massive. The multimedia contents of these networks are normally shared between common interest groups. However, big and important events attract a lot of attention and consequently a lot of content is shared across these networks. When a certain video or event goes viral, this sharing will eventually burden the network as the requested content would have to travel along the network on its way to the servers. A solution to such a problem was suggested by the authors of [123]. They suggested monitoring popular and social media websites, analysing the data, identifying if there is a growing interest in certain content, by which age category, and caching the popular data for a specific BS. BDA can be of major use in this situation by employing it to do the required analysis. The result would be cached content available to the users faster (reduced provisioning delay) and without burdening the network.

3.5.3.4 Proactive Cashing in 5G Networks

Cache-enabled BS can serve cellular subscribers. This is done by predicting the most strategic contents and storing them in their cache. Thus, minimising both the amount of time and the consumed network bandwidth, which can payoff in other ways (i.e., less congestion and less resource utilisation).

An approach, proposed by the authors in [89], used BDA and ML is to develop a proactive caching mechanism by predicting the popularity distribution of the content in 5G cellular networks. They demonstrated that this approach can achieve efficient utilisation of network resources (backhaul offloading) and an enhanced user experience. After collecting the raw data, i.e., the user traffic, the big data platform (Hadoop) has the task of predicting the user demands by extracting the useful information, like Location Area Code (LAC), Hyper Text Transfer Protocol (HTTP) request-Uniform Resource Identifier (URI), Tunnel Endpoint Identifier (TEID)-DATA, and TEID for control and data planes. Then using this information to evaluate the content popularity from the previously collected raw data. Experimentally testing this work on 16 BSs, as part of an operational cellular network, resulted in 100% request satisfaction and 98% backhaul offloading.

3.5.4 Network Optimisation

3.5.4.1 Big data-driven Mobile Network Optimisation Framework

When thinking about optimising a cellular network, it is important to collect as much information as possible. Large networks, as well as their users, generate a plethora of data, for which the use of BDA is vital to analyse the colossal amount of data generated.

The authors in [3] proposed a mobile network optimisation framework that is Big Data Driven (BDD). This framework includes several stages, starting from the collection of big data, managing storage, performing data analytics, and the last stage of the process is the network optimisation.

Three case studies were used to show that the proposed framework could be used for mobile network optimisation.

1- Managing resources in HetNets:

The Mobile Network Operators (MNOs) may use big data to provide real time and history analysis across users, mobile networks, and service providers. MNOs can benefit from BDD approaches in the operation and deployment of their network, and this can be done in several stages:

A) **Network Planning**: Due to a deficiency in the level of sufficient statistical data, evolved Node B (eNB) sites are not optimally optimised. BDA can help MNOs reach better decisions concerning the deployment of eNB in the mobile network. The authors in [3] suggested the use of the network and anonymous users' data (e.g., dynamic position information and other service features). Providing a relation between the data and their events can offer a better understanding of the traffic trends.

B) **Predictive Resource Allocation**: Resource requirements change depending on the density and usage patterns of mobile network subscribers. Predicting where and when mobile users are using the network can help in preparing for sudden significant traffic fluctuations. The authors in [3] suggested the use of BDA to examine behavioural and sentiment data from social networks and other sources and to predict the traffic in highly populated areas. Using the cloud RAN architecture [128], the right place at the right time can be served through the predictive resource allocation keeping a minimal service disruption.

C) **Interference Coordination**: HetNets with small cells can be used to conduct interference coordination among macro and small cells. This coordination has to be carried out in the time domain instead of the frequency domain. Schemes like the enhanced Inter-Cell Interference Coordination (eICIC) in LTE-Advanced [129] efficiently enable resource allocation among interfering cells. eICIC allows interference to be avoided between Macro cells evolved Node B (MeNB) and its neighbouring Small cell eNBs (SeNBs) by having data transmitted in isolated subframes called the Almost Blank Subframe (ABS). The ABS subframes carries minimum (and most essential) control information, transmitted with reduced power [129], so that the network operator can control the configuration of that subframe.

Many factors contribute to the determination of the ABS ratio of the macro cell to the small cell, such as the traffic load in a specific area, the service type, and so on. The deployment of BDD optimisation functions at the MeNB would enable them to collect and analyse eNB-originated raw big data and enable a quick response. As a result, the performance optimisation of each cell and the users can be fulfilled. Optimising ICIC parameters (e.g., ABS ratio) can be achieved by processing raw data in a periodic manner. Furthermore, the location and user traffic demands of multiple eNBs can be optimised, offering the deactivation of a SeNB due to elevated Signal-to-Interference-plus-Noise Ratio (SINR) to avoid the interference caused by a nearby SeNB that would also result in reducing the energy consumption.

2- Deployment of cache server in mobile CDN

Popular content (e.g., movies) can be delivered through a Content Delivery Network (CDN), which is a method that is considered efficient by many MNOs. Distributed cache servers should be located near the users to achieve a fast response as well as to reduce the delivery cost. In hierarchical CDN, it is vital to place cache servers in an optimal location. Due to the unique features that RAN has, it was the primary interest of the authors in [3].

It is expected that there will be an enhanced backhaul capability in 5G networks, and this would result in minimising the concerns related to the latency and traffic load of backhaul transmissions. Therefore, not all MeNBs would require a dedicated distributed cache server. In addition, a SeNB can have a distributed cache server. Optimal cache server placement depends on several factors, such as the features and load of traffic in a given area, as well as the cost of storage and streaming equipment. To help the MNOs decide where to deploy their cache servers, data analytics methods can be regarded as a feasible solution. However, this would require the collection of all the above-mentioned factors over a long period in the related coverage area.

3- QoE modelling for the support of network optimisation:

The authors of [3] believed that the management of services and applications needed more than just relying on the QoS parameters. Instead, they suggested taking the quality (i.e., QoE), as perceived by the end users, to be regarded as the optimisation objective. Accurate and automatic real time QoE estimation is important to realise the optimisation objective. In addition to the technical factors, non-technical factors (e.g., user emotions, habit, and expectations, etc.) can affect the QoE. A profile for each particular user is composed comprising the above non-technical factors. This can be assembled by installing a profile collection engine on the users' mobile devices. User activities are compared and tracked to recognise differences and similarities, and then stored in a database for additional processing. After profiling, the following step constitutes the use of ML to identify the relationship between QoE and the influencing factors. Data analytics can be used to discover what impacts the QoE in users' devices, as well as the services and network resources. The next step is for network optimisation functions to react to determining what caused the problem and select the optimal action accordingly.

3.5.4.2 Improve QoS in Cellular Networks through Self-configured Cells and Self-optimised Handover

Cellular networks have a crucial element on which the concept of mobility depends. This element is the handover success rate, which ensures call continuity while the user moves from one cell to another. Failing in that particular element would impact the quality of the service, thus putting the operator into a questionable situation. Operators try to make sure that each cell has a list of manually configured and optimised neighbour cells. Hence, it is vital to note the high probability of these cells failing to adapt when a rapid response is required due to a sudden network change. The authors in [130] presented two methods that used BDA to introduce a self-configured and self-optimised handover process, the first was associated with newly introduced cells, while the latter was concerned with the already existing cells. The analysis started by collecting and archiving predefined handover KPIs. A dispatcher process is run after the collection period, and its aim is to check the files to see if they were marked as new cells (where Self-Configuration Analytics is started) or not (where Self-Optimisation Analytics is started):

1. NCL self-configuration for new cells:

Newly installed BSs require Neighbour Cell List (NCL) to be configured on the new cells. The selection process takes into consideration the antenna type, the azimuth angle (for directional cells), the geographic location of the candidate cells, and the process concludes by selecting cells with a minimum distance and maximum traffic load to be the top candidate cells. The NCL is configured via Network Management System (NMS) Configuration Management (CM) tools.

2. NCL self-optimisation for existing cells:

The process starts by collecting KPI measurement statistics for the failed and successful handovers, and this task is done by the Performance Management (PM) or the NMS. Cells with a handover failure rate below a predefined threshold are excluded from the NCL, while unlisted neighbouring cells with a successful rate above a predefined threshold are considered as new neighbouring cells.

3.5.4.3 Optimising the Resource Allocation in LTE-A/5G Networks

The overall system performance evaluation in advanced wireless systems, like LTE, depends on KPIs. In a quest to enhance the user experience, the authors of [2] proposed an approach that utilises user and network data, such as configuration and log files, alarms, and database entries/updates. This approach relies on the use of BDA to process the above-mentioned data. The ultimate goal is to provide an optimal solution to the problem of allocating radio resources to RAN users and guarantee a minimal latency between requesting the resource and allocating it. This is done through user and network behaviour identification, which is a task well-matched for BDA.

The proposed framework involves three stages:

First stage: This process is carried out in the eNB system, processing the data from the cellular and core network side. Binary values are acquired by comparing cellular level KPIs to their respective threshold values, thus keeping the binary matrix updated. This procedure is repeated at fixed intervals.

Second stage: Repeat the same steps as in the first stage. However, this process is carried out on subscriber level data to acquire subscriber KPI, and maintain a binary matrix.

Third stage: This is activated when a user initiates a resource allocation request. A binary pattern is generated based on the user requirements. This pattern is later handed over to stage 2 to update the binary matrix (if required) and incorporate the new values in the row that represents the requested bandwidth. After generating the updated row, it is transferred to the first stage for comparisons with the current Physical Resource Block (PRB) groups. To identify which PRBs suit the user, the fuzzy binary pattern-matching algorithm [131] was used for that purpose. Using this algorithm, the execution time increased linearly for an exponential increase in the number of comparison patterns.

3.5.4.4 Framework Development for Big Data Empowered SON for 5G

The authors of [115] proposed a framework called Big data empowered SON (BSON) for 5G cellular networks. Developing an end-to-end network visibility is the core idea of BSON. This is realised by employing appropriate ML tools to obtain intelligence from big data.

According to the authors, what makes BSON distinct from SON are three main features:

- Having complete intelligence on the status of the current network.
- Having the ability to predict user behaviour.
- Having the ability to link network response and network parameters.

The proposed framework contains operational and functional blocks, and it involves the following steps:

1- *Data Gathering*: An aggregate data set is formed from all the information sources in the network (e.g., subscriber, cell, and core network levels).

2- *Data Transformation*: This involves transforming the big data to the right data. This process has several steps, starting from:

   a. *Classifying* the data according to key Operational and Business Objectives (OBO), such as accessibility, retainability, integrity, mobility, and business intelligence.
   b. *Unify/diffuse* stage, and the result of this stage is more significant KPIs, which are obtained by unifying multiple Performance Indicators (PIs).
   c. According to the KPI impact on each OBO, the KPIs are *ranked*.
   d. *Filtration* is performed on the KPIs impacting the OBO less than a pre-defined threshold.

e. *Relate*, for each KPI and find the Network Parameter (NP) that affects it.

f. *Order* the associated NP for each KPI according to their association strength.

g. Cross-correlate each NP by finding a vector that quantifies its association with each KPI.

3- *Model*: Learn from the right data acquired in step 2 that will contribute to the development of a network behaviour model.

4- *Run SON engine*: New NPs are determined and new KPIs are identified using the SON engine on the model.

5- *Validate*: If a new NP can be evaluated by expert knowledge or previous operator-experience, proceed with the changes. Otherwise, the network simulated behaviour for new NPs is determined. If the simulated behaviour tallies with the KPIs, proceed with the new NPs.

6- *Relearn/improve*: If the validation in step 5 was unsuccessful, feedback to the *concept drift* [132] block, which will update the behaviour model. To maintain model accuracy, concept drift can be triggered periodically even if there was a positive outcome in the validation step.

3.5.4.5 Network Flexibility using Consumption Prediction

Consumption analysis is concerned with two factors: customer locations and type of service. Consumption trends can be classified in a timely manner into long-term, seasonal, and short-term. To reach an accurate prediction, the authors in [123] implied that user data (e.g., GPS location and service usage) can be correlated with other data (e.g., news, social network, events, and weather conditions). Using BDA to analyse these correlations, operators would be able to decide when and where to place their nodes without affecting the subscribers' satisfaction.

Finally, a summary for the surveyed research topics is depicted in Table 3-2.

Table 3-2: Research summary

Network Type	Research Category	Ref.	Proposed or Deployed Technique
Wireless	Failure Prediction, Detection, Recovery, and Prevention	[118]	Analysed inter-technology (2G-3G) failed handovers.
		[82]	Used XDR data to discover network failures and present a solution advice.
		[119]	Developed CADM which uses CDRs to identify anomalous sites.
		[120]	Presented three case studies of self-healing using BDA.
		[123]	Suggested the analysis of the bandwidth trends to predict equipment failure.
	Network Monitoring	[124]	Developed a Hadoop-based system to monitor and analyse network traffic.
		[125]	Developed a solution powered by big data platforms with distributed storage and distributed database to solve the issues of data analysis and acquisition.
	Cache and Content Delivery	[81]	Utilised big data to form a cluster made up of nearby users that share the BS's wireless channel.
		[127]	Analysed the data that resides within the cache nodes to enhance the determination, allocation, and distribution of cache nodes.
		[123]	Suggested monitoring and analysing social media and popular sites, to predict and cache certain contents, according to age category, at the predicted locations where these contents are highly

			demanded.
		[89]	Proposed the use of BDA and ML techniques to proactively cache popular content in 5G networks.
	Network Optimisation	[3]	Presented three case studies in which a proposed network optimisation framework is efficiently utilised. In particular, the work suggested: 1) The use of BDA to manage resources in HetNets. This is done in three stages (network planning, resource allocation, and interference coordination). 2) The deployment of cache servers in mobile CDN. 3) The optimisation of networks with QoE in mind.
		[130]	Proposed NCL self-configuration/optimisation algorithms to achieve an automatic, self-optimised handover. The work relied on the processing of CM and PM KPIs using BDA platform.
		[2]	Developed a three-stage framework that utilises the network and user KPIs to reach an optimal allocation of radio resources (PRBs).
		[115]	Presented a framework that uses big data collected from the cellular network to empower SON. They also presented a case study on how to detect sleeping cells using this framework.
		[123]	Correlated location data, service usage, and other contextual data to predict the consumption trends and select the optimal node location.

3.6 BDA in the Industry

There are several companies that offer network solutions based on BDA. These companies and solutions are highlighted in Table 3-3. It should be noted that these solutions are enabled by research conducted in their corresponding areas. We have added academic research papers related to each solution in Table 3-3.

Due to the proprietary nature of industrial products, the exact algorithms or methods behind these products is not available in the open literature. Therefore, academic papers with related concept(s) are cited. *NetReflex IP* and *NetReflex MPLS* utilises BDA [3, 82, 133] to provide services like anomaly analysis and traffic analysis. Nokia provided several solutions targeting the wireless field. For example, *Traffica* introduces itself as a real-time traffic monitoring tool that analyses user behaviour to gain network insights, similar approaches were presented in academia by the authors of [124, 134]. The *Wireless Network Guardian* detects user anomalies in mobile networks where a comparable topic was discussed in [135]. *Preventive Complaint Analysis* makes use of BDA to detect behavioural anomalies in mobile network elements where the authors in [136] provided a similar approach. *Predictive Care* utilises BDA to identify anomalies in network elements before affecting the user, a comparable academic approach is presented in [135, 137]. HP presented *Vertica*, a solution that exploits CDRs for network planning, optimisation, and fault prediction purposes.

The authors in [119, 138] researched akin approaches. Amdoc's *Deep Network*

Analyzer provides predictive maintenance and proactive network deployment for cellular networks. The authors in [139] presented a similar approach. Log analytics can be used for a variety of purposes. Aprevi's *ARLAS* solution provided real-time collection and storage of network logs. Related academic research was presented by the authors in [140-142].

Examining the above solutions, one can note that the majority of the solutions are in the wireless field. This, in fact, coincides with the orientation of the academically-researched topics. Sampling through the offered solutions, we noticed the increased interest in anomaly prediction and network node deployment.

Thus, offering the customer a service that is as close to optimal as possible, while minimising network expansion expenditures.

Table 3-3: BDA-powered industrial solutions

No.	Manufacturer	Solution Name	Related Academic Papers	Usage, Functions and Capabilities
1	Juniper	NetReflex IP	[3, 82, 133]	Eliminates network errors.
				Monitors QoS/QoE.
				Capacity planning, traffic routing, caching, and other optimisations.
		NetReflex MPLS		Segment and trend MPLS and VPN usage to plan for congestion.
				Identifies traffic utilisation and trends to optimise operational cost.
				Ability to slice network performance according to VPN, Cost of Service (CoS), and Provider Edge (PE)-PE enabling more efficient planning.
2	Nokia	Traffica	[124, 134]	Real-time issues detection and network troubleshooting.
				Gain real-time, end-to-end insight on traffic, network, devices, and subscribers.
		Wireless Network Guardian	[135]	Improves end-to-end network analytics and reporting with real-time subscriber-level information.
				Detects anomalies and reports airtime, signalling, and bandwidth resource consumption.
				Proactive detection of issues, including automatic detection of user anomalies and low QoE score alerts.
		Preventive Complaint	[136]	Detects network elements' behaviour anomalies.
				Predicting where customer complaints might arise and prioritises network

		Analysis		optimisation accordingly.
		Predictive Care	[135, 137]	Used for network elements, and proved its effectiveness by helping Shanghai Mobile become more agile and responsive.
				Accuracy of the simplified alerts is around 98 percent, reducing operational workload.
3	HP (HPE)	Vertica	[119, 138]	Provides CDR analysis that can help Communication Service Provides (CSPs).
				Examines dropped call records and other maintenance data to determine where to invest in infrastructure.
				Failure prediction and proactive maintenance.
4	Amdocs	Deep Network Analytics	[139]	Combines RAN information with BSS and customer data to deploy the network proactively.
				Predictive maintenance.
5	Apervi	Apervi's Real-time Log Analytics Solution (ARLAS)	[140-142]	Collects, aggregates, and stores log data in real-time.

3.7 BDA-powered Design Cycle and Challenges

In this section, we highlight a common theme among most of the surveyed papers. This can be summarised as depicted in Figure 3-3. Also, we illustrate the challenges facing the implementation of BDA in network design and operation.

3.7.1 BDA Design Cycle

The quest for a well-designed communication network is never-ending. Researchers in the big data era rely on the capabilities offered by BDA to transform the way networks are designed. This includes employing BDA to predict and minimise the bandwidth utilisation, anticipate and prepare for upcoming failures, and predict the precise energy requirements. Hence, creating a network with fewer outages, higher user satisfaction, and an enhanced performance.

The network design process using big data can be outlined as shown in Figure 3-3. Big data is collected from the network, stored, and processed in a big data cluster to extract useful information, such as trends, patterns, and correlations (step 1). The resulting information is then transferred to the decision-making platforms where a new design decision for the network is evaluated by algorithms based on the inward inferred knowledge (step 2). Finally, the new design decision is sent as feedback configuration parameters to the network where re-configuration is implemented (step 3). It should be noted that the duration of the above-mentioned cycle might vary depending on the application type of the network, e.g., enterprise, healthcare, agriculture, or transportation. For instance, enterprise networks can generate large amounts of data over a short period and usually configuration faults could be undone anytime. On the other hand, healthcare networks usually generate less monitoring data over time, and they should not be re-configured until there is sufficient data available, as frequent reconfigurations may result in failures with severe impacts on peoples' health.

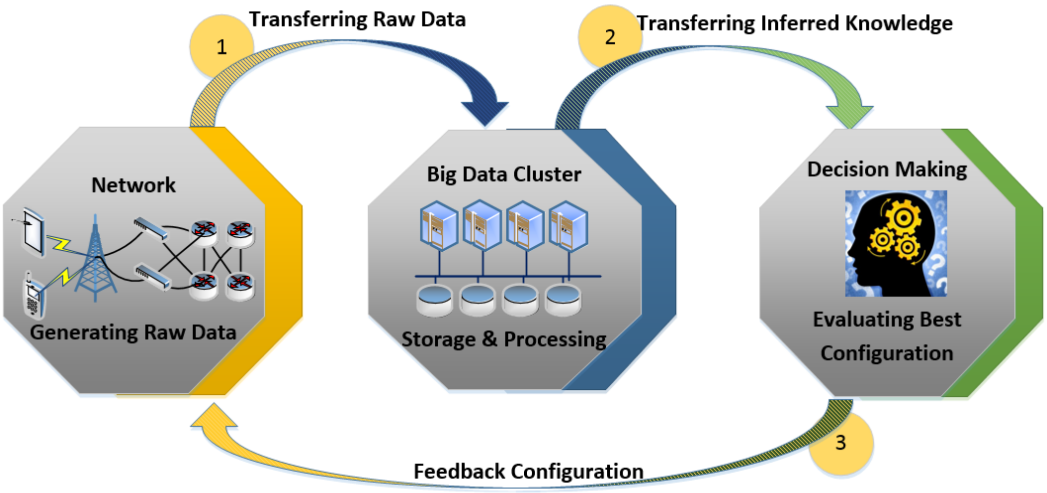


Figure 3-3: BDA-powered network design cycle

3.7.2 Challenges facing the use of BDA in Network Design

3.7.2.1 Network size vs BDA gains

Depending on the network size, the ease of redesigning a network through the feedback cycle we mentioned in Figure 3-3 is highly affected by the number of nodes. For instance, large data streams can be generated from the mass deployment of small Wireless Sensor Networks (WSNs) nodes and IoT [143]. The collected data may not carry a meaningful value until it is effectively analysed. However, analysing or mining that immense amount of data demands on finely tuned big data analytical capabilities, which turns out to be a challenging task [144]. Furthermore, these massive amounts of data require hierarchal communication and data processing solutions. The planning of such deployments in conjunction with the data processing framework is a challenging task [143].

Comparing optical to IoT networks, the former has a small number of nodes, hence they are easier to redesign, while the latter has a larger number of connected objects, and that can impose a problem.

3.7.2.2 Security and Privacy

Users' common patterns can be of great help. Network users can share certain patterns, like downloading some popular videos, retweeting about some certain upcoming game that would take place downtown, or even checking the same online channels. This information can be of a great value when used for network planning or optimisation. However, to use this information, access to user data has to be obtained, which is a thought that may cause unrest for many. When dealing with user data, there is always a flag raised, and that flag carries two issues: these issues are the security and the privacy of the data. This is why big data has to be protected from unauthorised access and release [90].

Big data security is a vital topic. If we want to label a system as "secured", it must meet the data security requirements, which are [145]:

1- *Confidentiality*: This implies the means to protect the data from unapproved disclosure.
2- *Integrity*: This implies the measures taken to protect the data from being modified improperly or without permission.
3- *Availability*: This is the system's ability to prevent and recover from hardware as well as software failures that might result in the database system being unavailable.

Privacy of data is an increasing concern. As a matter of fact, having accessible data does not mean it is ethical to access it [146]. Electronic health records have strict laws that precisely identify what can and cannot be accessed.

As an example, a user's location information can be tracked through cell towers and after a while, "a trail of crumbs" is going to be left by the user that could be used to link the user to their residence or office location, and to eventually determine the user's identity, private health information (e.g., attending a cancer treatment centre) or religious preferences (e.g., attending a church) may be discovered by tracking the user's movement over time [147], especially when we take into consideration the close correlation between an individual's identity and their movement patterns [148]. Some user data can be very valuable, for example, the estimated value of all global personal location data could reach $100 billion in revenue during the next 10 years for service providers, and when it comes to consumers and business end users, that figure can reach up to $700 billion [94].

With no obvious and secure way to handle the collected user data, BDA cannot be considered a reliable system. The security issues related to BDA can be divided into four concerns, starting with an input (e.g., handheld device, sensor, or even IoT device) where protecting the sensors from being compromised by attacks is regarded as an important security issue, as well as the other areas of data analysis, output, and communication with other systems [149]. It should be noted that these concerns are present in all steps throughout the design cycle shown in Figure 3-3.

A solution that was designed to address the big data security and privacy challenge is the integrated Rule-Oriented Data System (iRODS [150]). This novel technology was designed to ensure security and privacy in big data, and it has some technological features such as federated data grid or "intelligent clouds", distributed rule engine, “iCAT” metadata catalogue, storage access layer that facilitates common access, two ways of interfacing graphical and command line, and APIs to interact with the iRODS data grid [90, 150].

In a position paper, the authors of [151] noted a number of privacy-preserving challenges in the realm of BDA, and these challenges are classified as follows:

1- Individuals’ Interaction:
   a. *Transparency*: BDA is mostly associated with information collection and processing of specific individuals’ data. However, this means that each individual is entitled to know about the data processing operations conducted on his/her data, and the challenging part is in allocating that specific piece of information linked to that person’s identity
   b. *Individual’s Consent*: According to many privacy laws, an individual is entitled to the right to be asked for his/her *informed consent,* and such consent is a way of ensuring the individual is aware of the type of processing that is conducted. This type of consent, along with the explanation it requires is in fact considered challenging.
   c. *Consent Cancellation and Discarding Personal Data*: Granting consent, on one hand, should also allow the right of revoking it. However, if an individual wished for his/her consent to be cancelled, then this means all personal data has to be erased as well. This is a challenging requirement when considering the fact that the data might have been spread to various data collectors and data analysts.

2- Re-Identification Attacks: A user's identity may be compromised when correlating different types of datasets, and this type of attack was further classified:

    a. Correlation Attacks.
    b. Arbitrary Identification Attacks.
    c. Targeted Identification Attacks.

3- Probable vs. Provable Results: Different results can be produced by different queries conducted upon datasets. In this way, a provable link can turn out to be merely a probable one.

4- Economical Outcomes: Providing huge amounts of datasets in advance is essential for BDA to work. One way to provide such datasets is by buying them from data providers who offer to sell their users' data to their customers, thus privacy threats might appear. Context faults along with confusion and distraction are just two examples of other threats (i.e., fraud, censorship, and surveillance).

3.7.2.3 Data Centre Scalability

In the big data paradigm, data centres are not only a platform to concentrate data storage, but can also carry out further responsibilities, such as acquiring, managing, organising, processing and leveraging data values and functions. That would encourage the growth of the infrastructure and related software [91].

The continuous expansion in data volume, coupled with the ever greater demand for faster processing speeds, and the increasing complexity of Relational Database Management System (RDBMS) are considered the main elements to motivate the hunt for expandable (scalable) data centres to handle the data volume and parallel processing requirements; hence, a number of technical challenges have to be taken into consideration when we try to design a scalable data centre that can efficiently store, process, and analyse big data, these challenges can be mapped to the middle octagon (big data cluster) shown in Figure 3-3 and they are:

1. Taking into consideration the variety and sheer volume of the disparate data sources, just collecting and integrating data with scalability from scattered locations is a difficult task to accomplish.

2. Massive datasets must be mined by BDA at different levels and in either a real time or near real time fashion.
3. Massive and heterogeneous datasets are to be stored and managed by big data systems while providing the function and performance guarantees needed in terms of fast retrieval, scalability, and privacy protection. Facebook is a clear example, in that particular matter as it needs to store, access, and analyse over 30 petabytes of user-generated data [94].

Although some might claim that the current problem is not about storage (large volume), but it is about the online processing ability [66], a scalable data centre should also incorporate the ability to have a scalable storage system. Non-volatile memory (NVM) technologies are expected to have a promising role in future memory/storage designs [152]. An ideal storage platform has three vital points (constraints) to meet: it should support efficient data access in case of failure (network partitions and node failures), offer its clients a consistent view of the data, and provides high-availability. However, according to Brewer's CAP theorem [153], this ideal system cannot exist, which is due to the fact that it is impossible for the consistency to be guaranteed and for high-availability to be offered in the presence of network partitions. As a result, one of the above constraints has to be relaxed by distributed storage systems [152]. When it comes to securing the required processing speed, Chip Multiprocessors (CMPs) are expected to be the computational plotter for BDA [152]. Targeting the emerging trend, Datacentre-on-Chip (DoC) architectures were proposed by the authors of [94], with four usage models that depend on the state of the consolidating applications, if they were cooperating or not. Key scalability challenges were identified and addressed by cache hierarchies and shortage in performance isolation [152, 154].

3.8 Chapter Summary

There are many areas in which BDA can be utilised in the network design process. The concept of gathering network data and correlating them with user trends and service requirements can indeed create an adaptive and user-centric network design. Due to the subject at hand, we focused on the field of wireless communication networks design using big data. Delving deeper reveals that the field of 5G is getting the majority of the researchers' attention due to the new opportunities it has to offer. Furthermore, industrial efforts toward optimising networks based on BDA reflect the increasing trend toward employing AI-like approaches, such as pattern recognition and ML for network design. Some of the considered solutions handle big data in a batch manner while others are capable of performing real-time processing. Handling big data in a batch mode can offer more accurate information at the expense of delayed results due to the size of the processed data, while real-time processing offers fast results at the expense of accuracy. Hence, it would be an application-dependent decision whether to choose the former or the latter option.

We predict that the field of network design based on BDA will continue to flourish in the near future as more data are collected from the networks and processed to extract useful information regarding network behaviour. In the far future, or maybe quite soon, as some claim, employing quantum computing for ML purposes could help in dethroning Moor's law and provide more processing space per unit time. This extra space can be harnessed for BDA employed in network design.

Chapter 4
Data Preparation and Big Data Analytics Engine

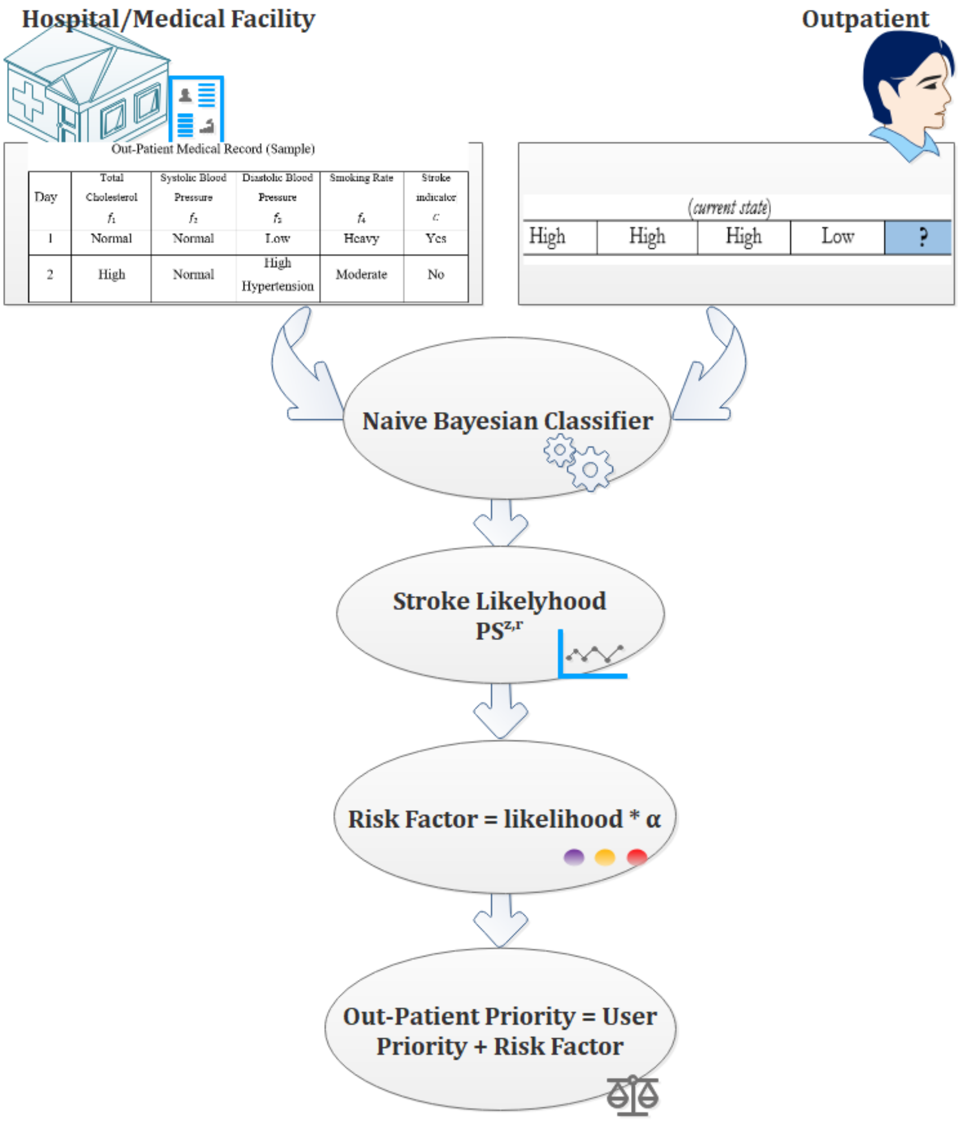
4.1 Introduction

Our goal is the prioritisation of OPs connected to a cellular network covering an urban environment. The OP prioritisation we seek to achieve is based on the severity of the OP's medical status. Big data harvested from the OPs' medical records, along with current readings from their body-connected medical IoT sensors are processed and analysed to predict the likelihood of a life-threatening medical condition, for instance, an imminent stroke. The OP prioritisation procedure is divided into two main parts; data analytics part, and the MILP-aided cellular optimisation part. In this chapter, we present the first part, illustrating the dataset preparation stages, the method used to calculate the stroke likelihood where the set of mathematical programming formulations that will be adopted throughout Chapter 5, Chapter 6, and Chapter 7 are presented. Finally, we outline the approach we employed to interpret the stroke likelihood as an effective user priority (i.e., weight) in the later optimisation stage.

4.2 BDA Engine

We consider an urban environment covered by a cellular network. A BDA engine is responsible for calculating the stroke likelihood of the OPs residing in the cellular network's coverage area. The goal is to prioritise the OPs over normal users in terms of radio resource allocation. OPs with a higher likelihood of stroke must transmit their data as soon as possible. However, if the OP was assigned a channel with a low SINR, the required medical response may not arrive in time.

The OPs' data is analysed in a cloud-located BDA engine running a naïve Bayesian (NB) classifier, one of BDA's algorithms [155]. The role of the NB classifier is illustrated in Figure 4-1. This engine is used to predict the stroke likelihood for an OP. Based on this likelihood, the OPs are assigned proportional weights (i.e. priorities) to grant them PRBs with an optimal SINR favouring them over normal (i.e., healthy) users.


Hospital/Medical Facility
Outpatient
Naive Bayesian Classifier
Stroke Likelyhood
PS^{z,r}
Risk Factor = likelihood * α
Out-Patient Priority = User Priority + Risk Factor

Out-Patient Medical Record (Sample)

Day	Total Cholesterol f_1	Systolic Blood Pressure f_2	Diastolic Blood Pressure f_3	Smoking Rate f_4	Stroke indicator c
1	Normal	Normal	Low	Heavy	Yes
2	High	Normal	High Hypertension	Moderate	No

(current state)

High	High	High	Low	?

Figure 4-1: BDA Engine / OP Priority Calculation Procedure

4.2.1 The NB Classifier

We used the NB classifier to determine the likelihood of occurrence of a certain incident c (e.g., a stroke) relying on a given set of independent feature variables f_i obtained from the OPs' big data (i.e., medical records). Given, a *current state* of a certain OP, the classifier can use the training dataset (medical record) to determine the likelihood that this OP would suffer a stroke and quantify it as a risk factor. These feature variables represent the vital readings (e.g., Systolic and Diastolic blood pressure, total cholesterol, and smoking rate) that can be collected by body-attached IoT sensors and fed to the BDA where the NB classifier resides. It is worth noting that this classifier is termed *naïve* due to the assumption it makes that the feature variables are conditionally independent [22].

4.2.2 The NB Classifier's Track Record

The NB classifier is preferred over other classifiers due to the following reasons; (i) The classifier's linearity [156] facilitates its direct joint use with the MILP while it exerts less computational burden due to its low complexity. Employing nonlinear classifiers imposes the use of additional linearisation procedures hence the model's complexity increases. This ultimately impedes further the system's development. Non-linear algorithms (e.g. artificial neural networks) can be computationally intensive by nature. Additionally, this can slow future model developments and scalable expansions; (ii) In a comprehensive study in [157], the authors stated that it is complicated to select a single tool for all types of disease analysis and they chose the NB classifier for heart disease problems; (iii) According to [158], the NB classifier was used for cardiovascular disease risk discovery and it was validated by a number of cardiologists where more than 80% of the respondents agreed with the classifier's accuracy; (iv) Its confirmed competitiveness when compared to other algorithms including NN and DT [22]; (v) The NB classifier requires a small training dataset [159]; (vi) It was the choice of many other researchers in cardiovascular disease risk prediction as in [159-166]; (vii) In the field of e-healthcare and disease risk prediction, the NB classifier proved to be one of the optimal (and sometimes the optimal) for such task, its accuracy surpassed DT, KNN and NN as discussed in [167]. The classifier gave higher accuracy when compared with DT in [168]. An intelligent heart disease prediction system was proposed in [169], the authors compared NB classifier, NN, and DT. The NB classifier proved to be the most effective as it had the highest percentage of correct predictions; (viii) it is optimal for any two-class concept with nominal features [160].

4.2.3 MILP Definitions

The following sets, parameters, and variables are defined to represent the developed MILP-compliant NB classifier.

Table 4-1: System Sets and Parameters

Sets	
$\boldsymbol{\mathcal{K}}$	Set of users.
$\boldsymbol{\mathcal{D}}$	Set of days.
$\boldsymbol{\mathcal{F}}$	Set of features in learning dataset.
$\boldsymbol{C}$	Set of classes in learning dataset.
$\boldsymbol{V_{F_i}^{r}}$	Set of values feature F_i can take in the learning dataset.
$\boldsymbol{V_{C_i}^{r}}$	Set of values a class variable C_i can take in the learning dataset.
$\boldsymbol{\mathcal{I}}$	Set of features and class variables.
$\boldsymbol{\mathcal{Z}}$	Set of outpatient users,($\mathcal{Z} \subset \mathcal{K}$).
Parameters	
$\boldsymbol{CP_{i,v}^{c,z}}$	The conditional probability that input feature i takes the value v given that outpatient z has class C considering input feature i of value v given class c for outpatient z.
$\boldsymbol{CS_i}$	The current state of the patient in feature *i* (e.g. Cholesterol value).
$\boldsymbol{V_{F_i}^{CS_i,z}}$	${CS_i}^{th}$ value taken by feature F_i for patient z.
$\boldsymbol{V_{C_i}^{CS_i,z}}$	${CS_i}^{th}$ value taken by class C_i for patient z.
$\boldsymbol{E_{F_i}^{j,d,z}}$	Binary variable, $E_{F_i}^{j,d,z} = 1$ if feature F_i takes the j^{th} value on day d for outpatient z, 0 otherwise.
$\boldsymbol{G_{C_i}^{r,d,z}}$	Binary variable, $G_{C_i}^{r,d,z} = 1$ if class C_i takes the r^{th} value on day d for outpatient z, 0 otherwise.
$\boldsymbol{S_{F_iC_i}^{j,r,d,z}}$	Binary variable, $S_{F_iC_i}^{j,r,d,z} = 1$ if $E_{F_i}^{j,d}$=1 and $G_{C_i}^{r,d,z}$=1 (Logical AND operation).
$\boldsymbol{UP_k}$	User priority (UP_k =1 for normal users whereas $UP_k > 1$ is granted for OPs depending on their risk factor).
$\boldsymbol{PS^{z,r}}$	The probability of stroke of outpatient z.

α	Tuning factor.
$\boldsymbol{NU}$	The total number of normal users.

4.2.4 Calculating the Stroke Likelihood

Developing the NB to work jointly with the MILP requires the reformulation of the stroke likelihood calculation method. The primary, MILP-noncompliant, mathematical formulation on which the NB classifier is based is depicted in equations (4-1) and (4-2).

Given a dataset comprised of a set of independent variables, called the ***feature variables***, and a set of dependent variables, called the ***class variables***. The *likelihood* of F given C is given as:

$$p(F_i = f_i | C = c) = \frac{\sum_{i=1}^{n}(C = c \wedge F_i = f_i)}{\sum_{i=1}^{n}(C_i = C_i)} \tag{4-1}$$

The NB classifier's *posterior probability* can be expressed as shown in equation (4-2).

$$p(C = c | F_i = f_i) = P(C = c) \prod_{i=1}^{n} P(Fi = fi \,| C = c) \tag{4-2}$$

where $P(C = c)$ represents *the prior probability* of stroke, in other words, it is the number of days in which a stroke occurred over the total number of days (i.e., observation period). While $\prod_{i=1}^{n} P(Fi = fi \,| C = c)$ represents the *joint probability*.

A dataset comprised of five columns is depicted in Table 4-2. The monitored body readings are stored in four columns represented by the ***feature variables*** $f_1, \ldots f_4$ reflecting the recorded state of each feature, whereas the fifth column represents the ***class variable*** C that registers whether a stroke (or a critical state) occurred in the corresponding day.

Table 4-2: Out-Patient Medical Record (Sample)

Day	**Total Cholesterol** f_1	**Systolic Blood Pressure** f_2	**Diastolic Blood Pressure** f_3	**Smoking Rate** f_4	**Stroke indicator** C
1	Normal	Normal	Low	Heavy	Yes
2	High	Normal	High Hypertension	Moderate	No
⋮	⋮	⋮	⋮	⋮	⋮
30	Optimal	High Hypertension	Pre-hypertension	Light	Yes
(current state)					
X	High	High Hypertension	High Hypertension	Light	?

The total number of rows represents the observation period for each OP and in this work, it is 30 which stands for 30 days. The total number of medical records is equivalent to the number of OPs, which in this thesis is three OPs. It should be noted that since the dataset is text-based with no multimedia components, its size is measured in kilobytes of data and this is harmonious with other datasets as in [170].

The classifier reads the OP's *medical record* (check Table 4-2) and uses the OP's *current state* (the lower part in Table 4-2) to predict the likelihood of an upcoming stroke. This likelihood is to be converted later (in the upcoming subsection) into a risk factor used to calculate the weight given to each OP to be prioritised among other users during PRB assignment which is implemented in this work using a MILP and a heuristic, as explained in the subsequent chapter. We also note that the terms "user weight" and "user priority" are used interchangeably throughout this thesis.

4.2.4.1 The Framingham Dataset

Since preserving the patient's privacy is of utmost importance for healthcare providers, it was not possible to acquire cardiovascular disease datasets of patients monitored over an extended period of time. The available datasets either reported statistics or were acquired through a collaboration with a medical institute that provided them with such datasets. Unfortunately, such datasets were not publicly accessible as in [158]. Thus, instead of generating a random dataset and risk having non-medically-compliant readings, we are fortunate in that the Framingham heart study in [171] has a big dataset that covers the features we needed. We populated our dataset by segmenting rows from the Framingham dataset and assigned each segment to an OP. Thus, the resulting dataset represents an observational period of 30 readings for each OP. It is worth noting that the Framingham cardiovascular cohort study started in 1948, and targeted adults residing in the town of Framingham, Massachusetts. The study is ongoing, and a new phase has started in 2002 with the enrolment of the third generation of participants [172]. The above–mentioned OP data has the characteristics of big data; hence, BDA algorithms can be used to predict the likelihood of occurrence of a certain incident (i.e. a stroke in our case).

4.2.4.2 Data Pre-processing

It should be noted that data reduction, data cleansing, and data generalisation are the data preparation steps that had to be carried out before applying the NB classifier to the Framingham dataset. Data preparation (or data pre-processing) is a vital stage to prepare the dataset before the use of BDA/ML algorithms [173], [174]. Moreover, having the dataset ready is a one-time process (i.e. before running the analysis [175]) as the rest of the procedure is for the NB to read the current state and to run its classification procedure against the outpatient's medical record (i.e. dataset) which is not time-consuming as we stated earlier. A similar process is done in relation to new incoming data (i.e., feature vector) from the outpatient. This feature vector is labelled "*current state*" in Table 4-2, which is only one row of data per user. Thus, the preparation time is negligible. As for adding the newly acquired readings to the dataset, those readings are added periodically:

1- Data Reduction

In this process, particular features are retained while others are excluded. There are three reasons for this; firstly, reducing the number of features has a direct effect on the dataset dimensions, thus, reducing the processor and memory utilisation while improving the classifier's accuracy [176]. This can be a crucial element in reducing the MILP's execution time. Secondly, in this work, we are targeting the main stroke contributors. Thus, according to [177, 178], Hyperlipidaemia (i.e. Total Cholesterol), blood pressure, and smoking are among the main contributors to a stroke. Thirdly, since each OP has a dataset comprised of their own readings, the inclusion of other fixed and very slowly-changing feature variables like weight, gender, age, and body mass index (BMI) can be avoided, hence, the selected features in this thesis (Cholesterol, blood pressure, and smoking). However, the impact of feature selection/ranking is to be investigated as a future extension to this work.

2- Data Cleansing

Incomplete, erroneous, and inconsistent entries were omitted. Thus, the resulting dataset is error-free and has a complete set of values across all entries.

3- Data Generalisation

The discretisation of data converts large numbers of continuous feature values into smaller ones. The purpose is to find concise data representations as categories [179]. The authors of [180] and [181] showed that the NB models' accuracy can be positively impacted by discretisation. Moreover, it is considered a data reduction mechanism because it reduces data from a large domain of numeric values to a subset of values that fall in categories [182].

Given the medical nature of the application and to stay in line with the medically-accredited ranges in the data discretisation stage, the ranges defined by the American National Institute of Health and the British Stroke Association in [183], [184] and [185] were adopted for the Systolic and Diastolic blood pressure values and total Cholesterol, respectively. As for the smoking rate, we categorised it into the levels: light, moderate, and heavy, respectively as in [186]. Consequently, the continuous values of the Framingham dataset were categorised as observed in Table 4-2 and according to their medically-accredited ranges shown in Table 4-3.

It should be noted that upon further examination we found that data can be discretised according to the European standards. However, investigating this is beyond the scope of this thesis.

Table 4-3: Feature Values and Their Corresponding Level

Feature	Range	Level
Total cholesterol Level (mg/dl) [185]*	<200	Optimal
	200-239	Normal
	240+	High
Systolic BP (mmHg) [183] [184] **	<120	Normal
	120-139	Pre-hypertension
	140+	High Hypertension
Diastolic BP (mmHg) [183] [184]**	<80	Normal
	80-89	Pre-hypertension
	90+	High Hypertension
Smoking rate (Cig/Day) [186]	1 - 10	Light
	11 - 19	Moderate
	20+	Heavy
*** Ranges adopted were according to the American National Institute of Health [185].** **** Ranges adopted were according to the American National Institute of Health and the British Stroke Association [183] [184].**		

4.2.5 Calculating the OP's Priority using MILP-Compliant NB Formulation

We developed the following formulations to include the NB classifier within the MILP model, where it calculates the stoke likelihood PS_z given a certain *current state* cs_i. The model then transforms this likelihood into an updated *user priority* (*weight*) UP_k indicated in equation (4-7).

Rewriting equation (4-1) in a mathematical programming formulation gives:

$$CP_{i,v}^{c,z} = P\left(F_i = V_{F_i}^{j,z} \middle| C_i = V_{C_i}^{r,z}\right) = \sum_{d=1}^{|D|} \sum_{F} \sum_{C} \frac{S_{F_iC_i}^{j,r,d,z}}{G_{C_i}^{r,d,z}} \qquad (4\text{-}3)$$

$\forall\, i \in \mathcal{I}, c \in C,\; z \in \mathcal{Z}$

where equation (4-3)is used to calculate the conditional probability $P(F_i|C_i)$ in the MILP model. The nominator represents the total number of days where the outpatient z has a certain reading $V_{F_i}^{j,z}$ that we want to test, *and* a stroke (indicated by $V_{C_1}^{1,z}$) where C_1 depicts the class stroke and $r = 1$ registers the stroke occurrence. The denominator represents the total number of stroke days.

$$S_{F_iC_i}^{j,r,d,z} \geq 0 \qquad (4\text{-}4)$$

$\forall\, z \in \mathcal{Z}, i \in \mathcal{I}, d \in \mathcal{D}$

$$S_{F_iC_i}^{j,r,d,z} = E_{F_i}^{j,d,z} + G_{C_i}^{r,d,z} - 1 \qquad (4\text{-}5)$$

$\forall\, z \in \mathcal{Z}, i \in \mathcal{I}, d \in \mathcal{D}$

Equations (4-4) and (4-5) achieve a logical AND operation in which the binary variable $S_{F_iC_i}^{j,r,d,z} = 1$ when both binary variables $E_{F_i}^{j,d,z} and\ G_{C_i}^{r,d,z}$ are equal to 1. This variable indicates that outpatient z with the j^{th} value of feature F_i has the r^{th} value of class C_i in day d.

Rewriting equation (4-2) gives:

$$PS^{z,r} = \left[\sum_{d=1}^{|D|} \frac{G_{C_i}^{r,d,z}}{|D|}\right] \prod_{i=1}^{\mathcal{I}} P\left(F_i = V_{F_i}^{CS_i,z} \middle| C_i = V_{C_i}^{CS_i,z}\right) \qquad (4\text{-}6)$$

$\forall\, z \in \mathcal{Z}$

Equation (4-6) represents the formulation we used to determine the probability of stroke $PS^{z,r}$. Given a *current state* CS_i , all feature variables F_i are considered. This means i has the range $i \leq |\mathcal{I}|$ (in this work $i = 1,..,4$). The L.H.S. represents the posterior probability that outpatient z has a stroke. The first term on the R.H.S. represents the prior probability of stroke and the second term on the R.H.S.

represents the joint probability that patient z has the given values of the features. The multiplication of the two terms on the R.H.S. shows the naïve nature of the NB estimate in this case where the features are assumed independent.

4.3 Results

Since the NB classifier produces probabilities of small magnitude, we multiplied the overall probability of stroke ($PS^{z,r}$) by a tuning factor α to produce an effective-yet-reasonable weight, which drives the objective function into favouring the imperilled outpatients. The user weight UP_k is calculated as shown in equation (4-7).

$$UP_k = 1 + \alpha \cdot PS^{z,r}$$
$$\forall\, k \in \mathcal{K}: z = k, k > NU \quad (4\text{-}7)$$

The NB calculated the OPs' stroke likelihood $PS^{z,r}$ of 0.0032, 0.0064, and 0.00208 for users 8, 9, and 10, respectively in a 10 user scenario. The use of tuning factor α yielded $1.104 \leq UP_k \leq 1.32$, $1.208 \leq UP_k \leq 1.64$, $1.312 \leq UP_k \leq 1.96m$, $1.52 \leq UP_k \leq 2.6$, $2.04 \leq UP_k \leq 4.2$ user priorities according to tuning factor values of 50, 100, 150, 250 and 500, respectively.

In order to test the classifier's accuracy, we employed the tenfold cross-validation method. The NB classifier's accuracy and precision were calculated for all outpatients' datasets. The NB classifier scored an accuracy of 60%, 63.3%, and 63.3% and precision of 65.2%, 66% and 71.6% for users 8, 9 and 10 (i.e., OP 1, 2, and 3), respectively.

4.4 Chapter Summary

This chapter illustrated the role of BDA in this thesis. It also gave an overview of the dataset used in this thesis, along with the data pre-processing stages the dataset underwent before the NB classifier can operate. Further, the reasons behind choosing the NB classifier to calculate the OP's stroke likelihood were illustrated. Moreover, the MILP-compliant NB mathematical formulations were given, and the mathematical formula used to transform the stroke likelihood into a meaningful priority asserting the favouring of the OPs over normal users in the optimisation process was presented. Finally, the classifier’s performance was inspected in terms of accuracy and precision.

Chapter 5
Patient-Centric Cellular Network Optimisation using Big Data Analytics

5.1 Introduction

BDA is one of the state-of-the-art tools to optimise networks and transform them from blind tubes that convey data, into cognitive, conscious, and self-optimising entities that can intelligently adapt according to the needs of their users. This, in fact, can be regarded as one of the highest forthcoming priorities of future networks. In this chapter, we propose a system for OP centric single-tier homogenous LTE-A network optimisation. The predicted stroke likelihood that is calculated in Chapter 4 is employed in this chapter to ensure that the OPs are assigned optimal LTE-A PRBs to transmit their critical data to their healthcare provider with minimal delay. To the best of our knowledge, this is the first time BDA are utilised to incorporate the topics of resource allocation, patient monitoring, disease risk prediction, and prioritisation to optimise a cellular network in an OP-conscious manner. The PRBs assignment is optimised using MILP and verified using a heuristic. Two approaches are proposed, a WSRMax approach and a PF approach. The approaches increased the OPs' average SINR by 26.6% and 40.5%, respectively. The WSRMax approach increased the system's total SINR to a level higher than that of the PF approach, however, the PF approach reported higher SINRs for the OPs, better fairness and a lower margin of error.

5.2 System Model

We consider an urban environment covered by an LTE-A cellular network. The area is populated with a number of users scattered at random distances from the BSs (between 300 and 600 meters). The users fall into two categories; normal (healthy) users and OPs as shown in Figure 5-1. As we previously indicated, cellular networks can provide an optimal way for OPs to have a connection when compared to Wi-Fi or wired connections. Since the OPs are randomly-located, different power levels (signal strengths) will be received from their mobile devices. We are assuming a system with a slow fading channel where the channel gain remains constant within one TTI. Thus, the coherence time is assumed to be greater than the duration of a TTI. To this end, the objective function of our optimisation model guarantees the allocation of high gain PRBs to the OPs and according to their likelihood of stroke. Aiming at maximising the total SINR received at the BS. Thus, enabling them to transmit their data as soon as possible, while preserving fairness among users to ensure such a resource allocation scheme will not negatively impact other users. We note that the terms ‘healthy user’ and ‘normal user’ are used interchangeably throughout the thesis.

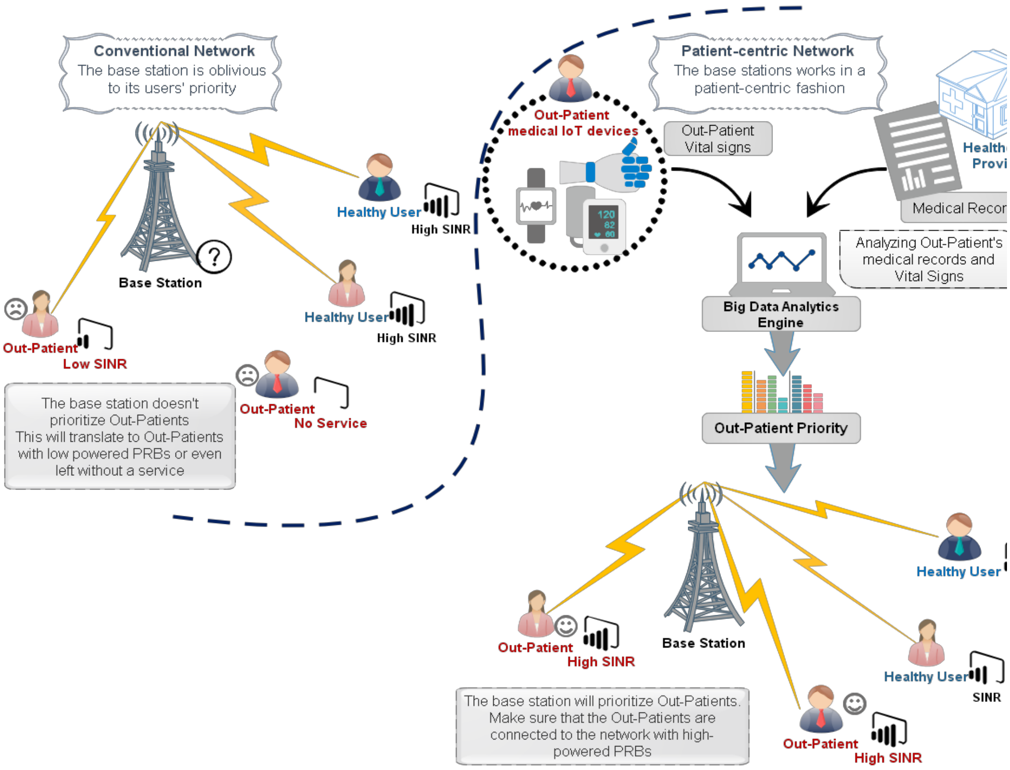

Conventional Network
The base station is oblivious to its users' priority
Healthy User
High SINR
Base Station
?
Healthy User
High SINR
Out-Patient
Low SINR
Out-Patient
No Service
The base station doesn't prioritize Out-Patients This will translate to Out-Patients with low powered PRBs or even left without a service
Out-Patient
medical IoT devices
120
82
60
Patient-centric Network
The base stations works in a patient-centric fashion
Out-Patient Vital signs
Health
Provi
Medical Recor
Analyzing Out-Patient's medical records and Vital Signs
Big Data Analytics Engine
Out-Patient Priority
Healthy User
Base Station
Out-Patient
High SINR
Healthy User
SINR
Out-Patient
High SINR
The base station will prioritize Out-Patients. Make sure that the Out-Patients are connected to the network with high-powered PRBs

Figure 5-1: Patient-Centric Cellular Network

The system will undergo the following stages, Further, we demonstrate those stages in a timeline as shown in and will have the timeline

1- The *Data Collection* stage; where the OP’s EHR and readings from the body attached medical IoT sensors are being aggregated, cleansed, and normalised. Erroneous and null entries are deleted in this stage and the dataset is prepared to be used to train a ML model.

2- The *training* stage; this is where the data collected from the previous stage is used to train the classifier(s) or the ML model(s). This stage takes place in a cloud-based BDA engine.

3- The *prediction* stage; It should be noted that this stage takes place in the cloud where each OP will have a dedicated classifier trained on the OP's own dataset. The output of this stage is the stroke likelihood for that OP.

4- The *network optimisation* stage; residing in the operator's core network side, the system utilises the stroke likelihoods acquired from the previous stage to convert them into priorities used during the radio resource allocation stage.

5- The *update and review* stage; in case it is no longer required to monitor a specific OP, or if the OP is still under monitoring, a periodic update to the OPs' EHR and thus the training dataset will take place. Hence, the classifier's introduced in the third stage must be retrained using the updated dataset. However, it should be noted that the frequency of the dataset update and classifier retraining is beyond the scope of this work and it is the subject of a future work.

It should be noted that the system's computational complexity is divided into two parts; post-operation and operational computational overheads. The former has no effect on the system's performance as it takes place before the system operation (i.e., applies to stages 1, 2, and 5 in Figure 5-2) timeline. Whereas the latter takes place during stages 3 and 4 in Figure 5-2 and it equals to $O(N^4 \log N)$ as we shall illustrate in detail in Section 5.5.3.

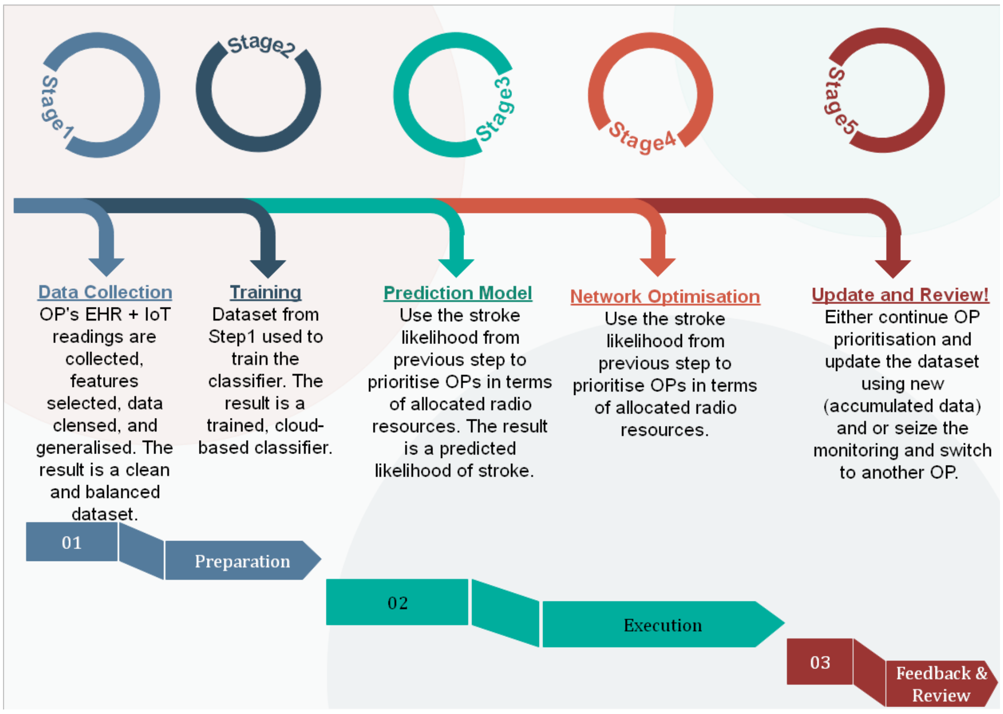

Stage1
Stage2
Stage3
Stage4
Stage5
Data Collection
OP's EHR + IoT readings are collected, features selected, data clensed, and generalised. The result is a clean and balanced dataset.
Training
Dataset from Step1 used to train the classifier. The result is a trained, cloud-based classifier.
Prediction Model
Use the stroke likelihood from previous step to prioritise OPs in terms of allocated radio resources. The result is a predicted likelihood of stroke.
Network Optimisation
Use the stroke likelihood from previous step to prioritise OPs in terms of allocated radio resources.
Update and Review!
Either continue OP prioritisation and update the dataset using new (accumulated data) and or seize the monitoring and switch to another OP.
01
Preparation
02
Execution
03
Feedback & Review

Figure 5-2: System Timeline and Operation Stages

5.3 Problem Formulation

We developed the following MILP models to optimise the cellular system resource allocation for OPs and normal users. We consider the OPs monitoring system to operate in a scenario of an LTE-A network comprising B BSs represented by set $\mathcal{B} = \{1, \dots, B\}$, operating at channels with 1.4 MHz bandwidth. Each BS b has N PRBs represented by set $\mathcal{N} = \{1, \dots, N\}$. The network serves K users (normal and OPs) represented by set $\mathcal{K} = \{1, \dots, K\}$ by allocating PRB n to connect to BS b in an instant in time. The goal is to optimise the uplink of the LTE-A network, so that the OPs are prioritised over normal users; thus, allocating them high-powered PRBs.

We formalise this problem as a MILP model. Table 5-1 defines the sets, parameters, and variables used in the network optimisation problem formulation.

Table 5-1: System Sets, Parameters, And Variables

Sets	
$\boldsymbol{\mathcal{K}}$	Set of users.
$\boldsymbol{\mathcal{N}}$	Set of physical resource blocks.
$\boldsymbol{\mathcal{B}}$	Set of base stations.
$\boldsymbol{C}$	Set of classes in learning dataset.
$\boldsymbol{\mathcal{Z}}$	Set of outpatient users,($\mathcal{Z} \subset \mathcal{K}$).
Parameters	
$\boldsymbol{UP_k}$	User priority (UP_k =1 for normal users whereas $UP_k > 1$ is granted for OPs depending on their risk factor).
$\boldsymbol{Q_{k,n}^b}$	Power received from user k using PRB n at base station b.
$\boldsymbol{H_{k,n}^b}$	Rayleigh fading with zero mean and a standard deviation equal to 1 experienced by user k using PRB n at base station b.
$\boldsymbol{A_k^b}$	Signal attenuation experienced by user k connected to base station b.
$\boldsymbol{PM}$	Maximum power allowed per uplink connection.
$\boldsymbol{P}$	Power consumed to utilise PRB n to connect user k to base station b.
$\boldsymbol{\lambda}$	An arbitrary, large positive value.
$\boldsymbol{\sigma_{k,n}^b}$	Additive White Gaussian Noise (AWGN) power in watts experienced by user k using PRB n at base station b.
$\boldsymbol{PS^{z,r}}$	The probability of stroke of outpatient z on the r^{th} value of class variable c
$\boldsymbol{m_{y,k}}$ $\boldsymbol{h_{y,k}}$	Piecewise linearisation equation coefficients for line y of user k.
$\boldsymbol{\alpha}$	Tuning factor.
$\boldsymbol{NU}$	The total number of normal users.
Variables	
$\boldsymbol{X_{k,n}^b}$	Binary decision variable $X_{k,n}^b = 1$ if user k is assigned PRB n in base station b, otherwise $X_{k,n}^b = 0$.
$\boldsymbol{T_{k,n}^b}$	The SINR of user k utilising PRB n at base station b.
$\boldsymbol{\phi_{m,n,k}^{w,b}}$	Non-negative linearisation variable where $\phi_{m,n,k}^{w,b} = T_{k,n}^b X_{m,n}^w$.

$\boldsymbol{S_k}$	SINR of user k.
$\boldsymbol{L_k}$	Logarithmic SINR of user k.

A user's SINR at the uplink side of an OFDMA network can be expressed as [18].

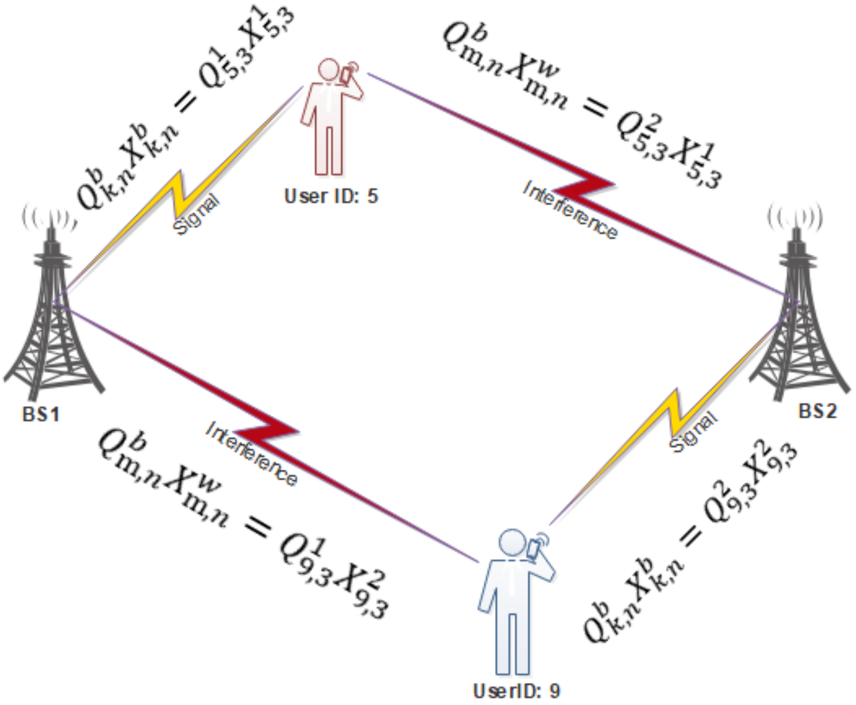
$$T_{k,n}^{b} = \frac{Signal}{Interference + Noise} = \frac{Q_{k,n}^{b} X_{k,n}^{b}}{\sum_{\substack{w \in \mathcal{B} \\ w \neq b}} \sum_{\substack{m \in \mathcal{K} \\ m \neq k}} Q_{\mathrm{m},n}^{b} X_{\mathrm{m},n}^{w} + \sigma_{k,n}^{b}} \qquad (5\text{-}1)$$

Examining the numerator (i.e. signal), $Q_{k,n}^{b} X_{k,n}^{b}$ represents the signal power received at the BS side from user k. The binary decision variable $X_{k,n}^{b} = 1$ indicates that user k is connected to BS b and occupies PRB n. The power received at BS b from the interfering user(s) $m, m \neq k$, on the same PRB is $Q_{\mathrm{m},n}^{b} X_{\mathrm{m},n}^{w}$; while $X_{\mathrm{m},n}^{w}$ indicates that the interfering user(s) m is connected to another BS $w, w \neq b$ on PRB n. The AWGN is annotated as $\sigma_{k,n}^{b}$. A graphical illustration of equation (5-1) is shown in Figure 5-3.

Rewriting equation (5-1):

$$\sum_{\substack{w \in \mathcal{B} \\ w \neq b}} \sum_{\substack{m \in \mathcal{K} \\ m \neq k}} T_{k,n}^{b} Q_{\mathrm{m},n}^{b} X_{\mathrm{m},n}^{w} + T_{k,n}^{b} \sigma_{k,n}^{b} = Q_{k,n}^{b} X_{k,n}^{b} \qquad (5\text{-}2)$$

$\forall\, k \in \mathcal{K}, n \in \mathcal{N}, b \in \mathcal{B}$



$Q_{k,n}^{b}X_{k,n}^{b} = Q_{5,3}^{1}X_{5,3}^{1}$
User ID: 5
Signal
$Q_{m,n}^{b}X_{m,n}^{w} = Q_{5,3}^{2}X_{5,3}^{1}$
Interference
BS1
BS2
Interference
$Q_{m,n}^{b}X_{m,n}^{w} = Q_{9,3}^{1}X_{9,3}^{2}$
Signal
$Q_{k,n}^{b}X_{k,n}^{b} = Q_{9,3}^{2}X_{9,3}^{2}$
UserID: 9

Figure 5-3: User Interference

The first term in (5-2) is nonlinear (quadratic) as it involves the multiplication of two variables (Continuous $T_{k,n}^{b}$ and Binary $X_{m,n}^{w}$). Therefore, linearisation is essential to solve the NP-hard model using a linear solver such as CPLEX, where the linearisation is given in (5-5) to (5-8).

We have developed two approaches to solve the resource allocation problem. The first approach uses an objective function that maximises the Weighted Sum-Rate of the SINRs experienced by the users. The second approach introduces fairness among the users by employing a Proportionally Fair (PF) objective function.

5.3.1 MILP Formulation for the WSRMax approach

The objective is to maximise the system's overall SINR. This can be realised through the maximisation of the individual users' SINRs.

5.3.1.1 Before Prioritising the OPs

The OPs' risk factors introduced in the previous chapters are scaled into priorities (i.e. weights) and used to prioritise the OPs over other users. The MILP model is formulated as follows:

Objective: Maximise

$$\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\sum_{b\in\mathcal{B}} T_{k,n}^{b}\, UP_{k} \tag{5-3}$$

The objective given in (5-3) aims to maximise the weighted sum of the users' SINRs. These weights (i.e. priorities) are higher for OPs compared to healthy users and proportional to the OPs calculated risk factor. Note that UP_k has an initial value of 1 for all users as shown in (5-4). However, the OPs will have updated values according to their risk factor. This will ultimately drive the system into prioritising the OPs over the healthy users during PRB assignment. The mathematical formulations related to the OP weight (priority) calculation were illustrated in Chapter 4.

$$UP_k = 1 \qquad \forall\, k \in \mathcal{K} \tag{5-4}$$

Constraints:

To maintain the model's linearity while performing the multiplication of the float variable $T_{k,n}^{b}$ by the binary variable $X_{m,n}^{w}$, we follow [187], and define a variable $\phi_{m,n,k}^{w,b}$ that includes all the indexes of both aforementioned (i.e., float and binary) variables as in equation (5-5). Constraints (5-6), (5-7), and (5-8) govern the multiplication procedure. As a result, the only two values satisfying the constraints are either zero (when x =0) or T (when x=1). It should be noted that **λ** is a large enough number where **λ** >>T:

Subject to:

$$\phi_{m,n,k}^{w,b} \geq 0 \tag{5-5}$$

Replacing the quadratic term $T_{k,n}^{b}X_{m,n}^{w}$ with the linearisation variable $\phi_{m,n,k}^{w,b}$ that incorporates all the indexes of the multiplied variables.

$$\phi_{m,n,k}^{w,b} \leq \lambda X_{m,n}^{w} \qquad \forall\, k,m \in \mathcal{K}, n \in \mathcal{N}, w,b \in \mathcal{B}\,, (m \neq k, b \neq w) \tag{5-6}$$

$$\phi_{m,n,k}^{w,b} \leq T_{k,n}^{b} \qquad (5\text{-}7)$$

$$\forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}\,, (m \neq k, b \neq w)$$

$$\phi_{m,n,k}^{w,b} \geq \lambda X_{m,n}^{w} + T_{k,n}^{b} - \lambda \qquad (5\text{-}8)$$

$$\forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}\,, (m \neq k, b \neq w)$$

After replacing $T_{k,n}^{b} X_{m,n}^{w}$ with $\phi_{m,n,k}^{w,b}$, equation (5-2) can thus be rewritten as in (5-9). $\phi_{m,n,k}^{w,b} = T_{k,n}^{b} X_{m,n}^{w}$ is equal to the SINR of user k connected to BS b with PRB n if there is an interfering user m connected to the other BS w with the same PRB n; it is zero otherwise.

$$\sum_{\substack{w \in \mathcal{B} \\ w \neq b}} \sum_{\substack{m \in \mathcal{K} \\ m \neq k}} Q_{\mathrm{m},n}^{b} \phi_{m,n,k}^{w,b} + T_{k,n}^{b} \sigma_{k,n}^{b} = Q_{k,n}^{b} X_{k,n}^{b} \qquad (5\text{-}9)$$

$$\forall\, k \in \mathcal{K}, n \in \mathcal{N}, b \in \mathcal{B}$$

$$\sum_{n \in \mathcal{N}} P\, X_{k,n}^{b} \leq PM \qquad (5\text{-}10)$$

$$\forall\, k \in \mathcal{K}, b \in \mathcal{B}$$

Constraint (5-10) ensures that the users do not exceed their maximum available amount of power per uplink connections (in case more than one PRB is utilised by the same user k). In the current work, the user is allowed a single PRB.

$$\sum_{k \in \mathcal{K}} X_{k,n}^{b} \leq 1 \qquad (5\text{-}11)$$

$$\forall\, n \in \mathcal{N}, b \in \mathcal{B}$$

Constraint (5-11) limits the assignment of each PRB to one user only.

$$\sum_{b \in \mathcal{B}} \sum_{n \in \mathcal{N}} X_{k,n}^{b} \geq 1 \qquad (5\text{-}12)$$

$$\forall\, k \in \mathcal{K}$$

Constraint (5-12) guarantees that each user is assigned at least one PRB from any BS. Thus, no user is left without service. Additionally, this prevents the MILP from blocking interfering users to maximise the total SINR.

5.3.1.2 After Prioritising the OPs

In this approach, OPs' risk factors introduced in the previous chapter are scaled into weights to prioritise the OPs over other users. The MILP model is formulated in the same way as mentioned in the previous subsection. However, equation (5-13) is included in this model to represent the OPs' weights (i.e. priorities) as follows:

$$UP_k = 1 + \alpha \cdot PS^{z,r} \qquad \forall\, k \in \mathcal{K}: z = k, k > NU \tag{5-13}$$

while (5-4) is replaced by (5-14) to cover the normal users only.

$$UP_k = 1 \qquad \forall\, k \in \mathcal{K}: 1 \leq k \leq NU \tag{5-14}$$

5.3.2 MILP Formulation for the PF Approach

In this approach, the objective is to maximise the logarithmic sum of the user's SINRs. Due to the nature of the natural logarithm, a slight decrease in the overall SINR might be observed but to the expense of preserving fairness among normal users.

5.3.2.1 Before Prioritising the OPs

In this case, all users are treated equally, thus there is no prioritisation in terms of resource allocation. However, keeping fairness among users still holds as a necessity. Since the only part that we are dealing with is the value of the individual user's SINR, and to simplify the manipulation of the equation before adding the natural logarithm part, we present the optimisation variable S_k, to serve as the SINR for each user k.

$$S_k = \sum_{n \in \mathcal{N}} \sum_{b \in \mathcal{B}} T_{k,n}^{b} \qquad \forall\, k \in \mathcal{K} \tag{5-15}$$

Equation (5-15) replaces the three-indexed variable $T_{k,n}^{b}$ with a single-indexed variable S_k.

$$L_k = \ln S_k \qquad (5\text{-}16)$$
$$\forall\, k \in \mathcal{K}$$

Equation (5-16) calculates L_k as a logarithmic function of the user's SINR S_k.

The objective is as shown in (5-17):

Objective: Maximise

$$\sum_{k \in K} L_k \qquad (5\text{-}17)$$

Constraints:

In addition to constraints (5-5)-(5-12) from the previous model, the PF satisfies the following constraint

Subject to:

$$L_k \leq m_{y,k} * S_k + h_{y,k} \qquad (5\text{-}18)$$
$$\forall\, k \in \mathcal{K}$$

Constraint (5-18) represents a set of piecewise linearisation relations implemented to linearize the concave function in equation (5-16). Note that constraint (5-18) corresponds to the line equation $y = mx + h$ where the line coefficients (i.e. $m_{y,k}$ and $h_{y,k}$) are selected as in [188]. It should be noted that the number of constraints used in the linearisation procedure is dictated by the total number of lines used to cover the linearized interval.

5.3.2.2 After Prioritising the OPs

In this case, the outpatients are prioritised. Equation (5-16) is rewritten to reflect the change.

$$L_k = \ln S_k \qquad (5\text{-}19)$$
$$\forall\, k \in \mathcal{K}\colon 1 \leq k \leq NU$$

Equation (5-19) shows that the log function is applied to normal users only. The OPs, on the other hand, are assigned weights instead.

Objective: Maximise

$$\sum_{k\in K,1\leq k\leq NU} L_k + \sum_{k\in K,k>NU} S_k UP_k \tag{5-20}$$

The multi-objective function in (5-20) (i) maximises the sum of the SINRs allocated to all users, (ii) Assigns OPs priority by allocating OPs PRBs with high SINRs that reflect their relative priority, and (iii) Implements Fairness: by assigning healthy users PRBs with comparable SINRs. These objectives were implemented by adding both the summation of a log function of the healthy users' SINRs (i.e. Proportional Fairness) and the weighted sum of the OPs' SINRs (OPs priority).

Constraints:

The model satisfies constraint (5-5)-(5-12) from the previous approach. In addition to equation (5-14) and:

$$L_k \leq m_{y,k} * S_k + h_{y,k}$$
$$\forall\, k \in \mathcal{K}, k \leq NU \tag{5-21}$$

Constraint (5-21) represents the same set of equations for the piecewise linearisation that was used in constraint (5-18), however, the difference is in the range of users it is applied to.

5.3.3 Calculating the Received Power

The received signal power (in Watts) $Q_{k,n}^{b}$ varies according to the channel conditions and the distance between the user and the BS. Considering Rayleigh fading denoted by $H_{k,n}^{b}$ and distance dependent path loss denoted by $A_{k,n}^{b}$ [19], the received signal power is given as:

$$Q_{k,n}^{b} = P\, H_{k,n}^{b} A_{k}^{b} \tag{5-22}$$

where $H_{k,n}^{b}$ denotes Rayleigh fading and A_{k}^{b} represents power loss due to attenuation (distance dependent path loss) and is given in (5-23) [19]:

$$A\ (dBm) = 128 + 37.6\ \log_{10} \frac{distance(meters)}{1000} \qquad (5\text{-}23)$$

To unify the units, equation (5-24) is used to convert the power to Watts.

$$A\ (\mathrm{m}w) = 10^{\frac{A(dBm)}{10}} \qquad (5\text{-}24)$$

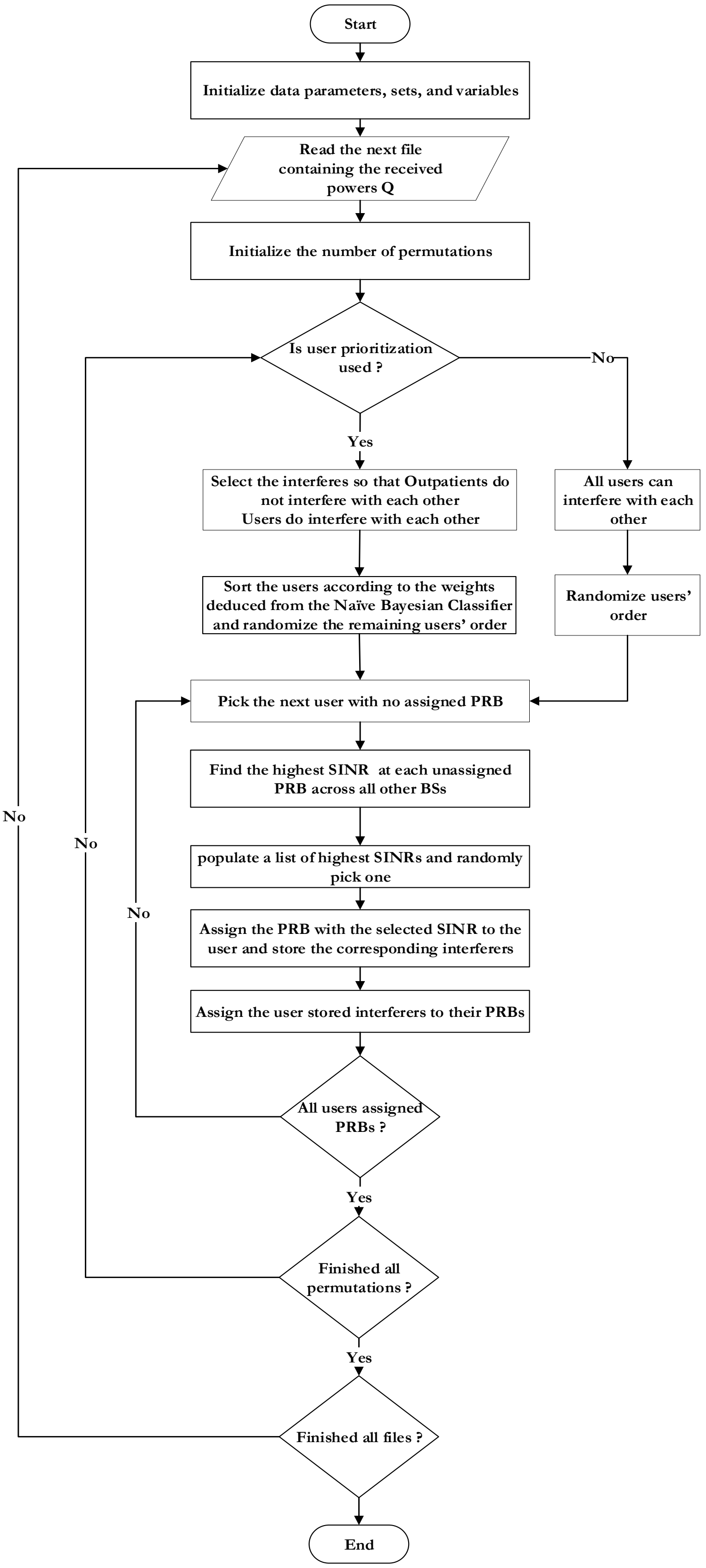
5.4 Heuristic

To provide a method to validate the MILP operation we developed a heuristic approach optimising the PRBs assignment based on the user's priority. The heuristic uses simple rules and therefore can be used in the cellular network control plane to carry out resource allocation in real time. The heuristic, as shown in the flowchart in Figure 5-4, starts by initialising the data parameters, sets, variables and reads the received power (Q) values from a separate file. A check for user prioritisation takes place. This affects the users' admittance order to the system. If user prioritisation is ON (i.e. BDA is used), the OPs will be arranged according to their priority such that the most critical OP will be served first. This kind of check is vital at this stage due to the sequential nature of the heuristic, thus, the first few users will be granted high SINRs due to the higher number of available channels. OPs do not compete with each other over the available PRBs, i.e. their interfering candidates are normal users only. Finding the PRB at which a user achieves a relatively-high SINR is done by assigning a PRB where interference is attributed to a subset of $|\mathcal{B}|$-1 interferers with minimum interfering power to that user at its PRB, where $|\mathcal{B}|$ is the number of BSs (the cardinality of $\mathcal{B}$). As the heuristic continues to run, the PRB availability is reduced. Once the PRBs are allocated to the OPs, the total number of allocated PRBs will equal to $(2 * \mathcal{Z})$. On the other hand, the number of free PRBs (FPRB) will be equal to $[\mathcal{B} * \mathcal{N}] - [2 * \mathcal{Z}]$ giving a total of 2^{FPRB}combinations. Finding an interfering user with the minimum power on each RB (i.e. maximum SINR) results in reducing the above number of combinations. Accordingly, a pool with the length $|FPRB|$ comprised of the highest achievable SINR on each PRB will be formed. The heuristic follows a semi-greedy approach [189]. Thus, one SINR will be randomly selected from the pool of best SINRs. The reasons behind this selection criterion are (i) to establish local fairness between the user and its interferer so that the interferer does not endure a huge impact by being assigned a very low-powered PRB; moreover, (ii) to conform to the objective function in which each individual user's SINR is maximised while maximising the overall system-wide SINR. Once the user is assigned a SINR, the corresponding PRB(s) is assigned to the user and the interferer. The heuristic repeats the above procedure for the remaining users. Due to its sequential nature, this heuristic was iterated 1000 times, randomising the users' admission order (serving sequence) to the system in each iteration, while maintaining the semi-deterministic nature of the interferer's PRB assignment stage. The users' average SINRs are then calculated. Thus, applying this heuristic over different realisations of the network adds fairness among users in the long run. Sensitivity analysis was carried out to calculate the 95% confidence interval. To that end, the heuristic was applied to over 100 files each containing different values representing the powers received from the BS. Concurring results between the heuristic and the MILP model operation can be observed, as will be shown in the results section.

It is of interest to compare the performance of the MILP which leads to the optimal solution with the performance of the heuristic which is sequential in nature and sub-optimal. In our optimisation model, the objective is to maximise the overall system's SINR by maximising the SINRs of all individual users *while* prioritising outpatient users over the healthy ones. This proceeds by allocating to user-A PRB-X at BS-1 which has a relatively high received power among the unassigned PRBs on that BS *while* choosing an unassigned interfering user-B to utilise the same PRB-X where the received power on BS-1 is one of the lowest. Such a scheme will be approached differently by the MILP and the heuristic as their method of operation differs in the following manner:

Given a certain objective and a number of constraints, the MILP produces a *feasible region* bounded by the constraints defined in the optimisation problem. All points within that region can *satisfy* the objective. However, only *one point typically* represents the *optimal* solution. The MILP tries all the points at the boundary of the feasible region for all the possible user-interferer combinations and chooses the *optimal* result which best satisfies the objective (i.e. either attaining the maximum or the minimum).

The heuristic, on the other hand, works on a sequential basis. In our case, it admits and examines the users and the interferers one by one (i.e., sequentially). The user admitted first will have the advantage of being able to select from a wide range of resource blocks that correspond to different potential interferers. This range decreases as PRBs are assigned to the users one by one. Therefore, first-served users have the highest SINRs. To assert fairness between users, we have randomised the user admission order to the system in each iteration and this fairness is demonstrated when comparing the heuristic and the MILP results in Figure 5-5, Figure 5-6, Figure 5-8, Figure 5-9, Figure 5-10, and Figure 5-12.



Start
Initialize data parameters, sets, and variables
Read the next file containing the received powers Q
Initialize the number of permutations
Is user prioritization used ?
No
Yes
Select the interferes so that Outpatients do not interfere with each other Users do interfere with each other
All users can interfere with each other
Sort the users according to the weights deduced from the Naïve Bayesian Classifier and randomize the remaining users' order
Randomize users' order
Pick the next user with no assigned PRB
Find the highest SINR at each unassigned PRB across all other BSs
populate a list of highest SINRs and randomly pick one
Assign the PRB with the selected SINR to the user and store the corresponding interferers
Assign the user stored interferers to their PRBs
All users assigned PRBs ?
No
Yes
Finished all permutations ?
No
Yes
Finished all files ?
No
End

Figure 5-4: The heuristic flowchart

5.5 Results and Discussion

Before delving into the results of the MILP model and heuristic, the parameters indicated in Table 5-2 should be noted. We consider a cellular network that operates in an urban environment, hence Rayleigh fading channel model with path loss. The results evaluate two scenarios; the first represents the state of the network before using BDA to prioritise the OPs. In this case, all the users were given equal base priority (i.e. weight) of 1. The second scenario represents the network state after using BDA where the OPs' priorities are updated according to their risk factor and the value of the tuning factor α.

The proposed system assumes a cloud-based setup with each OP having their own dataset comprised of their daily observations. The proposed system employs a dataset of daily observations over the course of a month, with a requirement to append additional observations periodically. In this work, we have assumed that the update frequency is daily. Additionally, the proposed system considers a system that is in operation. Here the dataset and the trained model are operational and the OP current reading is utilised by the NB classifier with the dataset to evaluate their current medical condition. Moreover, we would like to highlight that the classifier's role in this thesis is to calculate the *probability* of stroke. Since the outpatients are all under continuous monitoring, they are favoured according to their probability of stroke as long as the system is operational. The OPs' stroke likelihood $PS^{z,r}$ were 0.0032, 0.0064, and 0.00208 for users 8, 9, and 10, respectively.

We have employed the tenfold cross-validation method. The classifier's accuracy and precision were calculated for all outpatients' datasets. The classifier scored an accuracy of 60%, 63.3%, and 63.3% and precision of 65.2%, 66% and 71.6% for users 8, 9 and 10 (i.e., OP 1, 2, and 3), respectively. The use of equation (5-13) produced $1.104 \leq UP_k \leq 1.32$, $1.208 \leq UP_k \leq 1.64$, $1.312 \leq UP_k \leq 1.96m$, $1.52 \leq UP_k \leq 2.6, 2.04 \leq UP_k \leq 4.2$ user priorities according to tuning factor values of α of 50, 100, 150, 250 and 500, respectively.

Table 5-2: Model Parameters

Parameter	Description
LTE-A system bandwidth	1.4 MHz
Channel Model	Path Loss [19] and Rayleigh fading [18]
No. of BS	2
Number of PRBs per BS	5
Number of users	10
Number of normal users (NU)	7
Number of OPs	3
AWGN ($\sigma_{k,n}^{b}$)	-162 dBm/Hz [19]
The distance between user k and BS b	(300 - 600) m
Maximum transmission power per connection PM	23 dBm [19]
UE transmission power per PRB	17 dBm
Base (i.e. normal user priority) weight	1
Outpatient priority UP_k calculation method	Naïve Bayesian classifier
OP observation period	30 Days
Tuning factor (i.e., α) values	50, 100, 150, 250, and 500

5.5.1 The WSRMax Approach

5.5.1.1 Before Prioritising the OPs

In this scenario, BDA is not employed to prioritise the OPs, i.e., all users have equal weights equivalent to the *base user weight* (i.e. 1). Observing Figure 5-5, it can be seen that the OPs (represented by users 8, 9, and 10, in both the MILP and heuristic results) are assigned PRBs with near average SINR as the MILP and heuristic strive to maximise the overall SINR.

Analogous SINR values can be observed in Figure 5-5 for both the MILP and the heuristic. The average SINRs computed through the heuristic and the MILP approaches are comparable at around 5.4 and 5.5, respectively.

As a measure of fairness, i.e. to quantify how close the SINR values are to the mean, we considered accentuating the Standard Deviation (SD) for the users' SINRs. The results are 0.4 and 0.3 for the heuristic and the MILP, respectively. Thus, the results confirm that the heuristic can approach the MILP and provide an acceptable level of fairness among the users by implementing the described permutation over independent realisations of the channel, at the expense of slightly sacrificing the overall SINR. An extensive sensitivity analysis was carried out, and 95% confidence intervals for each user's SINRs are depicted in Figure 5-5. The average SINR lied between 5.1 and 6 for the MILP results, and between 4.5 and 5.7 for the heuristic results.

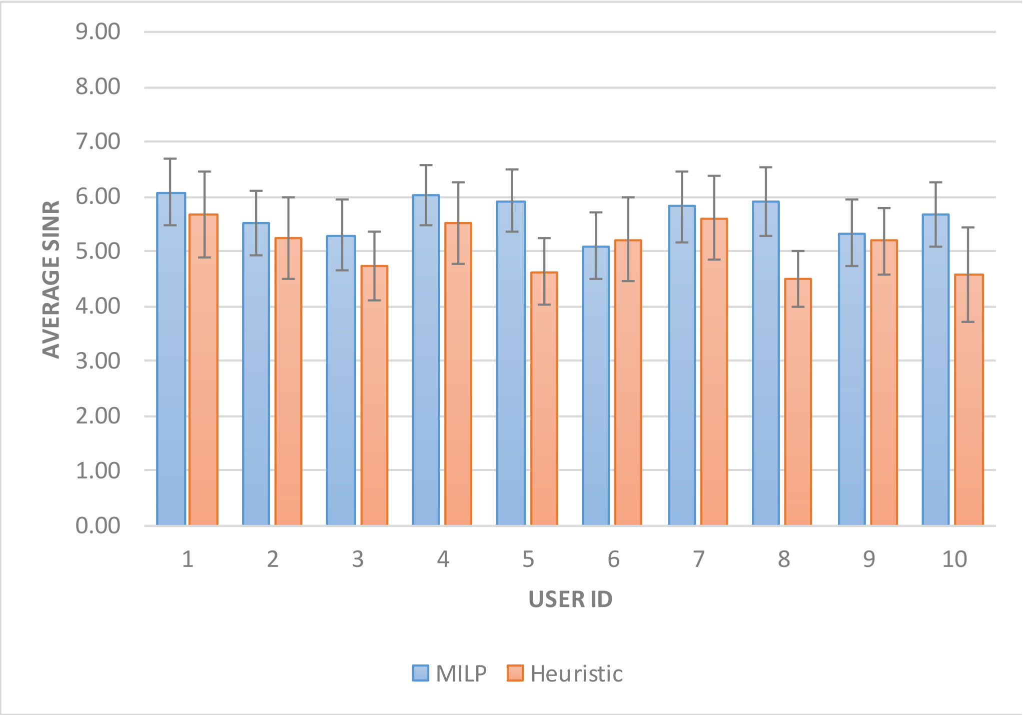


Figure 5-5: Users' SINR before using BDA (WSRMax Approach)

5.5.1.2 After Prioritising the OPs

In this scenario, the use of BDA resulted in assigning OPs higher priority than normal users by means of the NB classifier. The results shown in Figure 5-6 clearly demonstrate that all the OPs (users 8, 9, and 10) were assigned PRBs with high SINRs compared to their previous SINRs in Figure 5-5. The system-wide performance is a trade-off (*optimally* selected) between the task of assigning higher SINRS to OPs versus a reduction in the average SINR in this scenario (between 0.3% (α=50) and 6% (α=500)) compared to the average SINR in the first scenario. This reduction in the average SINR is due to the fact that the system was forced to choose a PRB assignment scheme that prioritises the maximisation of OPs' individual SINRs over the total SINR. The results also show that the heuristic approaches the MILP performance, with a very comparable SINRs, however, the heuristic mostly displayed a marginally higher OP SINRs. This is due to the sequential nature of the heuristic which forced the system to serve the OPs first after further arranging them according to their priorities. This challenge was mitigated by preparing a list of highest achievable SINRs and randomly selecting one. The selection criterion of the user and its interferer was conducted on a sequential and a semi-deterministic manner, respectively to adds fairness between users as illustrated in Section 5.4 .

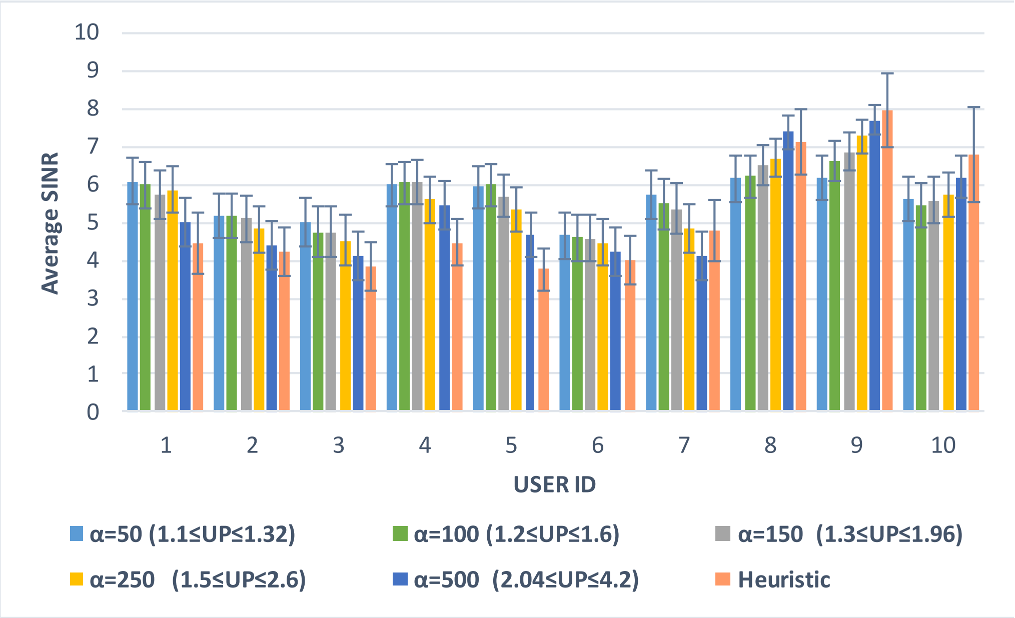


Figure 5-6: Users' SINR after user prioritisation (WSRMax Approach)

The results in Figure 5-6 depict an agreement in terms of the average SINR between the heuristic (5.1) and the MILP (ranged from 5.3 to 5.6 depending on the value of α). This approach slightly impacted the fairness between normal users as will be shown in the upcoming subsection. In this approach, the impact of converting the probability of stroke $PS^{z,r}$ (<<1) into a risk factor using α can be seen when comparing the users' average SINRs when α=50 to the ones associated with α=500. An OP (user 10) was granted an average SINR value very comparable to other healthy users (as in user 7) and sometimes less than the SINR of healthy users as the case with users 1, 4, 5, and 7. While that same OP had an average SINR higher than all healthy users when $\alpha = 500$ is used.

The average SINR of an individual user ranged between 4 and 7.6 for the MILP (α=500), and between 3.7 and 7.9 for the heuristic. A clearer illustration can be observed in Figure 5-6 where the confidence interval for each individual user's SINRs is shown.

5.5.1.3 The Impact of α on Fairness and SINR

The proposed model can be fine-tuned using the parameter α (i.e. tuning factor) introduced in equation (5-13). This parameter enables the reciprocity between the achievable fairness among users quantified by the SD and the average SINR. We examined the effect on the average SINR and the SD of using different values of α as illustrated in Figure 5-7 and in Figure 5-8.

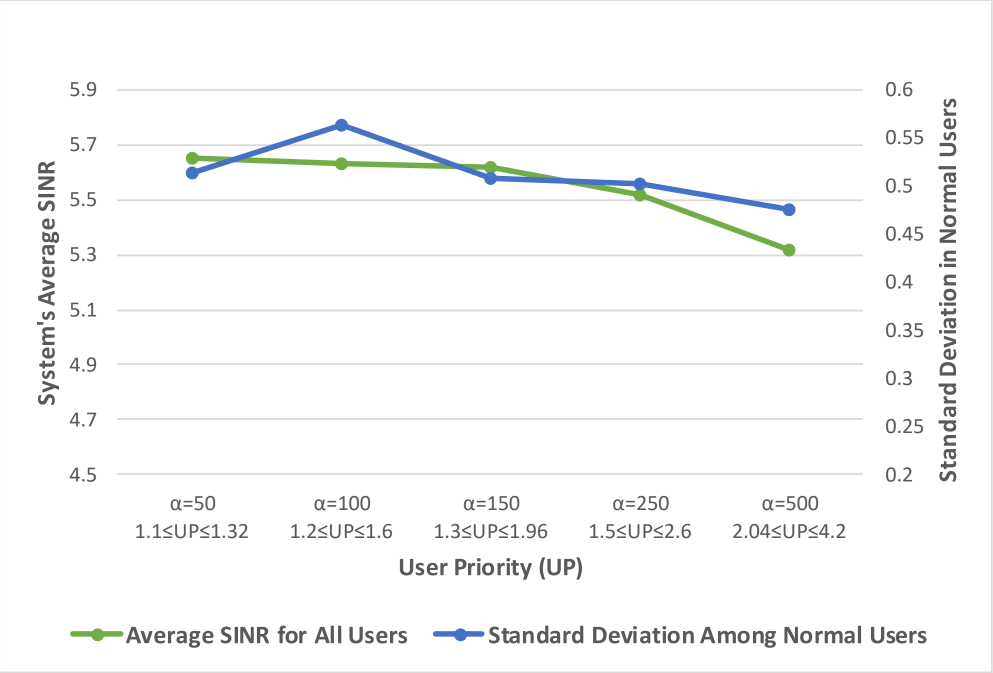


Figure 5-7: The effects of changing α on fairness and average SINR (WSRMax Approach)

Increasing the value of α directs the system to focus more on the OPs; consequently, a trade-off takes place resulting in lower values of the system's average SINR as seen in Figure 5-8 to increase the SINR of the selected users (i.e. the OPs), negatively affecting fairness as illustrated by the increasing SD in Figure 5-7.

It should be noted that the individual SINRs for the OPs correspond to the weights given to each OP using the NB Classifier. Sorting the users according to these weights produces an order that conforms to the values depicted in

Figure 5-8. The highest SINR was granted to user 9 which is the OP with the highest probability of stroke; thus, the highest priority, while the lowest among the three OPs was user 10 who also happened to be the one with the least priority among the OPs (nevertheless still higher than the normal users).

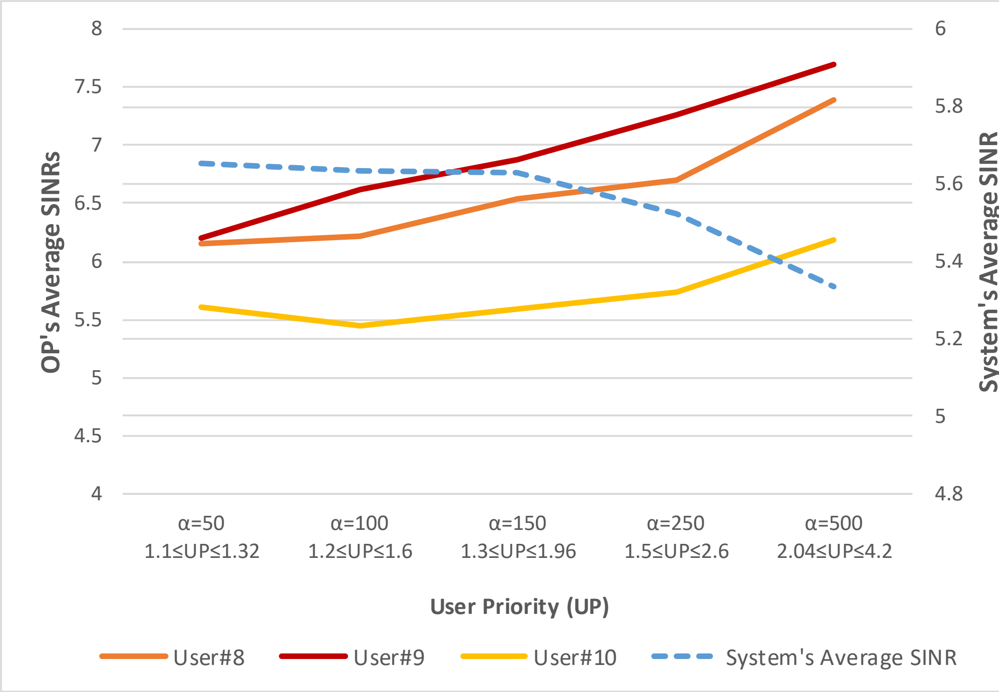


Figure 5-8: The impact of α on both user and average SINRs (WSRMax)

5.5.2 The PF Approach

5.5.2.1 Before Prioritising OPs

The objective function in (5-17) is applied to this scenario. The goal is to maximise the summation of the log of the users' SINRs while ensuring fairness without prioritising a certain subset of users. The results shown in

Figure **5-9** show a trend similar to the one depicted in Figure 5-5. However, due to the nature of the log function used in the objective function, fairness was maintained between the users (SD of 0.3 and 0.4 for the MILP and the heuristic, respectively), while the total SINR was reduced by 7% compared to the one produced by the MILP in the WSRMax approach.

The average SINRs for the heuristic and the MILP approaches are comparable at around 5.1 and 5.3, respectively. Sensitivity analysis was performed (95% confidence interval) where the average SINR achieved by the MILP ranged between 4.4 and 6.1, and between 4.1 and 6.4 for the heuristic results.

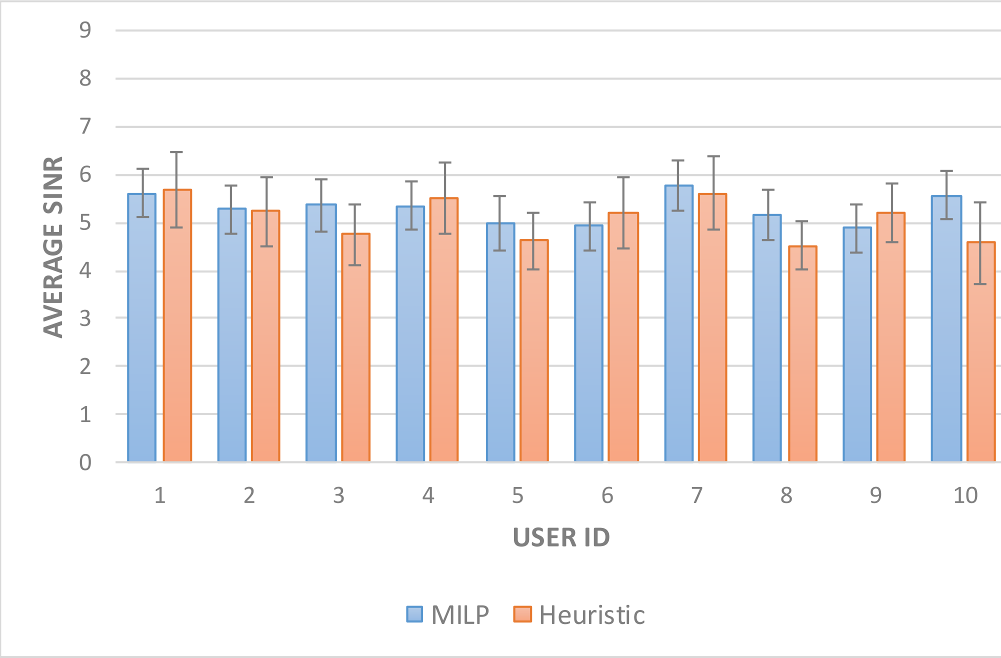


Figure 5-9: Users' SINR before user prioritisation (PF Approach)

5.5.2.2 After Prioritising OPs

In this scenario, the OPs' priorities (i.e. weights) are updated according to the stroke likelihood determined through the use of BDA. The objective function in (5-20) is used; consequently, the model grants the OPs high powered PRBs as can be noted in Figure 5-10. Comparing the PF approach to the WSRMax approach, it is evident that this approach grants the OPs higher SINRs (traded off with the other users). Furthermore, this approach shows higher conformance between the heuristic and MILP than the previous one. However, this was accomplished by trading off the average SINR. The MILP scored an average SINR between 5.2 ($\alpha = 50$) and 4.9 ($\alpha = 500$) as can be seen in Figure 5-10, while the heuristic's average SINR is 5.1. In this approach, the impact of different risk factor values on the OPs is less in comparison with the WSRMax approach due to the use of the natural logarithm causing the SINR to reduce in favour of the OPs. Nevertheless, an increase in the average SINR can also be noted among the OPs as depicted in Figure 5-10.

Narrower confidence intervals can be noted when employing this approach. As a matter of fact, this is a good indication of the precision of the approach in hand, thus producing results with narrower margins of error than the previous approach.

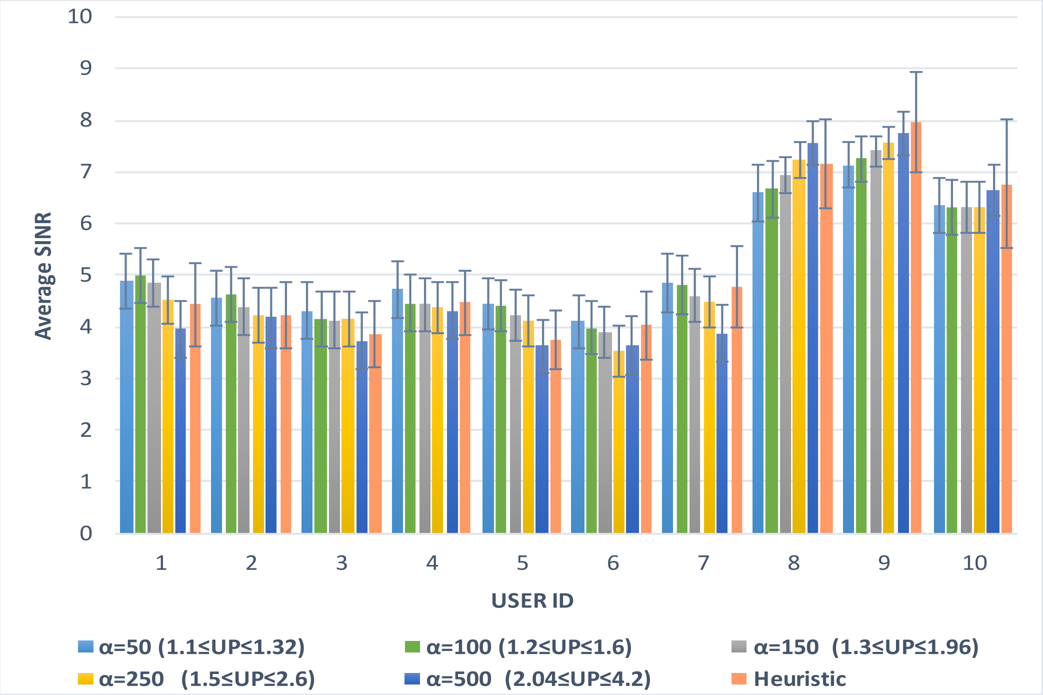


Figure 5-10: Users' SINR after user prioritisation (PF Approach)

5.5.2.3 The Impact of α on Fairness and SINR

Increasing the weights allocated to the OPs in this approach has similar effects to the ones in the previous section as shown in Figure 5-11 and in Figure 5-12. The reduction in the SINR is around 4%. However, the OPs were assigned higher SINRs. Furthermore, better fairness was reported among healthy users with an SD between 0.27-0.32 (depending on the value of α). Thus, offering a more stable approach.

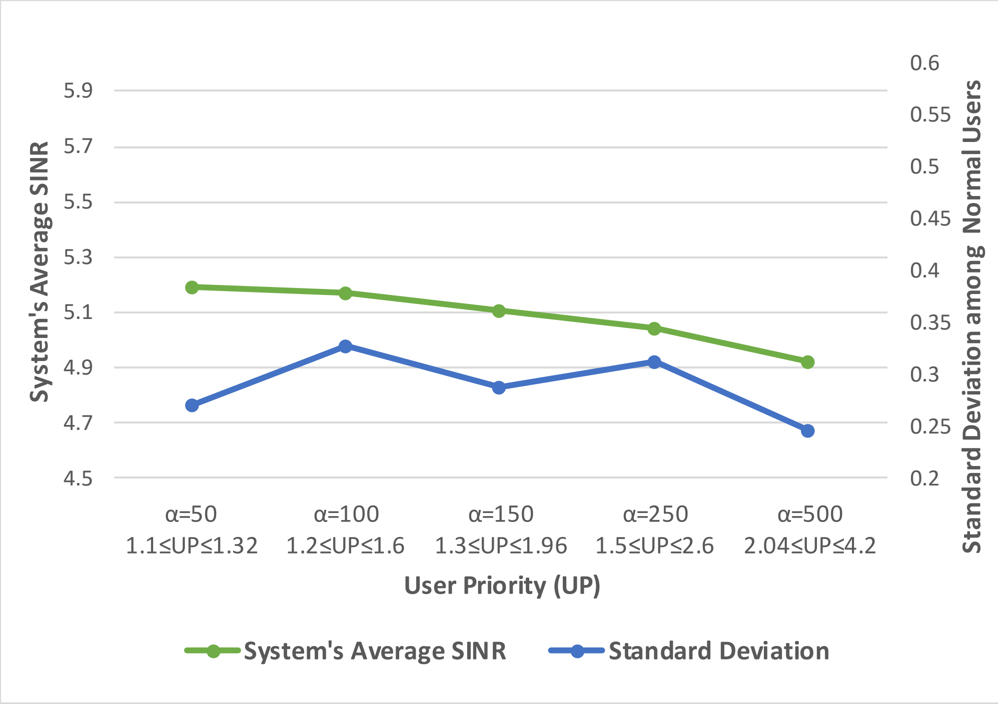


Figure 5-11: The effects of changing α on fairness and average SINR (PF Approach)

Further analysis of Figure 5-6 and Figure 5-10 reveals that the SINR sum achieved by the WSRMax approach is larger than that of the PF approach.

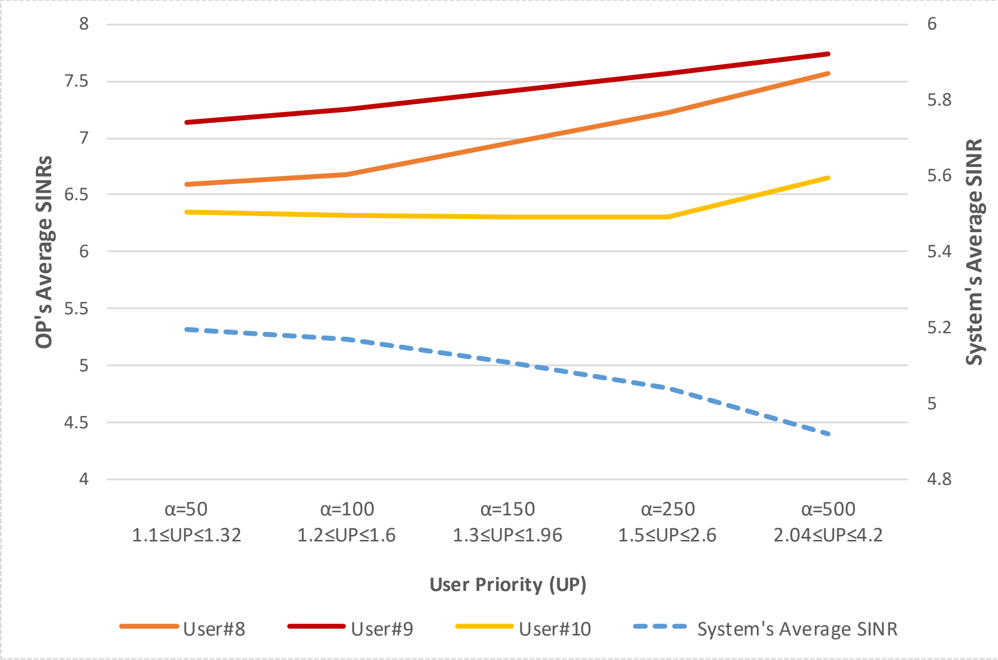


Figure 5-12: The impact of α on both user and average SINRs (PF Approach)

Since the WSRMax target is to maximise the sum rate (which is what an unregulated operator tries to do) while the PF approach introduces fairness, hence resources are not all allocated to the user with the best channel. The PF approach improves fairness but reduces the sum rate (which is the case of a regulated operator).

5.5.3 Testing the Heuristic's Scalability

Employing higher LTE-A system bandwidths enables the operator to serve more users creating a challenge for the developed heuristic to allocate resources to OPs with minimum delay to serve their urgent needs. To evaluate the scalability of the heuristic, elapsed time is considered.

We considered a scenario with six cases where the system operates at bandwidths of 1.4, 3, 5, 10, 15, and 20 MHz and increased the number of users, where all PRBs are occupied. For each case, we measured the time it takes the heuristic to allocate all users appropriate PRBs. The heuristic elapsed time was measured using the MATLAB functions tic and toc. Time calculation was carried out using two platforms: a Windows 10 computer equipped with Intel core i5-4460 3.2 GHz quad-core processor and 16 GB of RAM, and cloud-based MATLAB provided by MathWorks. The latter offers a measurement reference where calculations are made by relying on cloud-based resources, where such cloud resources are expected to play a key role in the control of future cellular networks. Given that it can take a stroke-suffering OP up to 8 hours before being administered with an anaesthetic, this heuristic's performance meets the requirements of this application. However, testing the heuristic's scalability in terms of other, more time-critical, applications is beyond the scope of this work. Figure 5-13 illustrates the heuristic's total elapsed time (in seconds) for both calculation methods versus the number of users. It should be noted that the worst-case scenarios are also considered and depicted in Figure 5-13.

The proposed heuristic tries to serve K users to be allocated to $K/2$ PRBs on each of the two BSs with another loop dedicated to interferer allocation. The first run contains a search of total K possible interferers (before satisfying the condition $k \neq m$). This means it requires $O(N * \frac{N}{2} * 2 * N)$ time. Additionally, the MATLAB sort function requires $O(N \log N)$ time [190]. Thus, the overall complexity is $O(N^4 \log N)$. The proposed heuristic provided a reduction in the run time to solve the NP-Hard problem [18] with a slight sacrifice in the accuracy of the results.

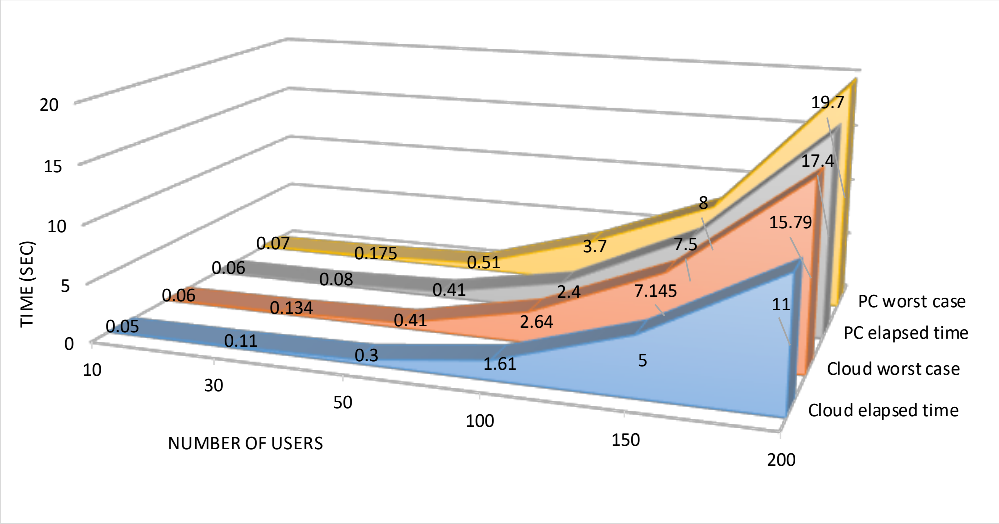


Figure 5-13: The Heuristic's Scalability

5.6 Chapter Summary

This chapter introduced a system that employs the power of BDA to optimise the uplink of an LTE-A cellular network. OP's medical record and readings from medical IoT sensors are processed in a BDA engine to find the likelihood of a stroke for an OP. The goal is to target OP users within the network to ensure they can always have access to the best wireless resources when in need. The proposed system achieves that with minimal impact on the wireless system-wide performance and SINR levels among healthy users in the network, thus improving the network utility for telecom operators while saving human lives and preserving fairness among normal users. Two approaches (WSRMax and PF) were presented and compared in terms of the average SINRs and fairness. The WSRMax approach improved the OPs' average SINR by up to 26.6%, whereas the PF approach increased them by 40.5%. The average SINR for normal users ranged between 5.5 and 4.6 using the WSRMax approach while the PF approach reported a range between 4.6 and 4 (depending on α). Fairness among users was quantified using SD. The WSRMax approach granted the healthy users SINRs with an SD between

0.47 and 0.56 (depending on α) while the PF approach ranged between 0.24 and 0.3 SD. Furthermore, we developed a heuristic to verify the MILP operation. The heuristic achieved comparable results to the MILP, and finally we demonstrated the heuristic's scalability.

Chapter 6 Using Machine Learning and Big Data Analytics to Prioritise Outpatients in HetNets

6.1 Introduction

In this chapter, we extend the work presented in the previous chapter by investigating the role of BDA to prioritise OPs according to their current health state in HetNet. Thus, providing, to the best of our knowledge, a novel incorporation of the topics of resource allocation, patient monitoring, disease risk prediction, and prioritisation in an optimisation model transforming **HetNets** to function in an OP-conscious manner. We use NB classifier to analyse data acquired from OPs' medical records, alongside data from medical IoT sensors that provide the current state of the OP. We use this ML algorithm to calculate the likelihood of a life-threatening medical condition, in this case an imminent stroke. An OP is assigned high-powered PRBs according to the seriousness of their current health state, enabling them to remain connected and send their critical data to the designated medical facility with minimal delay. Using a MILP formulation, we present two approaches to optimising the uplink of a HetNet in terms of user-PRB assignment: a WSRMax approach and a PF approach. Using these approaches, we illustrate the utility of the proposed system in terms of providing reliable connectivity to medical IoT sensors, enabling the OPs to maintain the quality and speed of their connection. Moreover, we demonstrate how system response can change according to alterations in the OPs' medical conditions.

6.2 System Model

We consider a HetNet comprised of a macro BS (MBS) and two neighbouring Pico BSs (PBSs) operating in an urban environment with a range of 40-100 meters. We assume that the network employs a spectrum partitioning strategy [191], and accordingly MBS users are not interfering with PBS users, hence, we consider here the intra-tier interference caused by users operating within the PBS range The users are randomly scattered and fall within two categories: healthy (normal) users, and

OPs as illustrated in Figure 6-1. Due to placing the users at random distances from the PBS, different power levels are received at the PBS from their UEs. If a low SINR channel is assigned to the OP, the health care provider may not be notified and the response may not arrive in time.

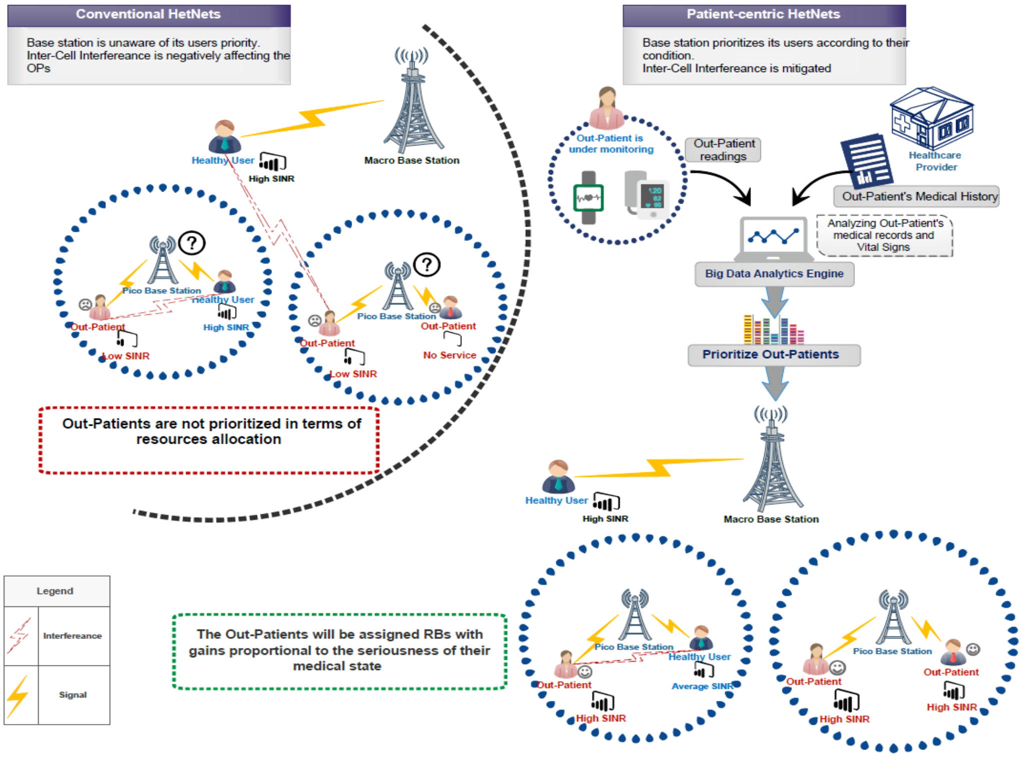
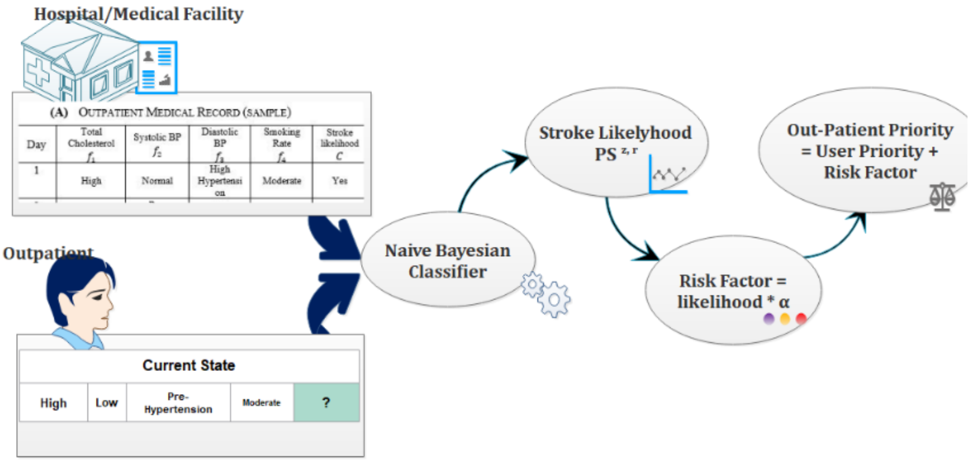

Conventional HetNets
Base station is unaware of its users priority.
Inter-Cell Interfereance is negatively affecting the OPs
Healthy User
High SINR
Macro Base Station
Pico Base Station
Healthy User
High SINR
Out-Patient
Low SINR
Pico Base Station
Out-Patient
Out-Patient
Low SINR
No Service
Out-Patients are not prioritized in terms of resources allocation
Patient-centric HetNets
Base station prioritizes its users according to their condition.
Inter-Cell Interfereance is mitigated
Out-Patient is under monitoring
Out-Patient readings
Healthcare Provider
Out-Patient's Medical History
Analyzing Out-Patient's medical records and Vital Signs
Big Data Analytics Engine
Prioritize Out-Patients
Healthy User
High SINR
Macro Base Station
Legend
Interfereance
Signal
The Out-Patients will be assigned RBs with gains proportional to the seriousness of their medical state
Pico Base Station
Healthy User
Average SINR
Out-Patient
High SINR
Pico Base Station
Out-Patient
Out-Patient
High SINR
High SINR

Figure 6-1: Patient-Aware HetNet

The goal is to allocate high-gain PRBs to OPs proportional to the severity of their medical status (i.e., stroke likelihood) as calculated in a cloud-located BDA engine according to the steps shown in Figure 6-2 (thus prioritising the OPs over normal users). OPs with high SINR values have greater spectral efficiency for their connection, because spectral efficiency is directly proportional to throughput, and the OPs will be able to send their data faster, hence minimising the delay.



Hospital/Medical Facility
Outpatient
Naive Bayesian Classifier
Stroke Likelyhood PS z, r
Risk Factor = likelihood * α
Out-Patient Priority = User Priority + Risk Factor

(A) OUTPATIENT MEDICAL RECORD (SAMPLE)

Day	Total Cholesterol f_1	Systolic BP f_2	Diastolic BP f_3	Smoking Rate f_4	Stroke likelihood C
1	High	Normal	High Hypertension	Moderate	Yes

Current State				
High	Low	Pre-Hypertension	Moderate	?

Figure 6-2: Outpatient Priority Calculation Procedure

6.3 Problem Formulation and Model Parameters

We developed a model to optimise PRB allocation in HetNets using MILP. Our scenario comprises a HetNet consisting of one MBS and two PBS. It is assumed that the network follows a spectrum partitioning strategy where Pico and macro users are on different PRBs (i.e., mitigating uplink inter-cell interference). Hence, interference occurs among Pico users only. Consequently, B PBSs are represented by the set $\mathcal{B} = \{1, \dots, B\}$. Each PBS has a total of N PRBs depicted by the set $\mathcal{N} = \{1, \dots, N\}$. A total of K users, both normal and OPs, represented by the set $\mathcal{K} = \{1, \dots, K\}$ are to be served in an instant of time by the PBSs using PRB n on PBS b. The target is to optimise the uplink of the network by maximising the overall system SINR while prioritising the OPs by allocating them high-gain PRBs.

We formalise this problem as a MILP model. Table 6-1 defines the sets, parameters, and variables used in the network optimisation problem formulation

Table 6-1: System Sets, Parameters, And Variables

Sets	
$\mathcal{K}$	Set of users.
$\mathcal{N}$	Set of physical resource blocks.
$\mathcal{B}$	Set of base stations.

$\mathcal{Z}$	Set of outpatient users,($\mathcal{Z} \subset \mathcal{K}$).
Parameters	
$\boldsymbol{CS_i}$	The current state of the patient in feature *i* (e.g. Cholesterol value).
$\boldsymbol{UP_k}$	User priority (UP_k =1 for normal users whereas $UP_k > 1$ is granted for OPs depending on their risk factor).
$\boldsymbol{Q^b_{k,n}}$	Power received from user k using PRB n at base station b.
$\boldsymbol{H^b_{k,n}}$	Rayleigh fading with zero mean and a standard deviation equal to 1 experienced by user k using PRB n at base station b.
$\boldsymbol{A^b_k}$	Signal attenuation experienced by user k connected to base station b.
$\boldsymbol{PM}$	Maximum power allowed per uplink connection.
$\boldsymbol{P}$	Power consumed to utilise PRB n to connect user k to base station b.
$\boldsymbol{\lambda}$	An arbitrary, large positive value.
$\boldsymbol{\sigma^b_{k,n}}$	Additive White Gaussian Noise (AWGN) power in watts experienced by user k using PRB n at base station b.
$\boldsymbol{PS^{z,r}}$	The probability of stroke of outpatient z.
$\boldsymbol{m_{y,k}}$ $\boldsymbol{h_{y,k}}$	Piecewise linearisation equation coefficients for line y of user k.
$\boldsymbol{\alpha}$	Tuning factor.
$\boldsymbol{NU}$	The total number of normal users.
Variables	
$\boldsymbol{X^b_{k,n}}$	Binary decision variable $X^b_{k,n} = 1$ if user k is assigned PRB n in base station b, otherwise $X^b_{k,n} = 0$.
$\boldsymbol{T^b_{k,n}}$	The SINR of user k utilising PRB n at base station b.
$\boldsymbol{\phi^{w,b}_{m,n,k}}$	Non-negative linearisation variable where $\phi^{w,b}_{m,n,k} = T^b_{k,n}X^w_{m,n}$.
$\boldsymbol{S_k}$	SINR of user k.
$\boldsymbol{L_k}$	Logarithmic SINR of user k.

The SINR $T^b_{k,n}$ of user k connecting to PBS b using PRB n is given as:

$$T^b_{k,n} = \frac{Signal}{Interference + Noise} = \frac{Q^b_{k,n}X^b_{k,n}}{\sum_{\substack{w\in\mathcal{B}\\ w\neq b}}\sum_{\substack{m\in\mathcal{K}\\ m\neq k}} Q^b_{\mathrm{m},n}X^w_{\mathrm{m},n} + \sigma^b_{k,n}} \quad (6\text{-}1)$$

The numerator in (6-1) depicts the *signal* part of the equation, whereas the denominator consists of two parts, *interference* received from users connected to other PBSs on the same PRB calculated as $Q_{\mathrm{m},n}^{b}X_{\mathrm{m},n}^{w}$ while the AWGN *noise* is represented by $\sigma_{k,n}^{b}$. $X_{k,n}^{b}$ is a binary variable equal to 1 when user k is connected to the PBS b using PRB n; $m, m \neq k$ and $w, w \neq b$ denote the interfering user(s) and interfering PBS(s), respectively. However, in our case there is a single interfering PBS. Rewriting equation (6-1):

$$\sum_{\substack{w\in\mathcal{B}\\ w\neq b}}\sum_{\substack{m\in\mathcal{K}\\ m\neq k}} T_{k,n}^{b}Q_{\mathrm{m},n}^{b}X_{\mathrm{m},n}^{w} + T_{k,n}^{b}\sigma_{k,n}^{b} = Q_{k,n}^{b}X_{k,n}^{b} \tag{6-2}$$

$\forall\, k \in \mathcal{K}, n \in \mathcal{N}, b \in \mathcal{B}$

The first term in (6-2) is nonlinear (quadratic) as it involves the multiplication of two variables (Continuous $T_{k,n}^{b}$ and Binary $X_{\mathrm{m},n}^{w}$). Therefore, linearisation is essential to solve the NP-hard model using a linear solver such as CPLEX, where the linearisation is given in (6-5) to (6-8).

We have developed two approaches to solve the resource allocation problem. The first approach uses an objective function that maximises the Weighted Sum-Rate of the SINRs experienced by the users. The second approach introduces fairness among the users by employing a PF objective function.

6.3.1 MILP Formulation for the WSRMax Model

The objective is to maximise the system's overall SINR. This can be realised through the maximisation of the individual users' SINRs.

6.3.1.1 Before Prioritising the OPs

The OPs' risk factors introduced in the previous chapters are scaled into priorities (i.e. weights) and used to prioritise the OPs over other users. The MILP model is formulated as follows:

Objective: Maximise

$$\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}}\sum_{b\in\mathcal{B}} T_{k,n}^{b}\, UP_{k} \tag{6-3}$$

The objective given in (6-3) aims to maximise the weighted sum of the users' SINRs. These weights (i.e. priorities) are higher for OPs compared to healthy users and proportional to the OPs calculated risk factor. Note that UP_k has an initial value of 1 for all users as shown in (6-4). However, the OPs will have updated values according to their risk factor. This will ultimately drive the system into prioritising the OPs over the healthy users during PRB assignment. The mathematical formulations related to the OP weight (priority) calculation was illustrated in Chapter 3.

$$UP_k = 1 \qquad \forall\, k \in \mathcal{K} \tag{6-4}$$

Constraints:

To maintain the model's linearity while performing the multiplication of the float variable $T_{k,n}^{b}$ by the binary variable $X_{m,n}^{w}$, we follow [187], and define a variable $\phi_{m,n,k}^{w,b}$ that includes all the indexes of both aforementioned (i.e., float and binary) variables as in (6-5). Constraints (6-6), (6-7), and (6-8) govern the multiplication procedure. As a result, the only two values satisfying the constraints are either zero (when x =0) or T (when x=1). It should be noted that **λ** is a large enough number where **λ** >>T:

Subject to:

$$\phi_{m,n,k}^{w,b} \geq 0 \tag{6-5}$$

Replacing the quadratic term $T_{k,n}^{b} X_{m,n}^{w}$ with the linearisation variable $\phi_{m,n,k}^{w,b}$ that incorporates all the indexes of the multiplied variables.

$$\phi_{m,n,k}^{w,b} \leq \lambda X_{m,n}^{w} \qquad \forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}, (m \neq k, b \neq w) \tag{6-6}$$

$$\phi_{m,n,k}^{w,b} \leq T_{k,n}^{b} \qquad \forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}, (m \neq k, b \neq w) \tag{6-7}$$

$$\phi_{m,n,k}^{w,b} \geq \lambda X_{m,n}^{w} + T_{k,n}^{b} - \lambda$$

$$\forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}\,, (m \neq k, b \neq w) \tag{6-8}$$

After replacing $T_{k,n}^{b} X_{m,n}^{w}$ with $\phi_{m,n,k}^{w,b}$, equation (6-2) can thus be rewritten as in (6-9). $\phi_{m,n,k}^{w,b} = T_{k,n}^{b} X_{m,n}^{w}$ is equal to the SINR of user k connected to BS b with PRB n if there is an interfering user m connected to the other BS w with the same PRB n; it is zero otherwise.

$$\sum_{\substack{w \in \mathcal{B} \\ w \neq b}} \sum_{\substack{m \in \mathcal{K} \\ m \neq k}} Q_{\mathrm{m},n}^{b} \phi_{m,n,k}^{w,b} + T_{k,n}^{b} \sigma_{k,n}^{b} = Q_{k,n}^{b} X_{k,n}^{b}$$

$$\forall\, k \in \mathcal{K}, n \in \mathcal{N}, b \in \mathcal{B} \tag{6-9}$$

$$\sum_{n \in \mathcal{N}} P\, X_{k,n}^{b} \leq PM$$

$$\forall\, k \in \mathcal{K}, b \in \mathcal{B} \tag{6-10}$$

Constraint (6-10) ensures that the users do not exceed their maximum available amount of power per uplink connections (in case more than one PRB is utilised by the same user k). In the current work, the user is allowed a single PRB.

$$\sum_{k \in \mathcal{K}} X_{k,n}^{b} \leq 1$$

$$\forall\, n \in \mathcal{N}, b \in \mathcal{B} \tag{6-11}$$

Constraint (6-11) limits the assignment of each PRB to one user only.

$$\sum_{b \in \mathcal{B}} \sum_{n \in \mathcal{N}} X_{k,n}^{b} \geq 1$$

$$\forall\, k \in \mathcal{K} \tag{6-12}$$

Constraint (6-12) guarantees that each user is assigned at least one PRB from any BS. Thus, no user is left without service. Additionally, this prevents the MILP from blocking interfering users to maximise the total SINR.

6.3.1.2 After Prioritising the OPs

In this approach, OPs' risk factors introduced in the previous chapter are scaled into weights to prioritise the OPs over other users. The MILP model is formulated in the same way as mentioned in the previous subsection. However, equation (6-13) is included in this model to represent the OPs' weights (i.e. priorities) as follows:

$$UP_k \; = \; 1 \; + \; \alpha \cdot PS^{z,r} \qquad \forall \, k \in \mathcal{K}: z = k, k > NU \tag{6-13}$$

while (6-4) is replaced by (6-14) to cover the normal users only.

$$UP_k = 1 \qquad \forall \, k \in \mathcal{K}: 1 \le k \le NU \tag{6-14}$$

6.3.2 MILP formulation for the PF Model

In this approach, the objective is to maximise the logarithmic sum of the user's SINRs. Due to the nature of the natural logarithm, a slight decrease in the overall SINR might be observed but to the expense of preserving fairness among normal users.

6.3.2.1 Before Prioritising the OPs

In this case, all users are treated equally, thus there is no prioritisation in terms of resource allocation. However, keeping fairness among users still holds as a necessity. Since the only part that we are dealing with is the value of the individual user's SINR, and to simplify the manipulation of the equation before adding the natural logarithm part, we present the optimisation variable S_k, to serve as the SINR for each user k.

$$S_k = \sum_{n \in \mathcal{N}} \sum_{b \in \mathcal{B}} T_{k,n}^{b} \qquad \forall \, k \in \mathcal{K} \tag{6-15}$$

Equation (6-15) replaces the three-indexed variable $T_{k,n}^{b}$ with a single-indexed variable S_k.

$$L_k = \ln S_k \tag{6-16}$$

$\forall\, k \in \mathcal{K}$

Equation (6-16) calculates L_k as a logarithmic function of the user's SINR S_k. Since the natural log is a concave function, and to preserve the linearity of our model, piecewise linearisation was used as depicted in constraint (6-18).

The objective is as shown in (6-17):

Objective: Maximise

$$\sum_{k \in K} L_k \tag{6-17}$$

Constraints:

In addition to constraints (6-5)-(6-12) from the previous model, the PF satisfies the following constraint

Subject to:

$$L_k \leq m_{y,k} * S_k + h_{y,k} \tag{6-18}$$

$\forall\, k \in \mathcal{K}$

Constraint (6-18) represents a set of piecewise linearisation relations implemented to linearise the concave function in equation (6-16). Note that constraint (6-18) corresponds to the line equation $y = mx + h$ where the line coefficients (i.e. $m_{y,k}$ and $h_{y,k}$) are selected as in [188]. It should be noted that the number of constraints used in the linearisation procedure is dictated by the total number of lines used to cover the linearised interval.

6.3.2.2 After Prioritising the OPs

In this case, the outpatients are prioritised. Equation (6-16) is rewritten to reflect the change.

$$L_k = \ln S_k$$

$$\forall\, k \in \mathcal{K}: 1 \leq k \leq NU \tag{6-19}$$

Equation (6-19) shows that the log function is applied to normal users only. The OPs, on the other hand, are assigned weights instead.

Objective: Maximise

$$\sum_{k\in K,1\le k\le NU} L_k + \sum_{k\in K,k>NU} S_k UP_k \tag{6-20}$$

The multi-objective function in (6-20) (i) maximises the sum of the SINRs allocated to all users, (ii) Assigns OPs priority by allocating OPs PRBs with high SINRs that reflect their relative priority, and (iii) Implements Fairness: by assigning healthy users PRBs with comparable SINRs. These objectives were implemented by adding both the summation of a log function of the healthy users' SINRs (i.e. Proportional Fairness) and the weighted sum of the OPs' SINRs (OPs priority).

Constraints:

The model satisfies constraint (6-5)-(6-12) from the previous approach. In addition to equation (6-14) and:

$$L_k \le m_{y,k} * S_k + h_{y,k}$$

$$\forall\, k \in \mathcal{K}, k \le NU \tag{6-21}$$

Constraint (6-21) represents the same set of equations for the piecewise linearisation that was used in constraint (6-18), however, the difference is in the range of users it is applied to

6.3.3 Calculating the received power

The received signal power (in Watts) $Q_{k,n}^{b}$ varies according to the channel conditions and the distance between the user and the BS. Considering Rayleigh fading denoted by $H_{k,n}^{b}$ and distance dependent path loss denoted by $A_{k,n}^{b}$ [19], the received signal power is given as:

$$Q_{k,n}^{b} = P\, H_{k,n}^{b} A_{k}^{b} \tag{6-22}$$

where $H_{k,n}^{b}$ denotes Rayleigh fading and A_{k}^{b} represents power loss due to attenuation (distance dependent path loss) and is given in (6-23) [19]:

$$A\ (dBm) = 140.7 + 36.7\ \log_{10} \frac{distance(meters)}{1000} \tag{6-23}$$

To unify the units, equation (6-24) is used to convert the power to Watts.

$$A\ (\mathrm{m}w) = 10^{\frac{A(dBm)}{10}} \tag{6-24}$$

6.4 Results and Discussion

In this section, we used the parameters in Table 6-2 for a scenario of a network employing a spectrum partitioning strategy. The results illustrate two approaches to identifying the resource allocation problem: the WSRMax and the PF. The first approach targets the maximisation of the weighted sum rate of all users' SINRs, with its objective in (6-3). The second, however, enforces fairness among users through its objectives in (6-17) and (6-20) by maximising the logarithmic sum of the users' SINRs. The MILP optimisation was performed using AMPL/CPLEX software running version 12.5 on a PC with 16 GB RAM and a core i5 CPU.

Table 6-2 : Model Parameters

Parameter	Description
System bandwidth	3 MHz
Total number of RBs	15
Channel Model	Path Loss [19] and Rayleigh fading [18]
Number of MBS	1
Number of PBS	2
Number of PRB per MBS	10
Number of PRBs per PBS	5
Number of users	10
Number of normal users (NU)	7
Number of OPs	3
AWGN ($\sigma_{k,n}^{b}$)	-162 dBm/Hz [19]
Distance between user k and BS b	(40 - 100) m
Maximum transmission power per connection	23 dBm [19]
UE transmission power per PRB	17 dBm
Base (i.e. normal user priority) weight	1

Outpatient priority UP_k calculation method	Naïve Bayesian classifier
OP observation period	30 Days
Weight Parameter (α)	50, 500, and 1000

Furthermore, we considered seven different *current states* in terms of input feature variables, as displayed in Table 6-3. We run each model over all seven different *current states* for 400 data files each representing randomised users' locations (i.e., random received power levels at the PBSs in each data file) simulating 400 instances and showing the average SINR. The seven current states produce different probabilities of strokes. These probabilities, along with different weight parameter α values, will be reflected as different SINR levels as shown in Figure 6-4 and Figure 6-6, respectively.

Table 6-3: Outpatient Current States

Instance	Features				Class
	Total Cholesterol f_1	Systolic Blood Pressure f_2	Diastolic Blood Pressure f_3	Smoking rate f_4	Stroke C
1	Normal	Pre-hypertension	Normal	Heavy	?
2	High	High Hypertension	Normal	Light	?
3	Normal	High Hypertension	High Hypertension	Moderate	?
4	High	High Hypertension	High Hypertension	Heavy	?
5	Normal	High Hypertension	Pre-hypertension	Light	?
6	Normal	High Hypertension	High Hypertension	Light	?
7	High	High Hypertension	High Hypertension	Light	?

It should be noted that to simplify the SINR calculation, we converted all logarithmic units (i.e., dBm) into linear scale (i.e., m Watt), hence the resulting average SINR values in Figure 6-3, Figure 6-4, Figure 6-5, and Figure 6-6 are unit less.

6.4.1 The WSRMax Approach

6.4.1.1 Before Prioritising the OPs

In this scenario, all users have equal priority (i.e., $UP_k = 1$). The average SINR is 830 (i.e., around 29 dB). However, observing the OPs (i.e., users 8, 9, and 10) in Figure 6-3, one can note that they have comparable SINRs to other (healthy) users, and sometimes actually lower, such as when comparing OPs 8 and 9 to user 7.

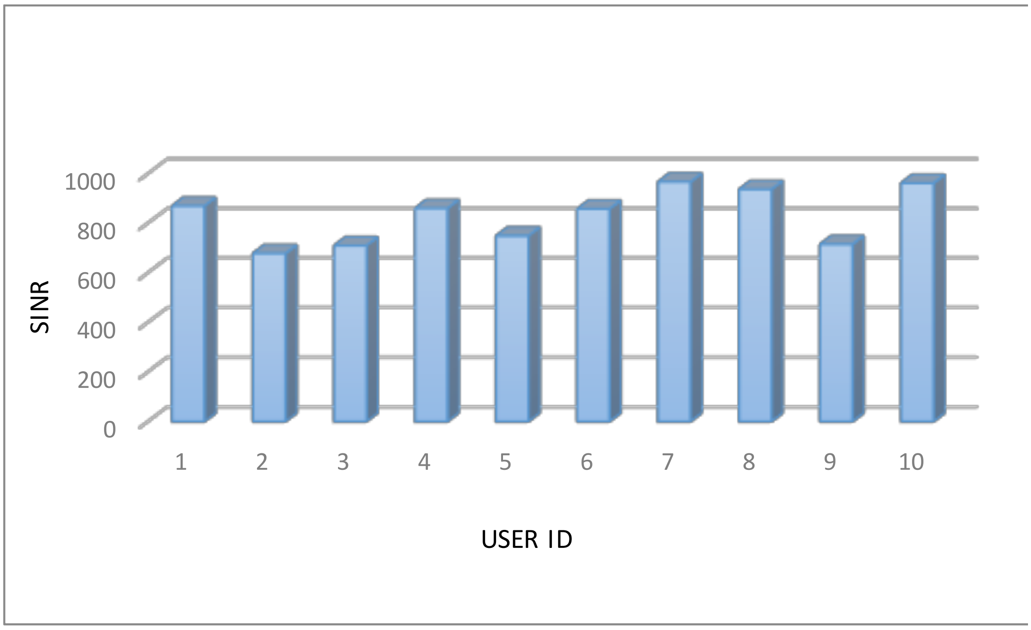
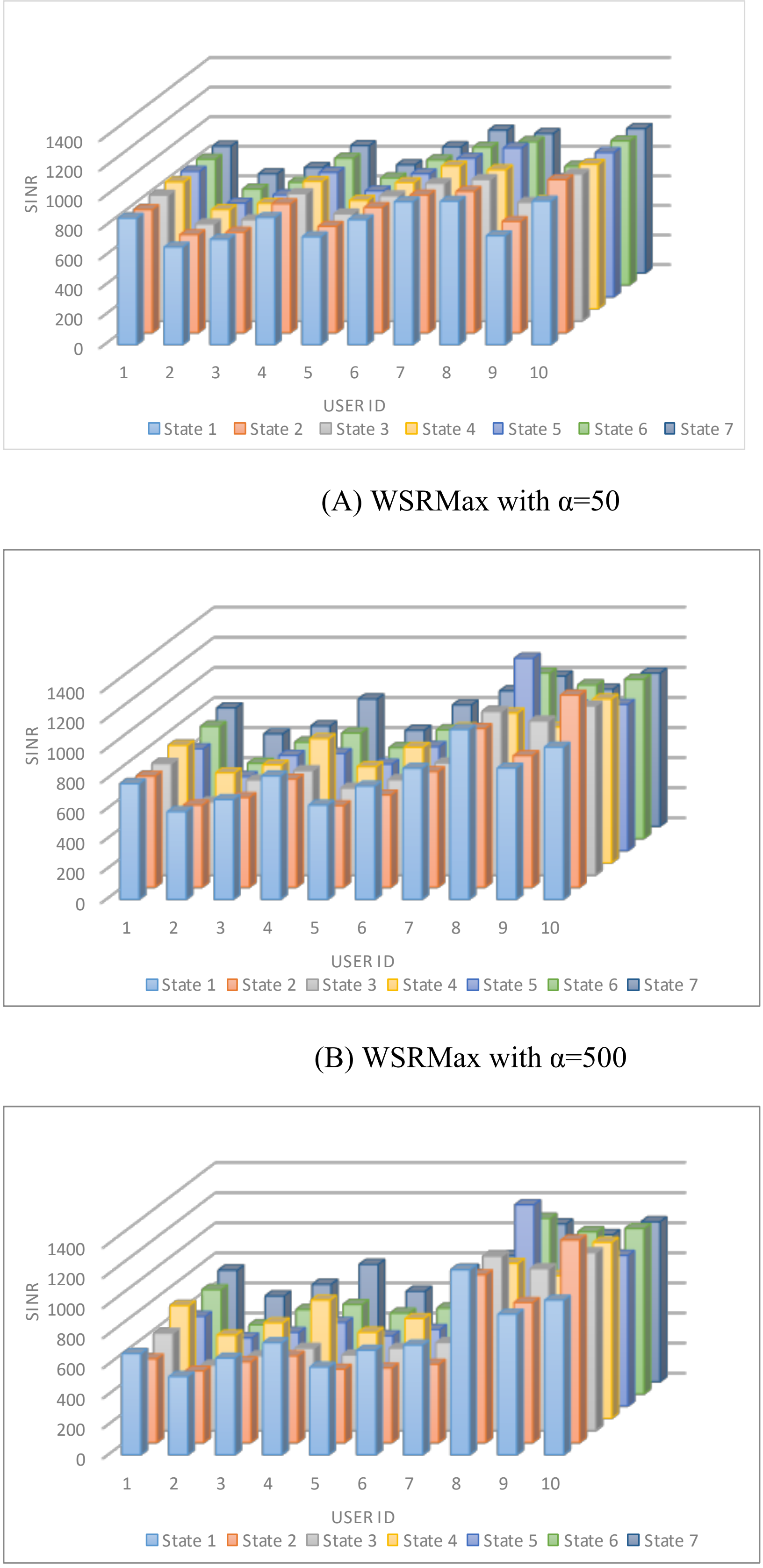


Figure 6-3: Users' SINRs before user prioritisation (WSRMax Approach)

6.4.1.2 After Prioritising the OPs

The OPs were granted high-gain PRBs according to their priority level. A negligible drop (0.3) in the average SINR is observed when selecting the weight parameter $\alpha = 50$. However, all OPs were granted above-average SINRs as shown in Figure 6-4 (A), (B), and (C). The OPs' SINRs increase with a focus on the OP with the highest priority in each state; moreover, we can notice that for $\alpha \geq 500$ all

OPs are assigned SINRs above the average, with 9% and 16% maximum SINR decrease when α =500 and 1000, respectively.



(A) WSRMax with α=50

(B) WSRMax with α=500

(C) WSRMax with α=1000

Figure 6-4: Users’ SINRs After user prioritisation (WSRMax Approach)

6.4.2 The PF Approach

6.4.2.1 Before Prioritising the OPs

The average SINR in this scenario is equal to 320 (around 25 dB) as illustrated in Figure 6-5. Users 9 and 10 are assigned less than the average SINR. A difference in the SINR levels can be observed between the two approaches. This is due to the use of the natural logarithm as well as the location of users with proximity to the PBS. When compared with the results in the previous chapter, we can clearly observe that the effect of the log differs. However, converting the SINRs to their logarithmic form (i.e., dB) shows that the SINR is still within the optimal range of operation.

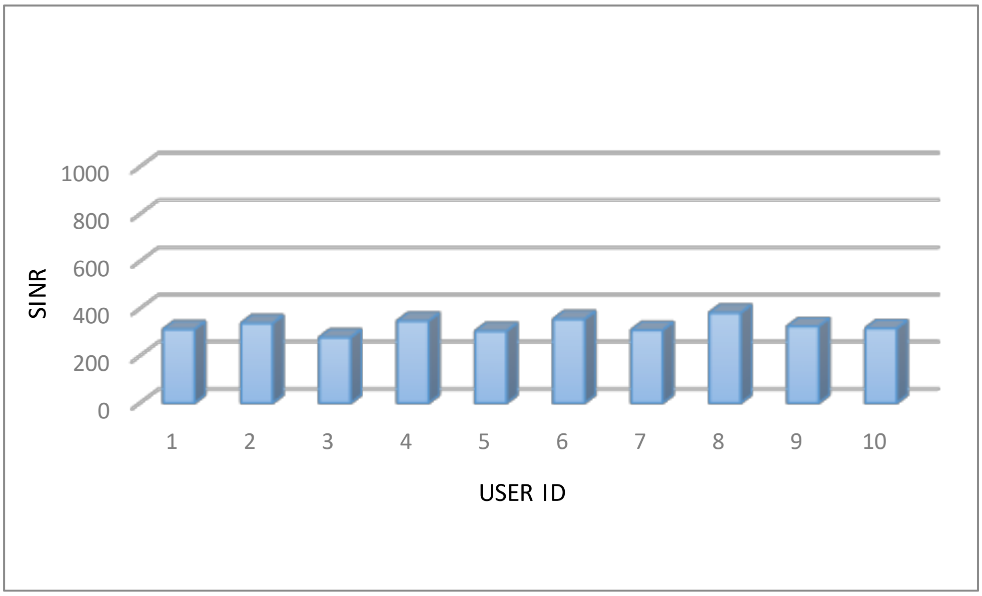
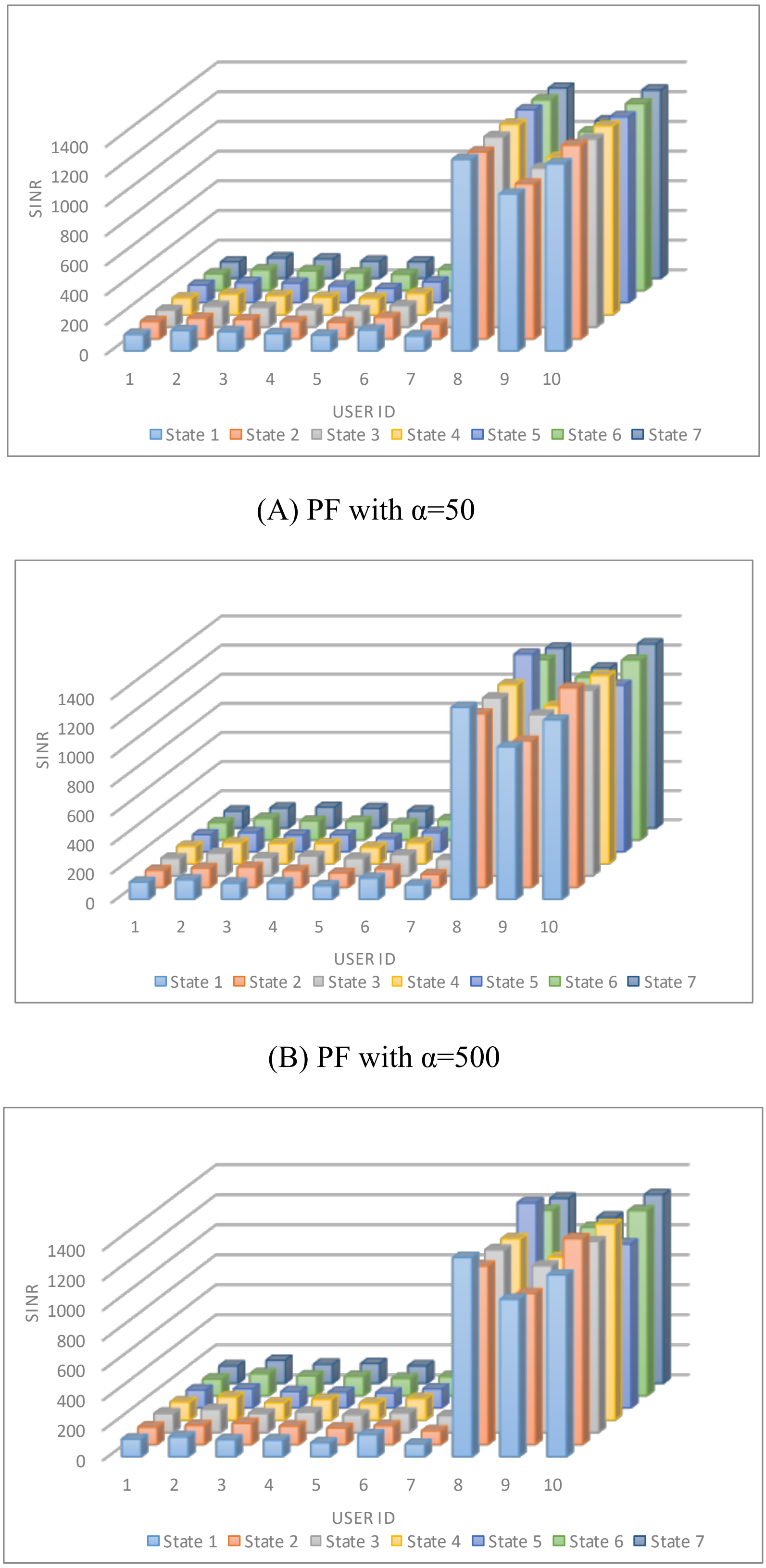


Figure 6-5: Users' SINRs before user prioritisation (PF Approach)

6.4.2.2 After Prioritising the OPs

In this scenario, the system's average SINR has increased due to the fact that only the normal users remain subjected to the logarithmic function. On the other hand, the OPs have high SINR levels, as shown in Figure 6-6 (A), (B), and (C). It should be noted that the effect of the increase of weight parameter α is minimal compared to the WSRMax approach.



(A) PF with α=50

(B) PF with α=500

(C) PF with α=1000

Figure 6-6: Users' SINRs After user prioritisation (PF Approach)

6.5 Chapter Summary

This chapter offered two multidisciplinary frameworks for patient-centric optimisation of HetNets. A BDA/ML algorithm was embedded in resource allocation optimisation and provided patient prioritisation in the e-health setting studied. The target is to prioritise stroke outpatients in HetNets according to their current medical condition based on readings acquired from body-attached and nearby IoT sensors. As a result, the developed ML-driven resource allocation frameworks granted these patients high-gain PRBs to ensure that they are always connected and can send their data with minimum delay. To that end, the WSRMax and PF approaches were presented and compared. The WSRMax approach maximises the OPs' SINRs with less impact on normal users when compared to the PF approach. The PF approach maximises the OPs' SINRs to a greater extent than the WSRMax approach, while a noticeable impact can be observed on normal users. With a false positive rate of 0.36, the current classifier can be further enhanced and compared to other algorithms to assess a patient's state, while the integration of more feature variables in a larger data set constitutes a basis for future work. Furthermore, investigating inter-cell interference as part of a larger model is currently being considered as a future direction.

Chapter 7
Beyond 5G: Patient-centric HetNets

7.1 Introduction

Having a cognitive and self-optimising network that proactively adapts not only to channel conditions, but also according to its users' needs can be one of the highest priorities of future HetNets. In this chapter, we introduce an interdisciplinary approach linking the concepts of e-healthcare, priority, radio resource optimisation, and BDA in a multi-tier 5G network. We employ three ML algorithms, namely, NB classifier, linear regression (LR), and decision trees (DT), working within an ensemble system to analyse historical medical records of stroke OPs and readings from body-attached IoT sensors to predict the likelihood of an imminent stroke. We convert the stroke likelihood into a risk factor functioning as a priority in MILP optimisation model. Hence, the task is to optimally allocate PRBs to HetNet users while prioritising OPs by granting them high gain PRBs according to the severity of their medical state. Thus, empowering the OPs to send their critical data to their healthcare provider with minimised delay. To that end, two optimisation approaches are proposed, the WSRMax approach and the PF approach. The proposed approaches increased the OPs' average SINR by 57% and 95%, respectively. The WSRMax approach increased the system's total SINR to a level higher than that of the PF approach, nevertheless, the PF approach yielded higher SINRs for the OPs, better fairness and a lower margin of error. The notion behind employing multiple ML algorithms is to; (i) check if the findings are consistent across different ML algorithms, (ii) select the optimal ML algorithm or set them to work together (which is what we proposed in the form of an ensemble system), (iii) the DT classifier fit the problem discrete nature, and (iv) the LR classifier is selected as it offers higher discrimination. However, it can have high sensitivity to feature vector noise, hence voting classifier is used.

7.2 System Model

In this work, we are considering a scenario of a HetNet consisting of a MBS and two neighbouring PBS operating in an urban environment. The MBS coverage range is 300-600 meters whilst the PBS has a range of 40-100 meters. In a previous work in [192], we assumed the adoption of a spectrum partitioning strategy [191] to mitigate the inter-tier interference on the PBS users caused by the MBS users. In this work, we are considering the effects of the inter-tier interference. The users belong to two categories: healthy (normal) users, and OPs as illustrated in Figure 7-1. As in a real-life scenario, the users are randomly scattered around the BSs at different distances which results in different received power levels at the BS from its UEs. If an OP is assigned a low-level SINR channel, the healthcare provider may not be notified and the response will be delayed. Here, a patient suffering a stroke loses 1.9 million neurons/min before the treatment starts [9]. Therefore, the objective is to assign high-gain PRBs to the OPs according to the severity of their medical status (i.e., stroke likelihood). The latter is computed in a cloud-based BDA engine according to the procedure shown in Figure 7-2. Thus, OPs that are prioritised over normal users will have higher spectral efficiency due to their high SINR values. This, in turn, will yield higher throughput (since spectral efficiency is directly proportional to throughout). Hence, the OPs will be able to send their data with minimal delay.

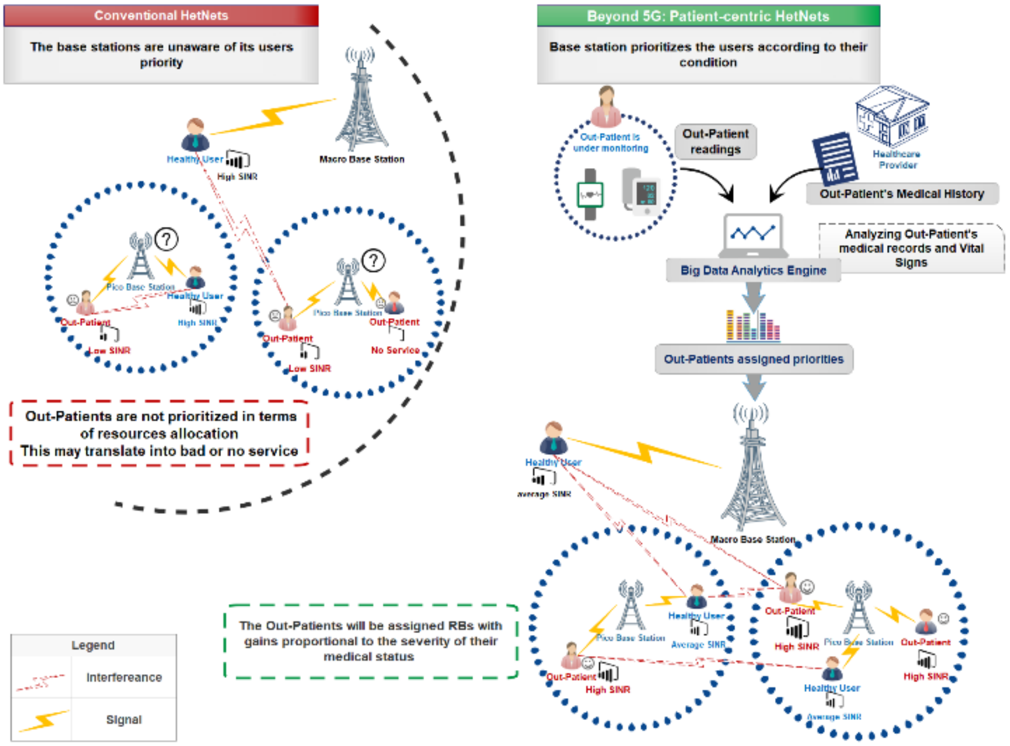
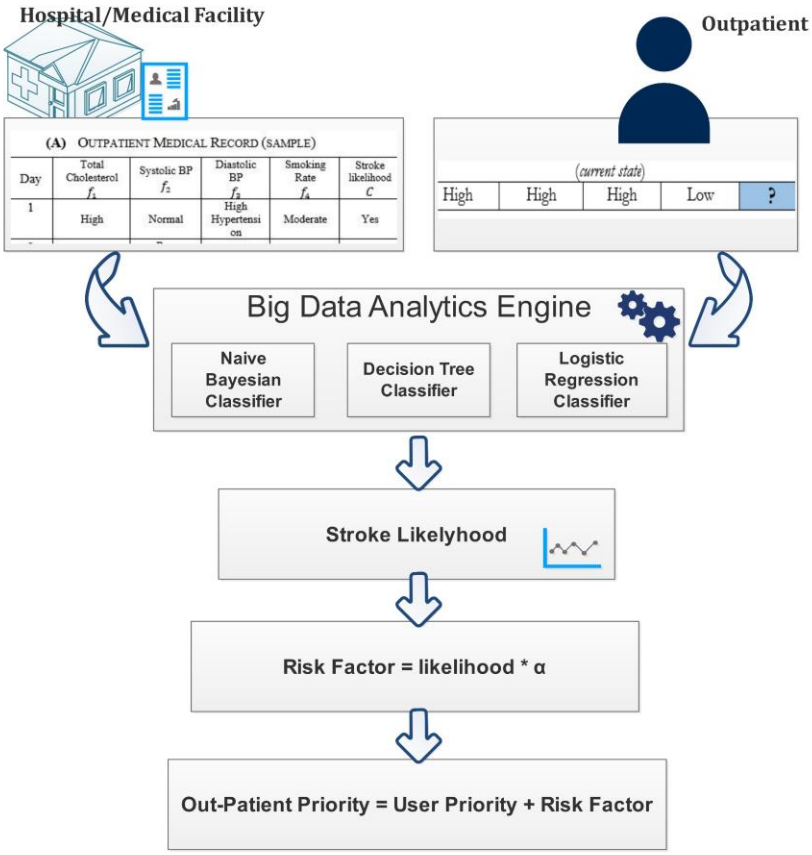

Conventional HetNets
The base stations are unaware of its users priority
Healthy User
High SINR
Macro Base Station
Pico Base Station
Healthy User
High SINR
Out-Patient
Low SINR
Pico Base Station
Out-Patient
Out-Patient
Low SINR
No Service
Out-Patients are not prioritized in terms of resources allocation
This may translate into bad or no service
Legend
Interfereance
Signal
The Out-Patients will be assigned RBs with gains proportional to the severity of their medical status
Beyond 5G: Patient-centric HetNets
Base station prioritizes the users according to their condition
Out-Patient is under monitoring
Out-Patient readings
Healthcare Provider
Out-Patient's Medical History
Analyzing Out-Patient's medical records and Vital Signs
Big Data Analytics Engine
Out-Patients assigned priorities
Healthy User
average SINR
Macro Base Station
Pico Base Station
Healthy User
Average SINR
Out-Patient
High SINR
Out-Patient
High SINR
Pico Base Station
Out-Patient
High SINR
Healthy User
Average SINR

Figure 7-1: Patient-Aware HetNet

In this work, we use an ensemble system comprising three supervised learning classifiers, namely, a NB classifier, a DT classifier, and a LR classifier that work on the OP's dataset and feed their predicted probabilities of stroke to a soft voting classifier. Given a certain feature vector (representing the OP's current state), each of the aforesaid classifiers yields a probability of stroke. Using ensemble learning, those classifiers can be combined into a single predictive model with higher accuracy, and thus, higher confidence is achieved in the predicted results.


Hospital/Medical Facility
Outpatient
Big Data Analytics Engine
Naive Bayesian Classifier
Decision Tree Classifier
Logistic Regression Classifier
Stroke Likelyhood
Risk Factor = likelihood * α
Out-Patient Priority = User Priority + Risk Factor

(A) OUTPATIENT MEDICAL RECORD (SAMPLE)

Day	Total Cholesterol f_1	Systolic BP f_2	Diastolic BP f_3	Smoking Rate f_4	Stroke likelihood C
1	High	Normal	High Hypertension	Moderate	Yes

(current state)

High	High	High	Low	?

Figure 7-2: Out-Patient Priority Calculation Procedure

7.2.1 Naïve Bayesian Classifier

The NB classifier is a probabilistic statistical classifier which uses a number of independent feature variables f_i (e.g. total Cholesterol and Blood pressure levels) obtained from a historical dataset (i.e., the OP's medical record) to determine the likelihood of an incident c (i.e. a stroke) as shown in Figure 7-2. The classifier is termed naïve because it assumes the feature variables are unrelated to each other [22]. This classifier is chosen for the following reasons; (i) it has a track record in disease risk prediction as in [158] and [193], (ii) its low complexity incur less computational burden, (iii) it is an ideal choice for any two-class concept with nominal features [160], (iv) it has proven accuracy in Cardio Vascular Disease (CVD) prediction compared to other approaches [166, 194], (v) it does not require large training datasets [159].

The classifier's *posterior probability* is given as

$$P(C = c|F_i = f_i) = P(C = c)\prod_{i=1}^{n} P(Fi = fi\ |C = c) \tag{7-1}$$

where $P(C = c)$ represents *the prior probability* of stroke, and the *likelihood* of F given C is given in (7-2)

$$P(F_i = f_i|C = c) = \frac{\sum_{i=1}^{n}(C = c \wedge F_i = f_i)}{\sum_{i=1}^{n}(C_i = C_i)} \tag{7-2}$$

where the term $\prod_{i=1}^{n} P(Fi = fi\ |C = c)$ depicts the *joint probability*.

7.2.2 Logistic Regression Classifier

The main distinctions between the NB classifier and the LR classifier is that it; (i) fast and a large change in response to the feature vector, (ii) it allows for large discrimination (i.e., a change in one feature may cause large effect). However, this also means that it suffers from high sensitivity to feature vector values. This classifier is a popular tool in disease prediction as in [195-197]. A *logistic model* is based on a mathematical form called the *logistic function* given in (7-3). This function equals zero when x is -∞, whereas the function equals 1 when x is + ∞.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{7-3}$$

This range is the primary reason for selecting the logistic model to estimate the probability. The index of combined features is x and it is given as a linear sum as shown in (7-4).

$$x = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + \ldots + \beta_n f_n \tag{7-4}$$

where β_0 represents the y intercept and $\beta_1 \ldots \beta_n$ are the regression coefficients, $f_1, \ldots f_n$ depict the feature variables, and n is the total number of features in the prediction model (in this work, $n = 4$) [198]. The conditional probability can be written as:

$$P(C = c|F_i = f_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n}\beta_i f i)}} \tag{7-5}$$

where $P(C = c|F_i = f_i)$ represents the conditional probability of a certain class variable $C = c$ given a feature vector FV. Therefore, if $C = 1$ then the conditional probability for $C = 0$ is $P(C = 0|F_i = f_i) = 1 - P(C = 1|F_i = f_i)$. The values of the line coefficients (i.e., $\beta_0 .. \beta_n$) cannot be solved analytically, therefore, we have to use solvers to navigate the search space.

7.2.3 Decision Trees Classifier

The DT construction procedure is done by splitting the dataset into descendant subsets. The splitting continues on repeated splits of the descendant subsets. The notion behind the tree methods is to have a set of partitions so that the best class can be determined. The partitions are performed so as to choose the splits in a way that guarantees that the leaves are *purer* than the parent node [199]. DT classifies vectors by sorting them, starting at the root of the tree down to some leaf nodes. In this tree, each node specifies a test of some input feature of the vector, and each branch descending from that node corresponds to one of the possible values for this feature The reasons for choosing DTs are; (i) their ability to implicitly perform feature selection or variable screening [168, 195, 200], (ii) they are uncomplicated to understand, interpret and, visualise, (iii) tree performance is not affected by nonlinear relationships between parameters, (iv) their track record in the stroke prediction literature as in [169, 201, 202] is good, where, in some cases, DTs yielded the highest accuracy.

The purity is measured using a *Gini index* which is used as an *attribute selection measure* where the ranking per attribute is given. The feature (attribute) with the best score is selected as the splitting feature for the given data subset. Splitting is done according to an impurity test conducted on a feature and a splitting subset (e.g., selecting two levels out of three $\{moderate, heavy\} \subset smoking$ or $\{moderate, heavy\} \subset V_{F_4}^r$ to be on a leaf while the remaining $\{low\} \subset V_{F_4}^r$ level is assigned to the other leaf). The binary split resulting in the maximum reduction in impurity (i.e., highest information gain) is selected as the splitting criterion. The *Gini* measure is given in (7-6).

$$Gini(\gamma) = 1 - \sum_{i=1}^{m} (p_i^\gamma)^2 \qquad (7\text{-}6)$$

where p_i^γ depicts the probability of a feature vector in training dataset γ belonging to class C_i^γ of a total number of m classes. The probability of an outcome of a certain class is given in (7-7) and the sum is calculated over m classes [203].

$$p_i^{\gamma} = \frac{|C_i^{\gamma}|}{|\gamma|} \tag{7-7}$$

It should be noted that the possible number of subsets is $2^{V_{F_i}^r} - 2$ (excluding the empty subset and the all $V_{F_i}^r$ subset), where $V_{F_i}^r$ represents the number of distinct values of feature F_i can have. However, in binary splits, this number is further reduced by omitting the cases where certain values are not included (e.g., assigning $\{moderate\} \subset V_{F_4}^r$ to one leaf and $\{\ heavy\} \subset V_{F_4}^r$ to another leaf and leaving the value $\{low\} \subset V_{F_4}^r$ unassigned. the weighted sum of the impurity is calculated for each resulting partition. Thus, if a feature F_i partitions the dataset γ into γ_1 and γ_2, then the Gini index of γ is given in (7-8).

$$Gini_{f_i}(\gamma) = \frac{|\gamma_1|}{|\gamma|} Gini(\gamma_1) + \frac{|\gamma_2|}{|\gamma|} Gini(\gamma_2) \tag{7-8}$$

The subset with the minimum impurity (i.e., Gini) for that feature is selected as its splitting subset. The same strategy is employed when using features with continuous values where each possible splitting point must be considered. Thus, extra computational resources will be required compared to the prior case.

The impurity reduction incurred by the binary split on feature F_i is given in (7-9).

$$\Delta Gini(f_i) = Gini(\gamma) - Gini_{F_i}(\gamma) \tag{7-9}$$

After forming the DT for an outpatient, the probability of a given vector of medical measurements is evaluated by tracing the decisions down the tree till the leaf where this vector belongs is reached. The probability in a given leaf is then evaluated as in (7-10).

$$P(C = c|F_i = \ f_i) = \frac{\Gamma_{z,C_i}}{\sum_{i=1}^{n} \Gamma_{z,C_i}} \tag{7-10}$$

where Γ_{z,C_i} denotes the number of samples in a leaf belonging to outpatient z having class C_i. The denominator represents the total number of samples of all classes in a given leaf.

7.2.4 Ensemble model

Ensemble methods train multiple learners on the same dataset to classify the same feature vector(s). The original goal of using ensemble systems is comparable to the way a person seeks advice from several trusted individuals. Hence, this reinforces the confidence that the decision made was the right one. Similarly, an ensemble of classifiers can be employed to increase the classification accuracy. Ensemble systems provide a method to incorporate various opinions, sometimes weighing them differently before reaching a concluding verdict. Individual classifiers may have different errors, however, they generally agree in terms of their their classification decision. Therefore, averaging the classifiers' outputs results in averaging the error component, and consequently reducing the classification error [204, 205] and balancing out the individual weaknesses of equally well-performing models [206]. The ensemble architecture of a soft voting (SV) classifier that we employed in this work is illustrated in Figure 7-3. The NBC, LR, and DT serve as base classifiers and their probabilities are then averaged to produce the voted probability denoted by P_{voting}. To calculate this probability, let the probability yielded by each base classifier CLF_i given in (7-1), (7-5), and (7-10) to be annotated as P_{CLF_1}, P_{CLF_2} and P_{CLF_3}, respectively. Since all base classifiers are treated evenly, the soft voting classifier calculates the probability as in (7-11).

$$P_{voting} = \frac{1}{|CLF_i|} \sum_{i=1}^{|CLF_i|} P_{CLF_i}(C = c|F_i = f_i) \tag{7-11}$$

where P_{voting} denotes the ensemble-calculated, averaged-conditional-probabilities.

In order to provide weights to the MILP so that the OPs are assigned higher gain PRBs, a base user priority UP_k of 1 is assigned to normal users while OPs are assigned the base weight *plus* another weight derived from the multiplication of a weight parameter α by the voted stroke likelihood P_{voting} thus, granting an effective-yet-reasonable priority.

$$\begin{gathered} UP_k = 1 + \alpha \cdot P_{voting} \\ \forall\, k \in \mathcal{K}: z = k, k \succ NU \end{gathered} \tag{7-12}$$

The OP's *updated* priority is given in (7-12). Using different values of α impacts the system response accordingly in terms of the OPs' SINR levels as shown in the results section.

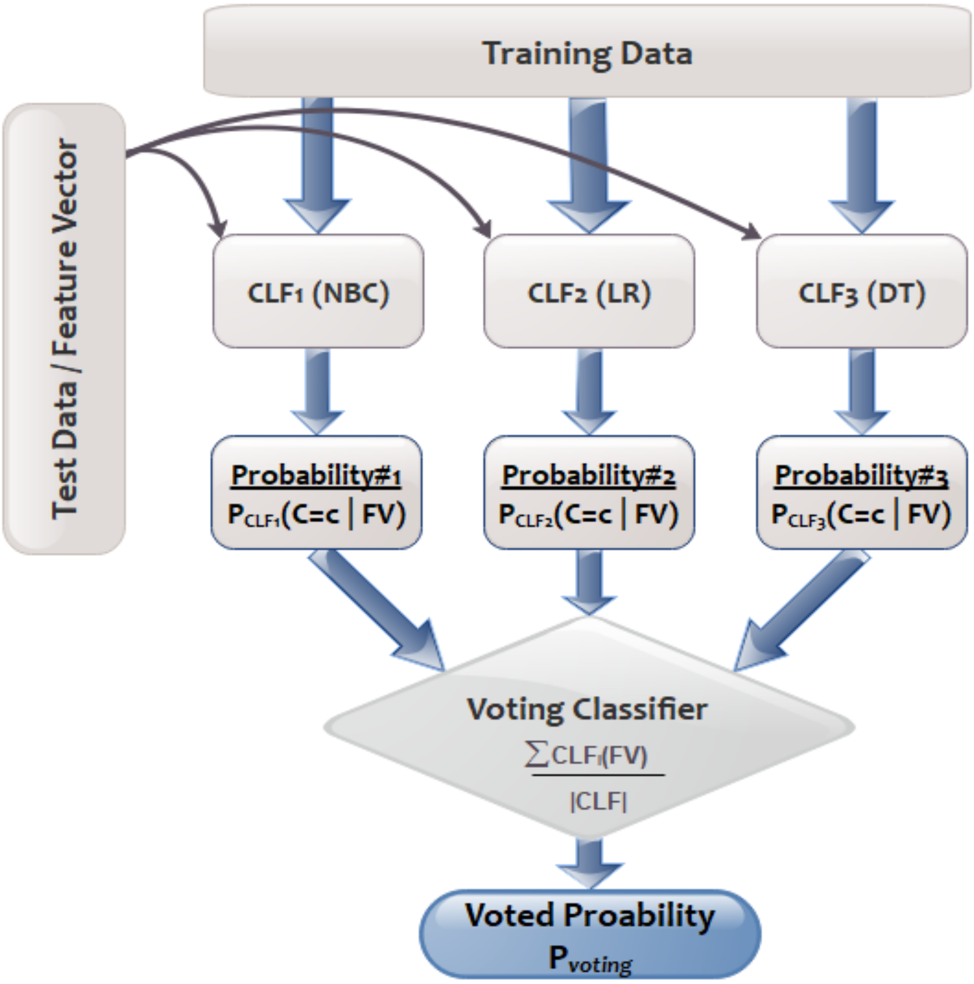


Figure 7-3: Ensemble Architecture

7.3 Problem Formulation

We developed the following MILP models to optimise the cellular system resource allocation for OPs and normal users. We consider the OPs monitoring system to operate in a scenario of a HetNet covered by B BSs denoted by the set $\mathcal{B} = \{1, \dots, B\}$ including both MBS and PBS types, operating at channels with 1.4 MHz bandwidth. Each BS b has N PRBs depicted by the set $\mathcal{N} = \{1, \dots, N\}$.

The network serves a total of K users (normal and OPs) denoted by set $\mathcal{K} = \{1, \dots, K\}$ by allocating PRB n to connect to BS b in an instant in time. The goal is to optimise the uplink of the HetNet, so that the OPs are prioritised over healthy users; hence, allocating them high-gain PRBs.

We formulate this problem as a MILP model. Table 7-1 defines the sets, parameters, and variables used in the network optimisation problem formulation.

Table 7-1: System Sets, Parameters, And Variables

Sets	
$\boldsymbol{\mathcal{K}}$	Set of users.
$\boldsymbol{\mathcal{N}}$	Set of physical resource blocks.
$\boldsymbol{\mathcal{B}}$	Set of base stations.
$\boldsymbol{\mathcal{D}}$	Set of days.
$\boldsymbol{\mathcal{F}}$	Set of features in the learning dataset.
$\boldsymbol{C}$	Set of classes in the learning dataset.
$\boldsymbol{\mathcal{Z}}$	Set of outpatient users,($\mathcal{Z} \subset \mathcal{K}$).
$\boldsymbol{CLF_i}$	Set of base classifiers
$\boldsymbol{V^r_{F_i}}$	Set of values that feature F_i can have in the learning dataset.
$\boldsymbol{V^r_{C_i}}$	Set of values a class variable C_i can take in the learning dataset.
Parameters	
$\boldsymbol{CS_i}$	The current state of the patient in feature i (e.g. Cholesterol value).
$\boldsymbol{UP_k}$	User priority (UP_k =1 for normal users whereas $UP_k > 1$ is granted for OPs depending on their risk factor).
$\boldsymbol{Q^b_{k,n}}$	Power received from user k using PRB n at base station b.
$\boldsymbol{H^b_{k,n}}$	Rayleigh fading with zero mean and a standard deviation equal to 1 experienced by user k using PRB n at base station b.

$\boldsymbol{A_k^b}$	Signal attenuation experienced by user k connected to base station b.
$\boldsymbol{PM}$	Maximum power allowed per uplink connection.
$\boldsymbol{P}$	Power consumed to utilise PRB n to connect user k to base station b.
$\boldsymbol{\lambda}$	An arbitrary, large positive value.
$\boldsymbol{\sigma_{k,n}^b}$	Additive White Gaussian Noise (AWGN) power in watts experienced by user k using PRB n at base station b.
$\boldsymbol{P_{votin}}$	The probability of stroke calculated at the voting classifier.
$\boldsymbol{m_{y,k}}$ $\boldsymbol{h_{y,k}}$	Piecewise linearisation equation coefficients for line y of user k.
α	Tuning factor.
$\boldsymbol{NU}$	The total number of normal users.
$\boldsymbol{\psi}$	The minimum SINR level.
Variables	
$\boldsymbol{X_{k,n}^b}$	Binary decision variable $X_{k,n}^b = 1$ if user k is assigned PRB n in base station b, otherwise $X_{k,n}^b = 0$.
$\boldsymbol{T_{k,n}^b}$	The SINR of user k utilising PRB n at base station b.
$\boldsymbol{\phi_{m,n,k}^{w,b}}$	Non-negative linearisation variable where $\phi_{m,n,k}^{w,b} = T_{k,n}^b X_{m,n}^w$.
$\boldsymbol{S_k}$	SINR of user k.
$\boldsymbol{L_k}$	Logarithmic SINR of user k.

The user's uplink SINR of an OFDMA network can be expressed as [18]:

$$T_{k,n}^b = \frac{Q_{k,n}^b X_{k,n}^b}{\sum_{\substack{w\in\mathcal{B}\\ w\neq b}} \sum_{\substack{m\in\mathcal{K}\\ m\neq k}} Q_{\mathrm{m},n}^b X_{\mathrm{m},n}^w + \sigma_{k,n}^b} \tag{7-13}$$

Examining the numerator (i.e. signal), $Q_{k,n}^b X_{k,n}^b$ signifies the signal power received at the BS from user k. $X_{k,n}^b$ is a binary decision variable, $X_{k,n}^b = 1$ denotes the connection of user k to PRB n in BS b. The power received at BS b from the interfering user(s) $m, m \neq k$, on the same PRB is $Q^{b}_{\mathrm{m},n}X^{w}_{\mathrm{m},n}$; while $X^{w}_{\mathrm{m},n}$ indicates an interfering user(s) m connected to another BS $w, w \neq b$ on PRB n. The AWGN is annotated as $\sigma^{b}_{k,n}$.

Rewriting equation (7-13):

$$\sum_{\substack{w \in \mathcal{B} \\ w \neq b}} \sum_{\substack{m \in \mathcal{K} \\ m \neq k}} T^{b}_{k,n} Q^{b}_{\mathrm{m},n} X^{w}_{\mathrm{m},n} + T^{b}_{k,n}\sigma^{b}_{k,n} = Q^{b}_{k,n} X^{b}_{k,n} \tag{7-14}$$

$\forall\, k \in \mathcal{K}, n \in \mathcal{N}, b \in \mathcal{B}$

The first term in (7-14) is nonlinear (quadratic) as it includes the multiplication of two variables (Binary $X^{w}_{\mathrm{m},n}$ and Continuous $T^{b}_{k,n}$). Hence, linearisation is vital to solve the model using a linear solver such as CPLEX, where the linearisation constraints are given in (7-17) - (7-20).

We have developed two approaches to solve the resource allocation problem. The first approach, named WSRMax, uses an objective function that maximises the Weighted Sum-Rate of the SINRs experienced by the users. The second approach implements fairness among cellular users by adopting a Proportionally Fair (PF) objective function.

7.3.1 Problem formulation for the WSRMax Model

In this approach, the objective is to maximise the system's overall SINR. This can be done by maximising the SINRs of individual users.

7.3.1.1 Before Prioritising the OPs

The OPs' risk factors introduced in the previous section are scaled into priorities (i.e. weights) and used to grant the OPs priority over other users. The MILP model is formulated as follows:

Objective: Maximise

$$\sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}} \sum_{b \in \mathcal{B}} T^{b}_{k,n}\, UP_{k} \tag{7-15}$$

The objective in (7-15) aims to maximise the weighted sum of the users' SINRs. The OPs have higher weights (i.e. priorities) than other healthy users and these weights are relative to the OPs' calculated risk factor.

Note that all the users share the same initial priority (i.e., $UP_k = 1$) as in (7-16). However, the OPs will have updated values according to their risk factor. This will ultimately drive the system into prioritising the OPs over healthy users during PRB assignment. The mathematical formulations related to the OP weight (priority) calculation was illustrated in Subsection 7.2.1 .

At this stage, all the users share the same initial priority (i.e., weight) as in (7-16).

$$UP_k = 1 \qquad \forall\, k \in \mathcal{K} \tag{7-16}$$

Constraints:

To ensure that the model holds its linearity while carrying out the multiplication of the binary variable $X_{m,n}^{w}$ by the float variable $T_{k,n}^{b}$, we follow [187], and define a variable $\phi_{m,n,k}^{w,b}$ that includes all the indexes of both aforementioned (i.e., binary and float) variables as in (7-17). Constraints (7-18), (7-19), and (7-20) govern the multiplication procedure. As a result, the only two values satisfying the constraints are either zero (when x =0) or T (when x=1). Note that **λ** is a large enough number where **λ** >>T:

Subject to:

$$\phi_{m,n,k}^{w,b} \geq 0 \tag{7-17}$$

The quadratic term $T_{k,n}^{b} X_{m,n}^{w}$ is replaced with the linearisation variable $\phi_{m,n,k}^{w,b}$ that incorporates all the indexes in the prior term.

$$\phi_{m,n,k}^{w,b} \leq \lambda X_{m,n}^{w} \qquad \forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}, (m \neq k, b \neq w) \tag{7-18}$$

$$\phi_{m,n,k}^{w,b} \leq T_{k,n}^{b} \qquad \forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}, (m \neq k, b \neq w) \tag{7-19}$$

$$\phi_{m,n,k}^{w,b} \geq \lambda X_{m,n}^{w} + T_{k,n}^{b} - \lambda \tag{7-20}$$

$\forall\, k, m \in \mathcal{K}, n \in \mathcal{N}, w, b \in \mathcal{B}, (m \neq k, b \neq w)$

After replacing $T_{k,n}^{b} X_{m,n}^{w}$ with $\phi_{m,n,k}^{w,b}$, equation (7-14) is rewritten as in (7-21). $\phi_{m,n,k}^{w,b} = T_{k,n}^{b} X_{m,n}^{w}$ equates the SINR of user k with PRB n connected to BS b if there is an interfering user m connected to the other BS w with the same PRB n; otherwise, it is zero.

$$\sum_{\substack{w \in \mathcal{B} \\ w \neq b}} \sum_{\substack{m \in \mathcal{K} \\ m \neq k}} Q_{\mathrm{m},n}^{b} \phi_{m,n,k}^{w,b} + T_{k,n}^{b} \sigma_{k,n}^{b} = Q_{k,n}^{b} X_{k,n}^{b} \qquad (7\text{-}21)$$

$\forall\, k \in \mathcal{K}, n \in \mathcal{N}, b \in \mathcal{B}$

$$\sum_{n \in \mathcal{N}} P\, X_{k,n}^{b} \ \leq PM \qquad (7\text{-}22)$$

$\forall\, k \in \mathcal{K}, b \in \mathcal{B}$

Constraint (7-22) ensures that the users do not exceed their maximum allocated power per uplink connection (in case more than one PRB is utilised by the same user k).

$$\sum_{k \in \mathcal{K}} X_{k,n}^{b} \ \leq 1 \qquad (7\text{-}23)$$

$\forall\, n \in \mathcal{N}, b \in \mathcal{B}$

Constraint (7-23) restrict the allocation of each PRB to only one user.

$$\sum_{b \in \mathcal{B}} \sum_{n \in \mathcal{N}} X_{k,n}^{b} \ \geq 1 \qquad (7\text{-}24)$$

$\forall\, k \in \mathcal{K}$

Constraint (7-24) guarantees that each user is allocated at least one PRB from any BS. Thus, no user is left without service. Furthermore, this stops the MILP from blocking interfering users to maximise the overall (network-wide) SINR.

7.3.1.2 After Prioritising the OPs

In this approach, OPs' risk factors introduced in the previous section are scaled into weights to prioritise the OPs over other users. The MILP model is formulated in the same way as mentioned in the previous subsection. However, equation (7-12) is included in this model to represent the OPs' weights (i.e. priorities) while (7-16) is replaced by (7-25) to cover the normal users only.

$$UP_k = 1 \qquad \forall\, k \in \mathcal{K}: 1 \leq k \leq NU \tag{7-25}$$

7.3.2 Problem formulation for the PF Model

Maximising the logarithmic sum of the user's SINRs is the objective in this approach. A slight decrease in the overall SINR might be observed (due to the nature of the natural logarithm) but with the benefit of preserving fairness among normal users.

7.3.2.1 Before Prioritising the OPs

All users, in this case, are treated evenly, thus there is no prioritisation in allocating the radio resources. However, keeping fairness among users still holds as a necessity. Since the only part that we are dealing with is the value of the individual user's SINR, and to simplify the manipulation of the equation before adding the natural logarithm part, we introduce the optimisation variable S_k, to serve as the SINR for each user k.

$$S_k = \sum_{n \in \mathcal{N}} \sum_{b \in \mathcal{B}} T_{k,n}^{b} \qquad \forall\, k \in \mathcal{K} \tag{7-26}$$

Equation (7-26) introduces single-indexed variable S_k which replaces the three-indexed variable $T_{k,n}^{b}$.

$$L_k = \ln S_k \qquad \forall\, k \in \mathcal{K} \tag{7-27}$$

Calculating L_k as a logarithmic function of the user's SINR S_k is indicated in (7-27). Since the natural log is a concave function, and to maintain the linearity of our model, piecewise linearisation was employed as in constraint (7-29).

The objective of this approach is given in (7-28):

Objective: Maximise

$$\sum_{k \in K} L_k \tag{7-28}$$

Constraints:

In addition to constraints (7-17)-(7-24) from the previous model, the PF satisfies the following constraint

Subject to:

$$L_k \leq m_{y,k} * S_k + h_{y,k} \tag{7-29}$$

$$\forall\, k \in \mathcal{K}$$

Constraint (7-29) represents a set of piecewise linearisation relations implemented to linearize the concave function in (7-27). It should be noted that constraint (7-29) follows the linear relation $y = mx + h$ where the line coefficients (i.e., $m_{y,k}$ and $h_{y,k}$) are selected as in [188]. It is worth noting that the number of constraints used in the linearisation procedure is dictated by the total number of lines used to cover the linearised interval.

7.3.2.2 After Prioritising the OPs

The outpatients are prioritised in this case, and equation (7-27) is rewritten to reflect the change.

$$L_k = \ln S_k$$
$$\forall\, k \in \mathcal{K}: 1 \leq k \leq NU \tag{7-30}$$

Equation (7-30) shows that the log function is applied to normal users only. The OPs, on the other hand, are assigned weights instead.

Objective: Maximise

$$\sum_{k \in K, 1 \leq k \leq NU} L_k + \sum_{k \in K, k > NU} S_k UP_k \tag{7-31}$$

The multi-objective function in (7-31) (i) Assigns OPs priority by allocating the OPs PRBs with high SINRs reflecting their relative priority, (ii) maximises the sum of the SINRs assigned to all users, and (iii) achieves fairness: by assigning healthy users PRBs with comparable SINRs. These objectives were implemented by adding both the summation of a log function of the healthy users' SINRs (i.e. Proportional Fairness) and the weighted sum of the OPs' SINRs (OPs priority).

Constraints:

The model satisfies constraints (7-17)-(7-24) from the previous approach. In addition to equation (7-25) and:

$$L_k \leq m_{y,k} * S_k + h_{y,k} \quad \forall\, k \in \mathcal{K}, k \leq NU \tag{7-32}$$

Constraint (7-32) represents the same set of equations for the piecewise leanariation that was used in constraint (7-29), however, the difference is in the range of users it is applied to.

7.3.3 Calculating the Received Power

The received signal power (in Watts) $Q_{k,n}^{b}$ varies according to two elements. Namely, the distance between the user and the BS and the channel conditions. The received signal power at the BS is given in (7-33):

$$Q_{k,n}^{b} = P\, H_{k,n}^{b} A_{k}^{b} \tag{7-33}$$

where $H_{k,n}^{b}$ denotes Rayleigh fading and A_{k}^{b} represents power loss due to attenuation (distance-dependent path loss) [19] and is given by equations (7-34) and (7-35), for the MBS and PBS, respectively.

$$A\,(dBm) = 128 + 37.6\ \log_{10} \frac{distance\ to\ MBS(meters)}{1000} \tag{7-34}$$

$$A\,(dBm) = 140.7 + 36.7\ \log_{10} \frac{distance\ to\ PBS(meters)}{1000} \tag{7-35}$$

Equation (7-36) is used to unify the units by converting the power to Watts, thus

$$A\,(\mathrm{m}w) = 10^{\frac{A(dBm)}{10}} \tag{7-36}$$

7.4 Results and Discussion

We consider a HetNet serving an urban environment, hence the Rayleigh fading channel model with path loss. The results evaluate two scenarios; the first depicts the HetNet state before prioritising the OPs. In this scenario, equal base priority (i.e., weight) of 1 is granted to all users. The second scenario shows the HetNet state after prioritising the OP through the updated priorities according to the value of the tuning factor α and their voted stroke likelihood.

A cloud-based arrangement is assumed where each OP has their personal dataset constructed from their medical history and daily observations over the course of 200 days, with the requirement to periodically extend the dataset by appending recent observations. Moreover, the proposed approach assumes a system that is in operation and the outpatient is being assessed by the voting system where multiple classifiers reside. We divided our dataset into two parts, a training set and a testing set, the training set comprised of 140 entries used to train/fit the classifiers, and the test set is 60 entries used to compare and verify the classifiers' performance. Furthermore, we would like to bring to the reader's attention that the ensemble's role in this work is to report the soft-voted stroke likelihood. Since the outpatients are all under continuous monitoring, they are favoured according to their probability of stroke as long as the system is operational. The OPs' stroke likelihood P_{voting} were 0.42, 0.84, and 0.65 for users 8, 9 and 10 (i.e., OP 1, 2, and 3), respectively. Moreover, the use of equation (7-12) produced $1.42 \leq UP_k \leq 1.84,\ 1.84 \leq UP_k \leq 2.68,\ 3.1 \leq UP_k \leq 4.25,\ 5.2 \leq UP_k \leq 9.4$ user priorities according to tuning factor values of α of 1, 2, 5, and 10, respectively.

7.4.1 Classifiers Comparison and Evaluation

In this section, we investigate the performance of the methods described in the previous section. There are several performance matrices for ML algorithms and certain metrics are known by more than one name. Since we have a binary classification problem, we refer to a prediction as "positive" if a classifier predicted $P(C = c|F_i = f_i) \geq 0.5$, indicating the occurrence of an *event* (e.g., stroke). Alternatively, if $P(C = c|F_i = f_i) < 0.5$ then the classifier predicted a *no-event* (e.g., no stroke), hence is translated as a "negative" prediction. In order to investigate the classifiers' performance, we use a test dataset of 60 entries where the outcome of all entries (i.e., feature vectors) are known (i.e., observed) to us and register the prediction results. Consequently, there will be four outcomes; (i) a correct positive prediction, named true positive (TP), indicating $P(C|F_i) \geq 0.5$ and an observed output of 1, (ii) an incorrect positive prediction, named false positive (FP), indicating $P(C|F_i) \geq 0.5$ and an observed output of 0, (iii) a correct negative prediction, named true negative (TN), indicating $P(C|F_i) < 0.5$ and an observed output of 0, and (iv) an incorrect negative prediction, named false negative (FN), indicating $P(C|F_i) < 0.5$ and an observed output of 1. The following matrices are computed through the use of these outcomes.

1. **Accuracy**, which is the ratio of true (i.e., correct) predictions to the total number in the dataset and is given in (7-37). Accuracy measures how well the classifier did in predicting the occurrence of an *event* as well as *no-event*.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \qquad (7\text{-}37)$$

2. **Sensitivity**, true positive rate, or recall, is the classifier's ability to pick an *event* of interest. Thus, accurately classifying actual positive values by labelling them as TP (i.e., stroke=1), and it is given in (7-38). In this work, it measures the classifier's ability to correctly classify an individual as *at-risk*.

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \qquad (7\text{-}38)$$

Sensitivity is a vital measure when the cost of FN prediction is high, in our case, if a high-risk outpatient is misclassified as low-risk (i.e.$stroke = 0$). Hence, the cost will be extremely high.

3. **Specificity** or true negative rate, is the classifier's ability to pick the occurrence of a *no-event* of interest. In other words, it is the classifier's ability to accurately identify actual negatives (i.e., stroke=0) in the test dataset. Thus, accurately classify an individual as *risk-free*.

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{7-39}$$

4. **Precision** or positive predictive value (PPV), it answers the question of how many of those who we predicted as at risk are actually at risk? Thus, it is the ratio of accurate positive predictions to the total number of positively-classified feature vectors, as in (7-40).

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{7-40}$$

Precision is a vital measure when the FP's cost is high. In our case, granting a priority to an outpatient that is not really in a high risk.

5. **Negative predictive value** (NPV), it answers the question of how many of those who we predicted as at no risk are actually not at risk? Thus, it is the ratio of feature vectors accurately classified as negative (i.e., TN) to the total number of classifications belonging to class $stroke = 0$, as denoted in (7-41).

$$NPV = \frac{TN}{TN + FN} \times 100\% \tag{7-41}$$

6. **False-positive rate** (FPR) or false alarm ratio represents the rate of misclassifying a class $stroke = 0$ as $stroke = 1$. It measures the frequency of false alarm and it is given in (7-42).

$$FPR = \frac{FP}{FP + TN} \times 100\% \tag{7-42}$$

7. **False-negative rate** (FNR) is a measure telling how erroneous a classifier can be in missing events (i.e., stroke=1). It is the ratio of misclassified positives to the total number of positives, as in (7-43).

$$FNR = \frac{FN}{FN + TP} \times 100\% \tag{7-43}$$

8. **F1 Score** is a function of both precision and recall values given in (7-40) and (7-38), respectively. This score is a measure of the balance between precision and recall as the former highly focuses on TPs, whilst the latter focuses on TNs. Thus, providing an equal weight for both precision and recall as it is the average (i.e., harmonic mean) of the two rates as given in (7-44).

$$F1\,Score = \frac{2.precision.recall}{precision + recall} \tag{7-44}$$

It should be noted that since there are three separate datasets (one per outpatient), hence, there are not only four classifiers to investigate, but also to examine the performance of these classifiers over three datasets as illustrated in Table 7-2.

The proposed SV classifier achieved higher accuracy compared to the other classifiers. Moreover, it had the lowest combined FPR and FNR which motivates its employment in this work. We further scrutinised the proposed SV classifier for the three OPs' datasets using 10-folds cross-validation and the results yielded 87.5%, 85.5%, and 88.5%, respectively.

Table 7-2: Comparing the machine learning methods.

OP#1 Training Dataset								
Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	NPV (%)	FPR (%)	FNR (%)	F1 Score
NB	82	76	88.5	90	74	11.5	24	83
LR	88	85	92	94	82.7	7.7	15	89
DT	90	88	92	94	86	7.6	11	91
SV	90	88	92	94	86	7.6	11	91
OP#2 Training Dataset								

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	NPV (%)	FPR (%)	FNR (%)	F1 Score
NB	76.7	68	84.4	79	75	15.6	32	73
LR	80	75	84.4	81	79	15.6	25	78
DT	80	64	94	90	75	6.2	36	75
SV	81.7	75	87.5	84	80	12.5	25	79
OP#3 Training Dataset								
Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	NPV (%)	FPR (%)	FNR (%)	F1 Score
NB	86.6	76	94.3	90	84.6	5.7	24	83
LR	90	88	91.4	88	91.4	8.6	12	88
DT	91.7	84	97.1	95	89.5	2.8	16	89
SV	93	88	97.1	96	92	2.8	12	92

7.4.1.1 Demystifying Performance Matrices

While it is significant to scrutinise the classifiers at hand and verify their performance. However, given the nature of our work, there are several performance matrices that are more vital than others. Hence we are highlighting their importance in this section. Accuracy is an important metric to our work due to the fact that it gives a balanced insights on the classifier's overall performance. FNR is the most important metric from the point of view of saving a patient's life, i.e., it tells us the proportion of ill people who is miss-classified. The F1-score takes misclassified entries (i.e., FP and FN) into account. Depending on the application, it can be equally as important as accuracy as in our case. F1-score gives

Before proceeding into the results of the MILP model, it worth noting that we used the parameters indicated in Table 7-3.

Table 7-3: Model Parameters

Parameter	Description
LTE-A system bandwidth	1.4 MHz
Channel Model	Path Loss [19] and Rayleigh fading [18]
No. of MBS	1
No. of PBS	2
Number of PRBs per BS	5
Number of users	10
Number of normal users (NU)	7
Number of OPs	3
AWGN ($\sigma_{k,n}^{b}$)	-162 dBm/Hz [19]
The distance between user k and MBS b	(300 - 600) m
The distance between user k and PBS b	(40-100) m
Maximum transmission power per connection PM	23 dBm [19]
UE transmission power per PRB	17 dBm
Minimum SINR defined for the reliability-aware PF approach (ψ)	21 dB [207]
Base (i.e. normal user priority) weight	1
Outpatient priority UP_k calculation method	Soft Voting Classifier
OP observation period	200 Days
α values	1, 2, 5, and 10

7.4.2 The WSRMax Approach

7.4.2.1 Before Prioritising the OPs

This scenario mimics the operation of a conventional HetNet where all users share the same *base user weight* (i.e. priority) of 1. The results in Figure 7-4 indicate that the OPs (represented by users 8, 9, and 10) are assigned PRBs of comparable gains resulting in near-average SINRs. This is due to the fact that the MILP's aim is aiming to maximise the HetNet's overall SINR. In order to measure fairness, we considered accentuating the Standard Deviation (SD) of the users' SINRs, hence, to quantify how close the calculated SINR values are to the mean, in this case, the SD was 195. Moreover, an extensive sensitivity analysis was carried out for the 300 independent realisations of the channel and the results with 95% confidence intervals per user are indicated in Figure 7-4. The average SINR lied between 2166 and 2691.

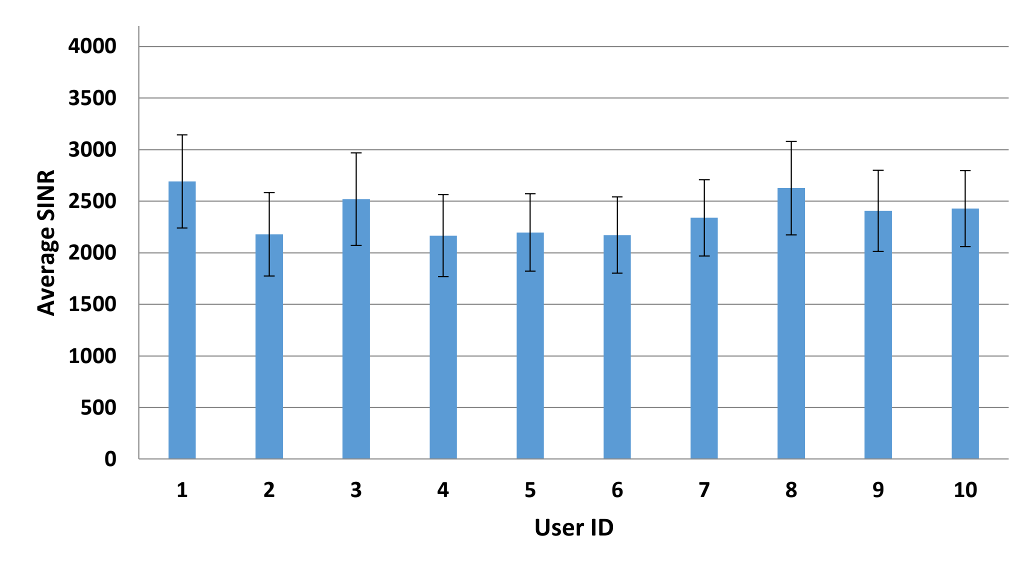


Figure 7-4: User SINR before OP Prioritisation (WSRMax Approach)

7.4.2.2 After Prioritising the OPs

The goal in this scenario is to utilise BDA/ML to prioritise the OPs over normal users by means of the ensemble system. As a result, high gain PRBs will be allocated to the OPs according to their risk factor, and guaranteeing them high-level SINRs. Comparing Figure 7-4 and Figure 7-5 clearly highlights that the OPs (i.e., users 8, 9, and 10) were granted PRBs with high SINRs. The overall system performance is a trade-off (*optimally-selected*) between guaranteeing the assignment of high SINRs to the OPs versus the decrease in the average SINR (between 2% ($\propto = 1$) and 19% ($\propto = 10$) in comparison to the SINR in the first scenario. The reduction in the average SINR is due to the system being was enforced to a PRB assignment scheme where the maximisation of the OPs' individual SINRs is prioritised over the total SINR. Fairness between normal users was marginally impacted in this approach as will be shown in the following subsection. The impact of converting the probability of stroke to a risk factor and using several values of the tuning factor (i.e.$\alpha = 1,2,5, and\ 10$) can be observed by comparing the increase in the OPs' average SINRs. Taking the case of user 9 (the most critical user with a probability of 0.84) having an SINR lower than users 1, 3, 8, and 10, the average SINR witnessed an increase from 17% ($\alpha = 1$) to 57% ($\alpha = 10$) granting this user an average SINR higher than all users. Individual users had an average SINR ranging from 1042 to 3776 for $\alpha = 10$.

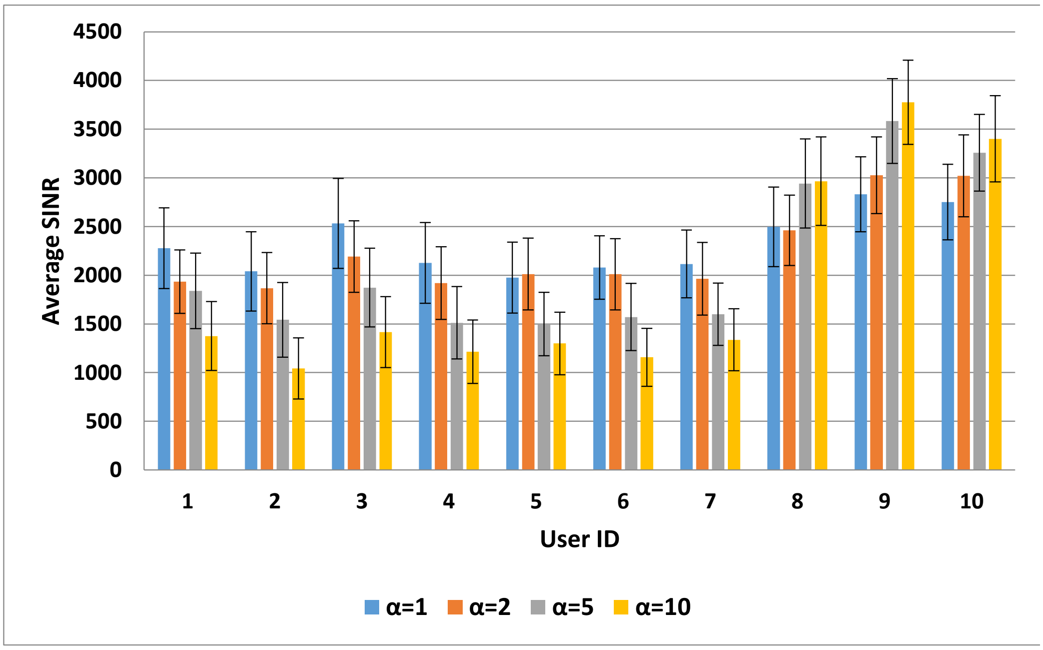


Figure 7-5: User SINR after OP Prioritisation (WSRMax Approach)

7.4.2.3 The Impact of α on Fairness and SINR

The parameter α is a tuning factor that is used to convert the minute value of the voted probability (i.e., P_{voting}) of stroke acquired from the ensemble system to a risk factor as depicted in equation (7-12). Moreover, this parameter enables the reciprocity between the average SINR and the attainable fairness among the users quantified by the SD. We used different values of α to study the effects on the SD and the average SINR. We examined the effects of using different vales of α on the SD and the average SINR as shown in Figure 7-6 and in Figure 7-7.

Increasing the value of α forced the system to concentrate on the OPs. Accordingly, the system's overall SINR was optimally traded-off to increase the OPs' SINRs while minimally impacting fairness among users as shown in Figure 7-6. It should be noted that examining the OPs' SINRs and comparing them against their corresponding risk factor values reveals an increase in the SINR in an order conforming to that depicted in Figure 7-7, where the PRB assignment granting the highest SINR was allocated to user 9 which is the user with the highest risk factor (priority). Furthermore, user 8 which has the lowest risk factor among the three OPs was given the lowest SINR among the OPs and very close to the system's average SINR. As the value of α increased (i.e., $\alpha = 5, 10$), user 8 is granted higher SINRs in comparison against other healthy users.

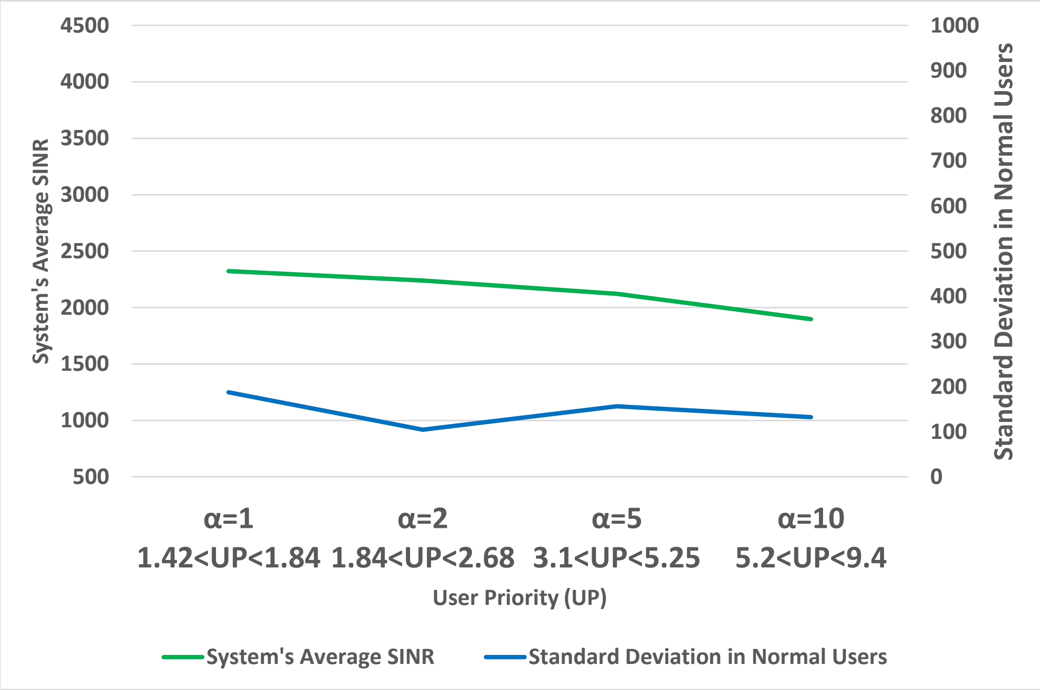


Figure 7-6: Effects of changing α on average SINR and fairness (WSRMax Approach)

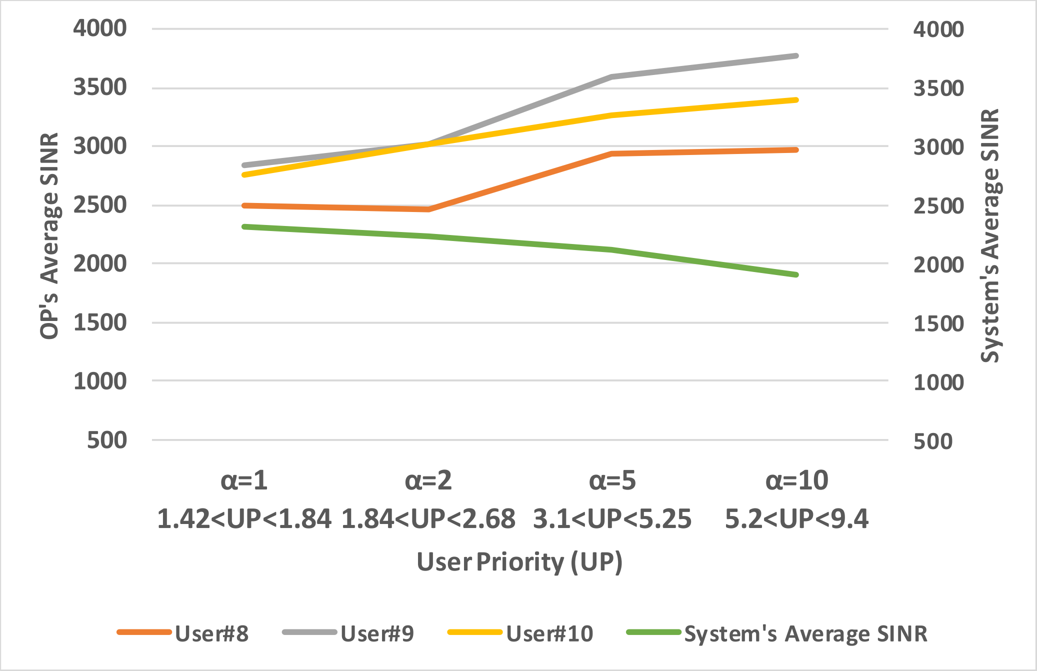


Figure 7-7: The impact of α both user and average SINR (WSRMax Approach)

7.4.3 The PF Approach

7.4.3.1 Before Prioritising the OPs

In this scenario, the goal is to maximise the logarithmic sum of the user's SINRs. Thus, no priority is given to any user in particular. Fairness is applied as a consequence due to the nature of the natural log in the objective function in (7-26. The results depicted in Figure 7-8 are in agreement with the ones depicted in Figure 7-4. However, a 46% reduction in the SD is reported when comparing this scenario and the one in Subsection 7.4.2.1 . The average SINR ranged between 1905 and 2251. Sensitivity analysis was implemented over 300 different realisations of the HetNet. The results with a 95% confidence interval are illustrated in Figure 7-8.

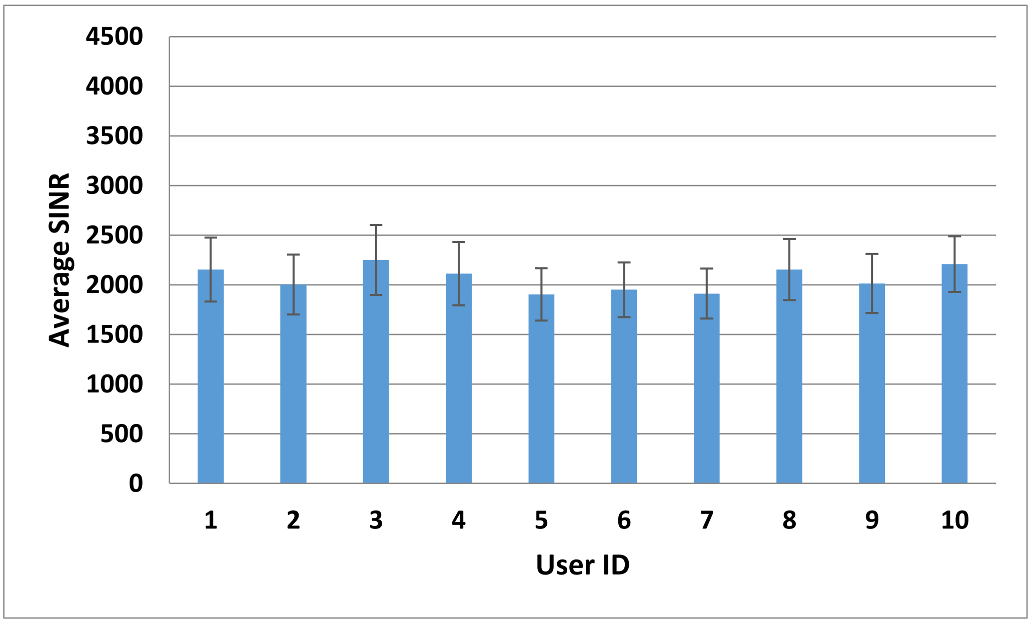


Figure 7-8: User SINR before OP Prioritisation (PF Approach)

7.4.3.2 After Prioritising the OPs

In this approach, the OPs are prioritised according to their risk factors using the objective function in (7-31). Therefore, the OPs are granted high-gained PRBs resulting in high SINRs as illustrated in Figure 7-9. The OPs' SINRs was boosted by up to 95% observed by user 9 with $\alpha = 10$. However, the average system SINR ranged between 1093 ($\alpha = 1$) and 1113 ($\alpha = 10$). The healthy users were noticeably affected by the intrinsic nature of the natural log, and the exclusion of the OPs from the logarithmic term in the objective function resulted in granting the healthy user lower SINRs in comparison to the OPs' SINRs. Figure 7-10 depicts the average users' SINR in a logarithmic scale where narrower confidence intervals can be observed in this approach.

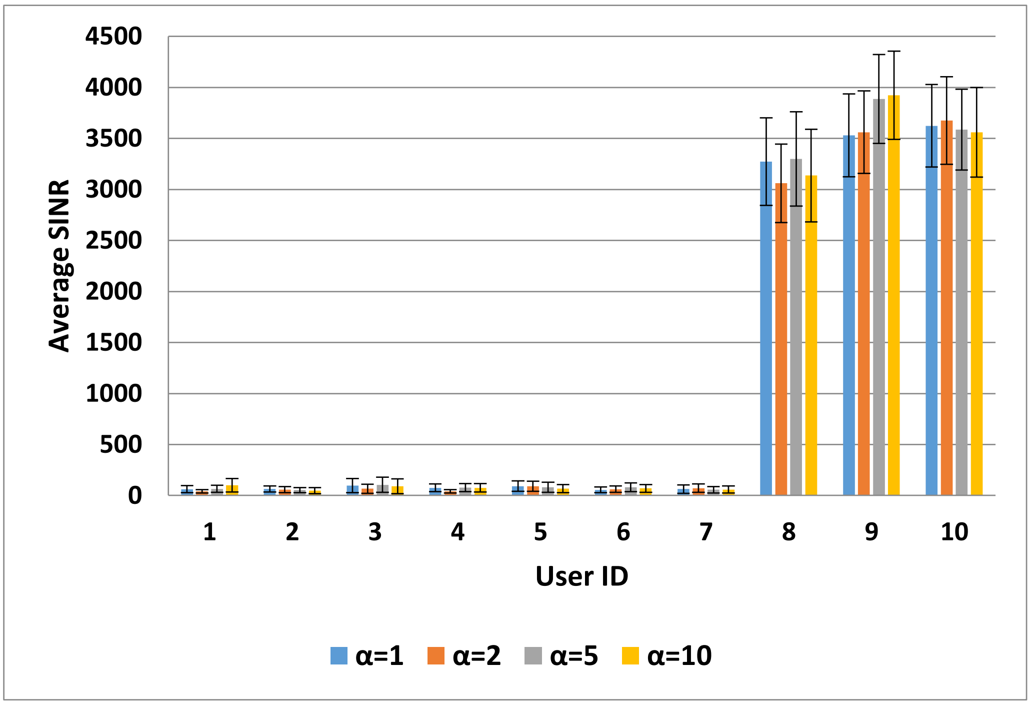
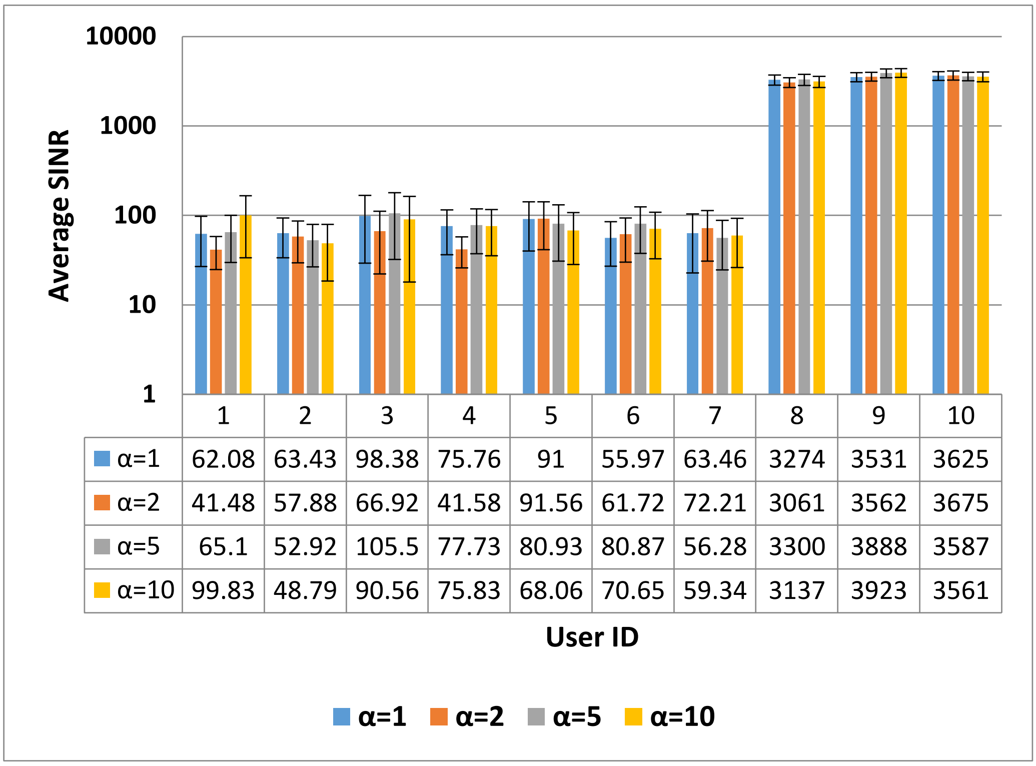


Figure 7-9: User SINR after OP Prioritisation in linear Scale (PF Approach)



	1	2	3	4	5	6	7	8	9	10
🟦 α=1	62.08	63.43	98.38	75.76	91	55.97	63.46	3274	3531	3625
🟧 α=2	41.48	57.88	66.92	41.58	91.56	61.72	72.21	3061	3562	3675
⬜ α=5	65.1	52.92	105.5	77.73	80.93	80.87	56.28	3300	3888	3587
🟨 α=10	99.83	48.79	90.56	75.83	68.06	70.65	59.34	3137	3923	3561

Figure 7-10: User SINR after OP Prioritisation in logarithmic Scale (PF Approach)

7.4.3.3 The Impact of α on Fairness and SINR

Increasing the OPs' priority by adjusting the tuning factor α has similar effects to the ones observed in Subsection 7.4.2.3 . Using the PF approach, boosts the OPs' SINRs by up to 95%, but has resulted in reducing the overall system SINR by up to 48% while maintaining a good fairness interpreted as a stable and very low SD as illustrated in Figure 7-11. Observing Figure 7-12, it can be clearly seen that the OPs' are granted SINRs approximately three times the system's average SINR. Furthermore, the analogy between the priorities (weights) granted to the OPs and the corresponding increase in their SINRs is highlighted.

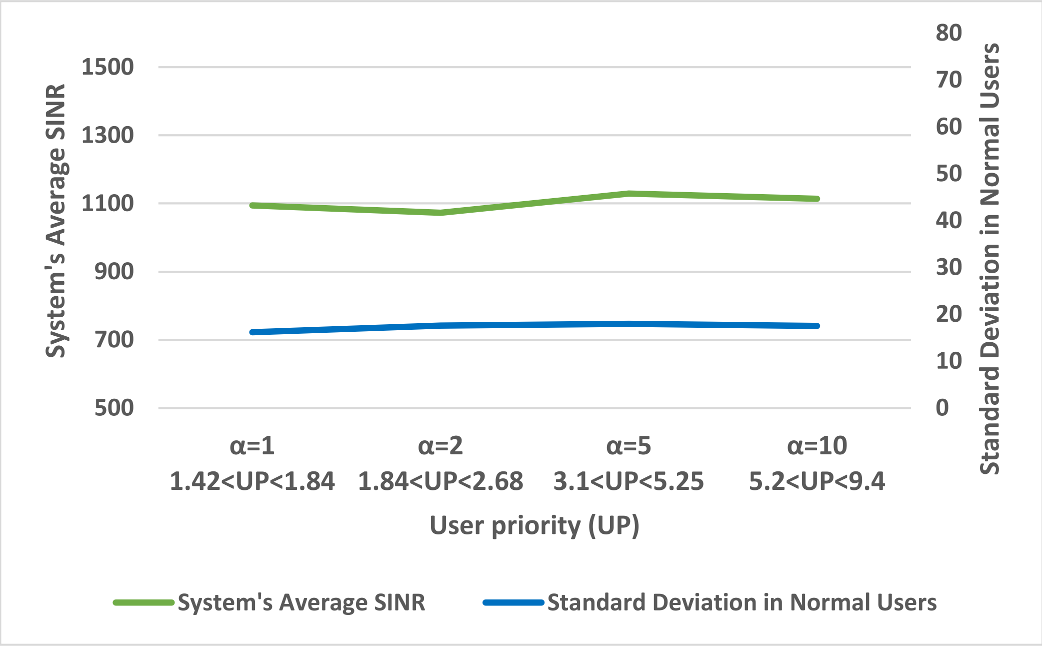


Figure 7-11: Effects of changing α on average SINR and fairness (PF Approach)

It should be noted that user 9, despite having a higher priority than user 10, it was assigned an SINR very close to the SINR of user 10 when $\alpha = 1\,, 2$. This is due to the fact that user 10 has already better channel conditions than user 9 as indicated in Figure 7-8. Thus, it would require higher values of the tuning factor α to bias the system towards user 9 and this can be seen in $\alpha = 5, 10$ in Figure 7-12.

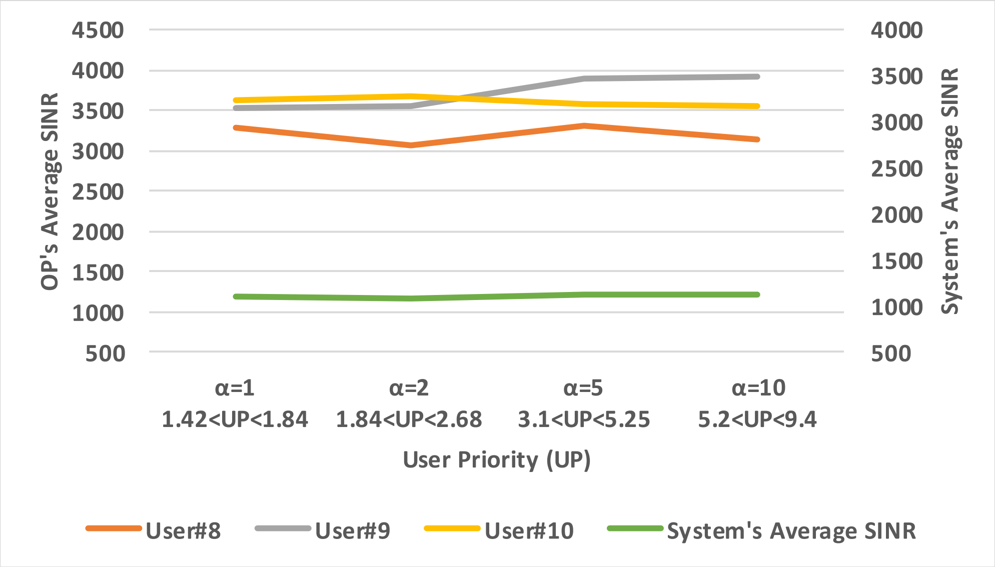


Figure 7-12: The impact of α both user and average SINR (PF Approach)

7.4.4 Reliability-aware PF Approach

In this approach, we are enhancing the SINR values for the normal users that are impacted by the logarithmic sum. This is done by setting a minimum SINR where the users that are subjected to this constraint will have a guaranteed reliable service levels [207].

7.4.4.1 Before Prioritising the OPs

This approach shares the same objective of the PF approach in section 7.4.3.1 . However, a constraint is added to the model guaranteeing a minimum SINR of 21 dB for all users. The results depicted in Figure 7-13 shows a similar trend to the ones illustrated in Figure 7-8. However, preserving a minimum SINR level with no prioritisation means there will be a slight impact on the system-wide SINR. Thusly, we registered a 5% decrease in the system's average SINR for the PF approach before and after introducing reliability.

$$S_k \geq \psi \qquad \forall\, k \in \mathcal{K} \tag{7-45}$$

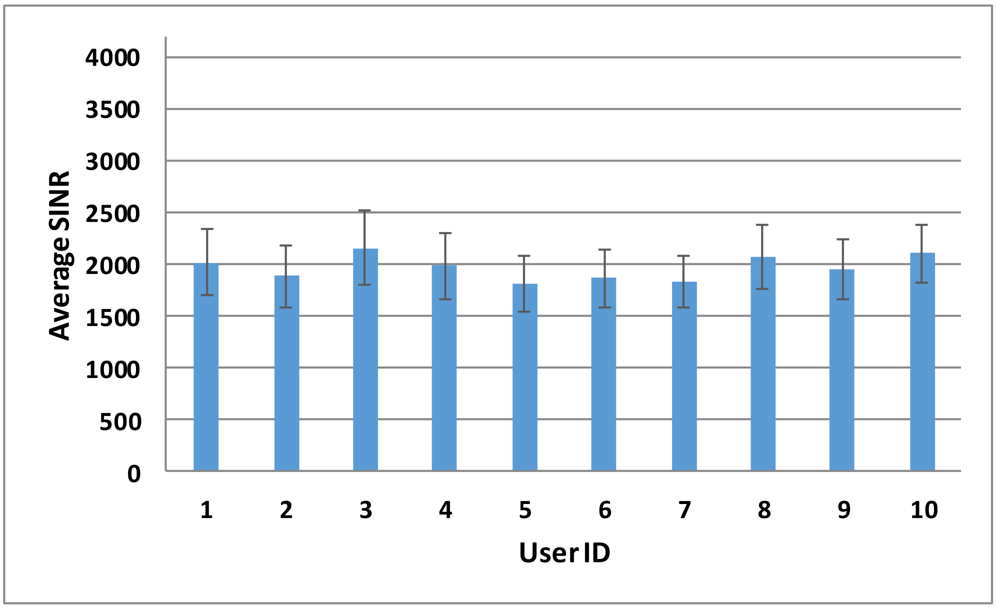


Figure 7-13: User SINR before OP Prioritisation (Reliability-aware PF Approach)

7.4.4.2 After Prioritising the OPs

The impact of the natural logarithm on healthy users motivated the inclusion of a constraint guaranteeing the minimum SINR level as in [207]. This results in an additional level of reliability with fairness in the PF approach.

$$S_k \geq \psi \qquad \forall\, k \in \mathcal{K}: 1 \leq k \leq NU \tag{7-46}$$

Constraint (7-46) works under the objective in (7-31) to guarantee a minimum SINR level specified by the parameter ψ. The result of introducing this constraint is shown in Figure 7-14.

The OPs' SINRs are boosted by up to 23% observed by user 9 with $\alpha = 10$. However, the OPs' SINRs are now reduced in comparison with the previous scenario before introducing reliability as shown in Figure 7-9. The results show narrower confidence intervals than under the WSRMax approach indicating a further reduction in the error values.

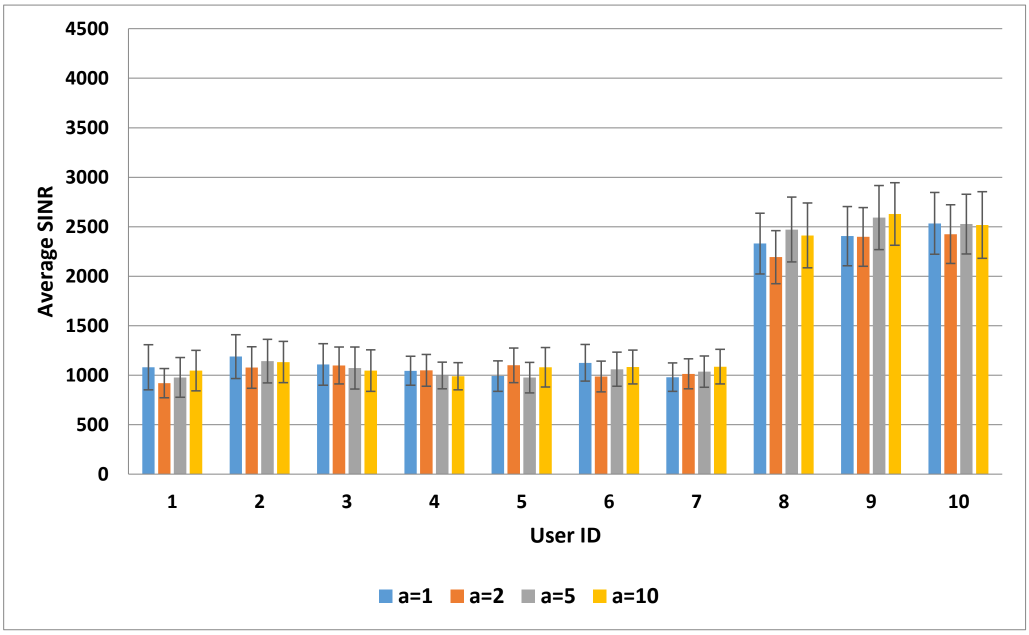


Figure 7-14: User SINR after OP Prioritisation (Reliability-aware PF Approach)

7.4.4.3 The Impact of α on Fairness and SINR

Introducing the reliability aspect to the PF approach resulted in improving the system's average SINR with a marginal increase in the SD. However, better fairness is observed when increasing the tuning factor α as indicated in Figure 7-15. Furthermore, the average SINR is increased by 32% in comparison to the reliability-unaware PF approach.

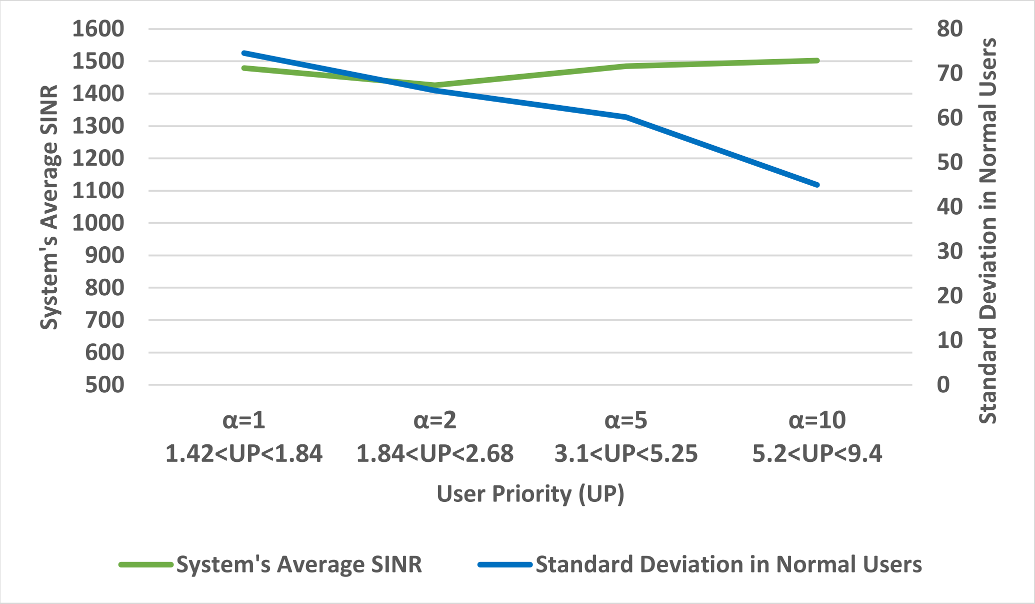


Figure 7-15: Effects of changing α on average SINR and fairness (Reliability-aware PF Approach)

The OPs' SINRs witnessed a 30% increase when employing the reliability-aware PF approach as shown in Figure 7-16. Moreover, the OPs were granted SINRs that are approximately 70% higher than the system's average SINR.

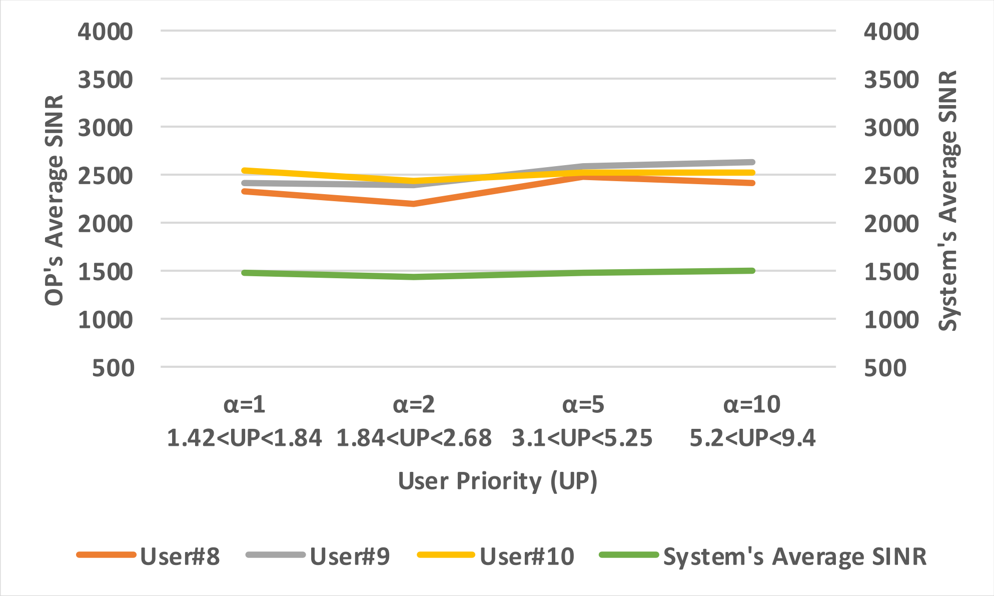


Figure 7-16: The impact of α both user and average SINR (Reliability-aware PF Approach)

7.5 Chapter Summary

This work introduced two interdisciplinary approaches to transform conventional HetNets by endowing them with a user-centric dimension. To that end, a BDA-powered framework was proposed to play part in uplink radio resource allocation optimisation model of a HetNet. The target is to prioritise stroke outpatients within the HetNet to provide them with the optimal wireless resources. Moreover, the assigned resources should be proportional to the severity of the patients' medical state (i.e., stroke likelihood), which is predicted using an ensemble system classifying readings of vital signs acquired from body-attached and nearby IoT sensors. Two approaches, namely, the WSRMax and the PF are presented and compared in terms of fairness and in terms of the average SINR (both at the system and the user level). The WSRMax approach enhanced the OP's average SINR by up to 57%, whereas the PF approach improved them by up to 95%. Depending on the value of tuning factor α, normal users reported an average SINR ranging between 2163 and 1263 using the WSRMax approach, while the reliability-aware PF approach attained an SINR ranging from 1089 to 1066 (depending on α). Using the SD to quantify fairness among users, the WSRMax scored between 104 and 156, while the reliability-aware PF approach ranged between 44 and 74. Furthermore, to add confidence in the estimated probability of stroke, the ensemble system is examined and the voting classifier yielded up to 93% accuracy, a false positive rate of 2.8% and a false negative rate of 11%.

Chapter 8
Conclusions and Future Work

This chapter summarises the work presented in this thesis and specifies the original contributions. In addition, this chapter suggests potential new directions for future research that could be conducted as a result of the work presented in this thesis.

8.1 Conclusions

This section summarises the work that has been performed in the present thesis and states its original contributions. This thesis investigates the use of BDA and ML algorithms in the design, operation, and optimisation of cellular networks. Thus, a new paradigm of user-centric cellular networks powered by BDA is introduced. We focus in this work on stroke patients due to the significance of their medical status and the intrinsic time requirements. In this thesis, we introduce an interdisciplinary approach to optimise the uplink in cellular networks while prioritising cellular-connected-OPs using BDA and MILP optimisation to grant the OPs suitable PRBs according to their current health condition. A dual role for the OP's data is envisioned, along with diagnosis, it guides the network operator to the OPs with the most urgent needs in order that resources can be directed towards them. We argue that ensuring high-quality connectivity between the OP-linked peripherals and their medical provider represents an important step toward highly personalised e-healthcare-centric services and applications.

A number of contributions are introduced in this thesis, starting with Chapter 2, we surveyed the role of ML in the radio resource optimisation of wireless networks. We highlighted the fact that most of the research relied on supervised as well as reinforcement learning methods and that the field of ML is receiving increased interest and is being incorporated in wireless network design for emerging technologies like 5G's new radio (NR) and cognitive radio. In Chapter 3, we surveyed the role BDA can play in wireless network design. Throughout our survey, we noticed a lot of focus on the field of 5G where it is getting most of the researchers' consideration due to the new prospects it has to offer. The contributions of this chapter helped identify the challenges and the opportunities facing the use of BDA in wireless network design. Thus, this chapter can help academic researchers save effort and time. Further, we also surveyed network equipment manufacturing companies offering network solutions using BDA. In Chapter 4, where we developed a seamless integration of the NB classifier that is jointly programmed with the MILP model used to optimise the uplink of the considered cellular network. This classifier uses real patient big data sets to determine the likelihood of a stroke. NB classifier scored an accuracy of 60%, 63.3%, and 63.3% and precision of 65.2%, 66% and 71.6% for users 8, 9 and 10 (i.e., OP 1, 2, and 3), respectively. In Chapter 5, we introduced a novel interdisciplinary approach incorporating the topics of resource allocation, disease risk prediction, patient monitoring, and prioritisation to optimise the uplink of a single-tier homogenous LTE-A network while prioritising cellular-connected-OPs using BDA and MILP optimisation to grant the OPs suitable PRBs according to their current health situation. Moreover, using MILP, two approaches to maximise the OPs' SINRs were developed, namely, the WSRMax approach and the PF approach. We compared the approaches in terms of the fairness achieved between the users and the percentage increase in the SINR. Furthermore, we developed a heuristic to verify the MILP results and we studied the computational complexity of this heuristic. We considered a high number of instances to reflect different network realisations and presented the results indicating a 95% confidence interval. The approaches increased the OPs' average SINR by 26.6% and 40.5%, respectively. The WSRMax approach increased the system's total SINR to a level higher than that of the PF approach, however, the PF approach reported higher SINRs for the OPs, better fairness and a lower margin of error. The work was extended in Chapter 6 to include a two-tier HetNet employing the spectrum partitioning strategy. Thus, mitigating the inter-tier interference. Moreover, we extended this work by considering higher number of instances. Thus, studying the system performance over an extended period of time. and testing the system response over different current states for each OP. The WSRMax and the PF approaches were considered, and the results were compared in terms of fairness and the overall system SINR where it is shown that the WSRMax approach can increase the OP’s SINRs by up to 16%, and the PF can achieve higher than that but with a higher impact on the normal users’ SINRs.

In Chapter 7, we studied the system performance under inter-and-intra-tier interference in a two-tier HetNet. We expanded the dataset to 200 entries and incorporated the concept of ensemble system (i.e., soft voting classifier) where DT, LR, and the NB classifiers were jointly used. Furthermore, we examined the classifiers’ performance by conducting various tests of accuracy, specificity, recall, false-positive rate, false-negative rate, negative prediction rate, precision, and F1 score. Furthermore, reporting the cross-validation test scores for all datasets. Moreover, we added a reliability-aware aspect to the PF approach. Further, we tested the fairness among users, and conducted the required sensitivity analysis over 300 instances. The results show that the WSRMax approach enhanced the OP’s average SINR by up to 57%, whereas the PF approach improved the SINR by up to 95%. Depending on the value of tuning factor α, normal users reported an average SINR ranging between 2163 and 1263 using the WSRMax approach, while the reliability-aware PF approach attained an SINR ranging from 1089 to 1066 (depending on α). Using the SD to quantify fairness among users, the WSRMax scored between 104 and 156, while the reliability-aware PF approach ranged between 44 and 74. Furthermore, to add confidence in the estimated probability of stroke, the ensemble system is examined and the voting classifier yielded up to 93% accuracy, a false positive rate of 2.8% and a false negative rate of 11%.

8.2 Future Research Directions

8.2.1 Choosing the Decision-making Entity

Choosing the optimal type and location of computing (e.g. cloud, fog, etc.) is a separate optimisation problem. Additionally, this may depend on other factors (or variables) like the ratio of OPs to normal users.

8.2.2 Testing the impact of the Feature Ranking Techniques

The current system treats the feature variables on an equal basis. However, we plan to further study the impact of each feature and correspondingly employ a suitable feature ranking technique. The impact of this technique can then be verified with clinical help.

8.2.3 Routing within Small Cells in 5G Networks with Privacy

The proposed solution can be integrated with 5G networks. Optimised routing algorithms can be developed to carry the OPs' traffic through the small cells with minimum latency. In addition, it is vital to protect the OPs' privacy through the traversed hops. This can be addressed by classifying the OPs' data in a ranking system, where the highest rank is treated as the most private medical data. Hence, a specific (secure) route is selected.

8.2.4 Impact of OP Mobility

Grouping the OPs into clusters with common mobility patterns allows the operator to know in advance if there are some areas with high OP density. Hence, prepare the network. This means deploying more nodes so that these OPs do not severely impact the network operation. In addition, our current system works on a given realisation of the patient data and channel conditions (although consideration is given to many realisations). However, in a real-world scenario, there is a constant change in the number of users accessing and leaving the BS coverage. Such dynamic behaviour should be addressed, possibly by OP weighted beamforming and beamsteering.

8.2.5 Use of Infrastructure Sharing and Game Theory

The use of infrastructure sharing can help ensure the widest coverage since the resulting area is the combination of all the local (or national) operators' coverage at a reduced cost. To encourage the operators to participate, game theory can be used to establish coalitions, such that, for example, the higher the number of OPs, the more revenue is awarded to the operator, e.g., reduced taxes.

8.2.6 Wireless energy transfer for Remote Drug Injection

Ensuring high-energy transfer in the downlink might be integrated with our approach to power the body sensors or to actuate a drug-injection mechanism. This can be used in the case of a sudden degradation in the health parameters especially in the case of critical conditions such as diabetes. The reliability of such an approach should be evaluated and improved. Moreover, the delay component from the time of data collection until administering the injection is crucial and has to be considered in the model.

8.2.7 Testing other Discretisation values

The current model uses three ranges to categorise the continuous feature values of the Framingham dataset according to medical entities like the American National Institute of Health and the British Stroke Association. However, other medical entities such as the European Society of Hypertension (ESH) and the European Society of Cardiology (ESC) [208] offer further discretisation ranges. In addition to comparing classification results, the use of different discretisation techniques can be expected to affect the classification bias and variance of generated NB classifiers [209].

8.2.8 Using other types of NB classifiers and increasing the number of features

Since feature selection can have a direct impact on the performance of a prediction model, we recommend a future expansion for the current work to include more feature variables. Upon which, a further system examination can be carried out investigating the classifiers' performance. Furthermore, future work may consider and compare other ML methods especially NB-variant classifiers such as the semi-naïve Bayesian classifier [210] or the locally weighted NB classifier [211].

8.2.9 Investigating the system response to other fading models

Studying other fading models can further enrich the proposed system. The use of different fading models in the current optimisation framework will simulates different working conditions and environments. Rician fading that has less severity than Rayleigh fading due to a dominant multi-path component (normally a light-of-Sight component) can be considered. Alternatively, Nakagami fading which is more general can be considered [212].

8.2.10 Examining the Impact of Network densification Multi-tier HetNet

The current model can be further extended to investigate the effect of network densification. Network densification techniques not only improve capacity and coverage but also enable carriers to maximise spectral efficiency. However, 5G cells may not be able to maintain the classic “always on” routine. Rather, an operational strategy for most 5G cells might be “turn on when required”. Therefore, the current system can be extended to examine the effects of “turn-on when needed” techniques.

List of References

[1] M. S. Hadi, A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Big Data Analytics for Wireless and Wired Network Design: A Survey," *Computer Networks,* 2018.

[2] P. Kiran, M. G. Jibukumar, and C. V. Premkumar, "Resource allocation optimization in LTE-A/5G networks using big data analytics," in *2016 International Conference on Information Networking (ICOIN)*, ed: IEEE, 2016, pp. 254-259.

[3] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Network,* vol. 30, pp. 44-51, 2016.

[4] M. M. Rathore, A. Ahmad, A. Paul, J. Wan, and D. Zhang, "Real-time Medical Emergency Response System: Exploiting IoT and Big Data for Public Health," *Journal of medical systems,* vol. 40, p. 283, 2016.

[5] R. Cortés, X. Bonnaire, O. Marin, and P. Sens, "Stream processing of healthcare sensor data: studying user traces to identify challenges from a big data perspective," *Procedia Computer Science,* vol. 52, pp. 1004-1009, 2015.

[6] M. Ballon, "Number Crunchers • Trojan Family Magazine", 2013. [Online]. Available: https://tfm.usc.edu/number-crunchers/. [Accessed: January 2020]

[7] N. S. Banu and S. Swamy, "Prediction of heart disease at early stage using data mining and big data analytics: A survey," in *Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 016 International Conference on*, 2016, pp. 256-261.

[8] D. Mozaffarian, E. Benjamin, A. Go, D. K. Arnett, M. J. Blaha, M. Cushman*, et al.*, "AHA statistical Update," *Heart Dis. stroke,* vol. 132, 2015.

[9] S. Association, "State of the nation: Stroke statistics," *Retrieved June,* 2018.

[10] N. Bui and J. Widmer, "Mobile network resource optimization under imperfect prediction," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2015 IEEE 16th International Symposium on a*, 2015, pp. 1-9.

[11] M. Al-Rawi, R. Jantti, J. Torsner, and M. Sagfors, "Channel-aware inter-cell interference coordination for the uplink of 3G LTE networks," in *2009 Wireless Telecommunications Symposium*, 2009, pp. 1-5.

[12] S. Sesia, M. Baker, and I. Toufik, *LTE-the UMTS long term evolution: from theory to practice*: John Wiley & Sons, 2011.

[13] A. Aijaz, M. R. Nakhai, and A. H. Aghvami, "Power efficient uplink resource allocation in LTE networks under delay QoS constraints," in *2014 IEEE Global Communications Conference*, 2014, pp. 1239-1244.

[14] F. Ghavimi, Y.-W. Lu, and H.-H. J. I. T. o. V. T. Chen, "Uplink scheduling and power allocation for M2M communications in SC-FDMA-based LTE-A networks with QoS guarantees," vol. 66, pp. 6160-6170, 2017.

[15] F. Moety, S. Lahoud, B. Cousin, and K. Khawam, "Joint power-delay minimization in 4g wireless networks," in *Wireless Days (WD), 2014 IFIP*, 2014, pp. 1-8.

[16] B. Bakhshi and S. Khorsandi, "On the performance and fairness of dynamic channel allocation in wireless mesh networks," *International Journal of Communication Systems,* vol. 26, pp. 293-314, 2013.

[17] R. V. Sathya, V. Venkatesh, R. Ramji, A. Ramamurthy, and B. R. Tamma, "Handover and SINR optimized deployment of LTE FEMTO base stations in enterprise environments," *Wireless Personal Communications,* vol. 88, pp. 619-643, 2016.

[18] P. Adasme, J. Leung, and A. Lisser, "Resource allocation in uplink wireless multi-cell OFDMA networks," *Computer Standards & Interfaces,* vol. 44, pp. 274-289, 2016.

[19] J. P. Muñoz-Gea, R. Aparicio-Pardo, H. Wehbe, G. Simon, and L. Nuaymi, "Optimization framework for uplink video transmission in hetnets," in *Proceedings of Workshop on Mobile Video Delivery*, 2014, p. 6.

[20] J. F. Borin and N. L. S. da Fonseca, "Admission control for WiMAX networks," *Wireless Communications and Mobile Computing,* vol. 14, pp. 1409-1419, 2014.

[21] T. Ohkubo, K. Asayama, M. Kikuya, H. Metoki, H. Hoshi, J. Hashimoto*, et al.*, "How many times should blood pressure be measured at home for better prediction of stroke risk? Ten-year follow-up results from the Ohasama study," *Journal of hypertension,* vol. 22, pp. 1099-1104, 2004.

[22] T. M. Mitchell, *Machine Learning* vol. 1: McGraw-Hill Boston, MA:, 1997.

[23] S. Raschka and V. Mirjalili, *Python machine learning*: Packt Publishing Ltd, 2017.

[24] E. Alpaydin, *Introduction to machine learning*: MIT press, 2009.

[25] EDUCBA, "Machine Learning Algorithms | Categories, Division and Top 8 Algorithms", 2019. [Online]. Available: https://www.educba.com/machine-learning-algorithms/. [Accessed: January 2020]

[26] J. P. Mueller and L. Massaron, *Machine learning for dummies*: John Wiley & Sons, 2016.

[27] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*: " O'Reilly Media, Inc.", 2016.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. J. A. S. e. n. Witten, "The WEKA data mining software: an update," vol. 11, pp. 10-18, 2009.

[29] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, *et al.*, "Orange: data mining toolbox in Python," vol. 14, pp. 2349-2353, 2013.

[30] S. Land and S. J. R.-I. G. Fischer, "Rapid Miner 5," 2012.

[31] C. Idoine, P. Krensky, E. Brethenoux, and A. Linden, "Magic Quadrant for Data Science and Machine Learning Platforms", 2019. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb/. [Accessed: January 2020]

[32] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Communications Surveys and Tutorials,* vol. 21, pp. 2224 - 2287, 2019.

[33] P. V. Klaine, M. A. Imran, O. Onireti, R. D. J. I. C. S. Souza, and Tutorials, "A survey of machine learning techniques applied to self-organizing cellular networks," vol. 19, pp. 2392-2431, 2017.

[34] M. Bkassiny, Y. Li, S. K. J. I. C. S. Jayaweera, and Tutorials, "A survey on machine-learning techniques in cognitive radios," vol. 15, pp. 1136-1159, 2012.

[35] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. J. I. W. C. Hanzo, "Machine learning paradigms for next-generation wireless networks," vol. 24, pp. 98-105, 2016.

[36] J. Moysen, L. Giupponi, and J. Mangues-Bafalluy, "A machine learning enabled network planning tool," in *2016 IEEE 27th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*, 2016, pp. 1-7.

[37] J. Moysen, L. Giupponi, and J. Mangues-Bafalluy, "On the potential of ensemble regression techniques for future mobile network planning," in *2016 IEEE Symposium on Computers and Communication (ISCC)*, 2016, pp. 477-483.

[38] J.-B. Wang, J. Wang, Y. Wu, J.-Y. Wang, H. Zhu, M. Lin*, et al.*, "A machine learning framework for resource allocation assisted by cloud computing," *IEEE Network,* vol. 32, pp. 144-151, 2018.

[39] J. Zhang, X. Xu, K. Zhang, B. Zhang, X. Tao, and P. Zhang, "Machine Learning Based Flexible Transmission Time Interval Scheduling for eMBB and uRLLC Coexistence Scenario," *IEEE Access,* vol. 7, pp. 65811-65820, 2019.

[40] J. Fu, G. Wu, Y. Zhang, L. Deng, and S. Fang, "Active User Identification Based on Asynchronous Sparse Bayesian Learning With SVM," *IEEE Access,* vol. 7, pp. 108116-108124, 2019.

[41] E. C. Santos, "A Supervised Machine Learning Mechanism for Traffic and Flow Control in LTE-A Scheduling," in *Proceedings of the 10th Latin America Networking Conference*, 2018, pp. 33-39.

[42] C.-J. Huang, W. K. Lai, R.-L. Luo, and Y.-L. Yan, "Application of support vector machines to bandwidth reservation in sectored cellular communications," *Engineering Applications of Artificial Intelligence,* vol. 18, pp. 585-594, 2005.

[43] Z. Zhou, H. Yu, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez, "Dependable content distribution in D2D-based cooperative vehicular networks: A big data-integrated coalition game approach," *IEEE Transactions on Intelligent Transportation Systems,* vol. 19, pp. 953-964, 2018.

[44] Y. Zang, F. Ni, Z. Feng, S. Cui, and Z. Ding, "Wavelet transform processing for cellular traffic prediction in machine learning networks," in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2015, pp. 458-462.

[45] P. Savazzi and L. Favalli, "Dynamic cell sectorization using clustering algorithms," in *2007 IEEE 65th Vehicular Technology Conference-VTC2007-Spring*, 2007, pp. 604-608.

[46] K. M. Thilina, K. W. Choi, N. Saquib, and E. Hossain, "Machine learning techniques for cooperative spectrum sensing in cognitive radio networks," *IEEE Journal on selected areas in communications,* vol. 31, pp. 2209-2221, 2013.

[47] Y.-Y. Liu and S.-J. Yoo, "Dynamic resource allocation using reinforcement learning for LTE-U and WiFi in the unlicensed spectrum," in *2017 Ninth*

International Conference on Ubiquitous and Future Networks (ICUFN), 2017, pp. 471-475.

[48] P. Baltiiski, I. Iliev, B. Kehaiov, V. Poulkov, and T. Cooklev, "Long-term spectrum monitoring with big data analysis and machine learning for cloud-based radio access networks," *Wireless Personal Communications,* vol. 87, pp. 815-835, 2016.

[49] P. T. Semov, V. Poulkov, A. Mihovska, and R. Prasad, "Increasing throughput and fairness for users in heterogeneous semi coordinated deployments," in *2014 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2014, pp. 40-45.

[50] H. Yang, A. Alphones, W.-D. Zhong, C. Chen, and X. Xie, "Learning-Based Energy-Efficient Resource Management by Heterogeneous RF/VLC for Ultra-Reliable Low-Latency Industrial IoT Networks," *IEEE Transactions on Industrial Informatics,* 2019.

[51] J. Jang, H. J. Yang, and S. Kim, "Learning-Based Distributed Resource Allocation in Asynchronous Multicell Networks," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 2018, pp. 910-913.

[52] Z. Chen and R. C. Qiu, "Q-learning based bidding algorithm for spectrum auction in cognitive radio," in *2011 Proceedings of IEEE Southeastcon*, 2011, pp. 409-412.

[53] H.-S. Lee, J.-Y. Kim, and J.-W. Lee, "Resource Allocation in Wireless Networks With Deep Reinforcement Learning: A Circumstance-Independent Approach," *IEEE Systems Journal,* 2019.

[54] C. Fan, B. Li, C. Zhao, W. Guo, and Y.-C. Liang, "Learning-based spectrum sharing and spatial reuse in mm-wave ultradense networks," *IEEE Transactions on Vehicular Technology,* vol. 67, pp. 4954-4968, 2017.

[55] A. T. Z. Kasgari, W. Saad, and M. Debbah, "Human-in-the-loop wireless communications: Machine learning and brain-aware resource management," *IEEE Transactions on Communications,* 2019.

[56] J. Qadir, N. Ahad, E. Mushtaq, and M. Bilal, "SDNs, Clouds, and Big Data: New Opportunities," in *2014 12th International Conference on Frontiers of Information Technology*, ed: IEEE, 2014, pp. 28-33.

[57] R. Tudoran, A. Costan, and G. Antoniu, "OverFlow: Multi-Site Aware Big Data Management for Scientific Workflows on Clouds," *IEEE Transactions on Cloud Computing,* vol. 4, pp. 76-89, 2015.

[58] S. Gole, "A survey of Big Data in social media using data mining techniques," presented at the 2015 International Conference on Advanced Computing and Communication Systems, 2015.

[59] Z. Nyikes and Z. Rajnai, "Big Data , As Part of the Critical Infrastructure," pp. 217-222, 2015.

[60] L. Null and J. Lobur, *The essentials of computer organization and architecture*: Jones & Bartlett Publishers, 2014.

[61] J. Shemer and P. Neches, "The genesis of a database computer," *Computer,* vol. 17, pp. 42-56, 1984.

[62] V. R. Borkar, M. J. Carey, and C. Li, "Big data platforms," *XRDS: Crossroads, The ACM Magazine for Students,* vol. 19, p. 44, 2012.

[63] S. Ghemawat, H. Gobioff, and S.-t. Leung, "The Google File System," 2003.

[64] D. J. DeWitt, B. Gerber, G. Graefe, M. Heytens, K. Kumar, and G. A. Muralikrishna, *A High Performance Dataflow Database Machine*: Computer Science Department, University of Wisconsin, 1986.

[65] S. Fushimi, M. Kitsuregawa, and H. Tanaka, "An Overview of The System Software of A Parallel Relational Database Machine GRACE," in *VLDB*, 1986, pp. 209-219.

[66] S. Yin and O. Kaynak, "Big Data for Modern Industry :," *Proceedings of the IEEE,* vol. 103, pp. 143-146, 2015.

[67] A. S. Alghamdi, I. Ahmad, and T. Hussain, "Big Data for C4I Systems : Goals , Applications , Challenges and Tools," presented at the International Conference on Innovative Computing Technology (INTECH), 2015.

[68] A. McAfee and E. Brynjolfsson, "Big Data. The management revolution," *Harvard Buiness Review,* vol. 90, pp. 61-68, 2012.

[69] V. Moreno-Cano, F. Terroso-Saenz, and A. F. Skarmeta-Gomez, "Big data for IoT services in smart cities," presented at the 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), 2015.

[70] K. Sravanthi and T. Subba Redy, "Applications of BIG Data in Various Fields," *International Journal of Computer Science and Technologies,* vol. 6, pp. 4629-4632, 2015.

[71] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," *IEEE Transactions on Intelligent Transportation Systems,* vol. 16, pp. 865 - 873, 2014.

[72] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of Big Data,* vol. 2, p. 24, 2015.

[73] H. Baek and S.-K. Park, "Sustainable Development Plan for Korea through Expansion of Green IT: Policy Issues for the Effective Utilization of Big Data," *Sustainability,* vol. 7, pp. 1308-1328, 2015.

[74] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," *2013 46th Hawaii International Conference on System Sciences,* pp. 995-1004, 2013.

[75] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowledge and Information Systems,* vol. 46, pp. 241-284, 2016.

[76] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in Scientific Data Infrastructure," presented at the Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013, 2013.

[77] J. Andreu-Perez, C. C. Poon, R. D. Merrifield, S. T. Wong, and G. Z. Yang, "Big data for health.," *IEEE journal of biomedical and health informatics,* vol. 19, pp. 1193-208, Jul 2015.

[78] L. Zhang, "A framework to model big data driven complex cyber physical control systems," in *2014 20th International Conference on Automation and Computing*, ed: IEEE, 2014, pp. 283-288.

[79] P. D. C. d. Almeida and J. Bernardino, "Big Data Open Source Platforms," in *2015 IEEE International Congress on Big Data*, ed: IEEE, 2015, pp. 268-275.

[80] C. Senbalci, S. Altuntas, Z. Bozkus, and T. Arsan, "Big data platform development with a domain specific language for telecom industries," *2013*

High Capacity Optical Networks and Emerging/Enabling Technologies, HONET-CNS 2013, pp. 116-120, 2013.

[81] B. Fan, S. Leng, and K. Yang, "A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks," *IEEE Network,* vol. 30, pp. 6-10, 2016.

[82] C.-L. I, Y. Liu, S. Han, S. Wang, and G. Liu, "On Big data Analytics for Greener and Softer RAN," *IEEE Access,* vol. 3, pp. 3068-3075, 2015.

[83] P. Russom, "Big data analytics," *TDWI Best Practices Report,* p. 38, 2011.

[84] A. Belle, R. Thiagarajan, S. M. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big Data Analytics in Healthcare," *Biomed Res Int,* vol. 2015, p. 370194, 2015.

[85] R. Buyya, K. Ramamohanarao, C. Leckie, R. N. Calheiros, A. V. Dastjerdi, and S. Versteeg, "Big Data Analytics-Enhanced Cloud Computing: Challenges, Architectural Elements, and Future Directions," pp. 75-84, 2015.

[86] P. Gölzer, L. Simon, P. Cato, and M. Amberg, "Designing Global Manufacturing Networks Using Big Data," *Procedia CIRP,* vol. 33, pp. 191-196, 2015.

[87] R. Kapdoskar, S. Gaonkar, N. Shelar, A. Surve, and P. S. Gavhane, "Big Data Analytics," vol. 4, pp. 518-520, 2015.

[88] C. Hu, H. Li, Y. Jiang, Y. Cheng, and P. Heegaard, "Deep semantics inspection over big network data at wire speed," *IEEE Network,* vol. 30, pp. 18-23, 2016.

[89] E. Bastug, M. Bennis, E. Zeydan, M. A. Kader, I. A. Karatepe, A. S. Er*, et al.*, "Big data meets telcos: A proactive caching perspective," *Journal of Communications and Networks,* vol. 17, pp. 549-557, 2015.

[90] B. Matturdi, X. Zhou, S. Li, and F. Lin, "Big Data security and privacy: A review," *China Communications,* vol. 11, pp. 135-145, 2014.

[91] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications,* vol. 19, pp. 171-209, 2014.

[92] A. Asahara, H. Hayashi, N. Ishimaru, R. Shibasaki, and H. Kanasugi, "International standard “OGC® moving features” to address “4Vs” on locational bigdata," in *2015 IEEE International Conference on Big Data (Big Data)*, ed: IEEE, 2015, pp. 1958-1966.

[93] L. He and P. Yue, "Moving towards intelligent giservices," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, ed: IEEE, 2015, pp. 1373-1376.

[94] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access,* vol. 2, pp. 652-687, 2014.

[95] B. Cyganek, M. Grana, A. Kasprzak, K. Walkowiak, and M. Wozniak, "Selected aspects of electronic health record analysis from the big data perspective," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, ed: IEEE, 2015, pp. 1391-1396.

[96] L. Cui, F. R. Yu, and Q. Yan, "When big data meets software-defined networking: SDN for big data and big data for SDN," *IEEE Network,* vol. 30, pp. 58-65, Jan-Feb 2016.

[97] Y. Demchenko, E. Gruengard, and S. Klous, "Instructional Model for Building Effective Big Data Curricula for Online and Campus Education," in

2014 IEEE 6th International Conference on Cloud Computing Technology and Science, ed: IEEE, 2014, pp. 935-941.

[98] M. A.-u.-d. Khan, M. F. Uddin, and N. Gupta, "Seven V's of Big Data understanding Big Data to extract value," in *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*, ed: IEEE, 2014, pp. 1-5.

[99] M. K. Pusala, M. A. Salehi, J. R. Katukuri, Y. Xie, and V. Raghavan, "Massive Data Analysis: Tasks, Tools, Applications, and Challenges," in *Big Data Analytics*, ed: Springer, 2016, pp. 11-40.

[100] S. Pyne, B. P. Rao, and S. B. Rao, *Big Data Analytics: Methods and Applications*: Springer, 2016.

[101] K. Lee, K. Jung, J. Park, and D. Kwon, "ARLS: A MapReduce-based output analysis tool for large-scale simulations," *Advances in Engineering Software,* vol. 95, pp. 28-37, May 2016.

[102] T. White, *Hadoop: The definitive guide*: " O'Reilly Media, Inc.", 2012.

[103] M. Lemoudden and B. E. Ouahidi, "Managing cloud-generated logs using big data technologies," in *2015 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, ed: IEEE, 2015, pp. 1-7.

[104] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," *Journal of Big Data,* vol. 2, p. 8, 2014.

[105] A. B. Ayed, M. B. Halima, and A. M. Alimi, "MapReduce Based Text Detection in Big Data Natural Scene Videos," *Procedia Computer Science,* vol. 53, pp. 216-223, 2015.

[106] N. Zhu, X. Liu, J. Liu, and Y. Hua, "Towards a cost-efficient MapReduce: mitigating power peaks for Hadoop clusters," *Tsinghua Science and Technology,* vol. 19, pp. 24-32, 2014.

[107] "Big Data in the Enterprise : Network Design Considerations," *White Paper,* pp. 1-33, 2011.

[108] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, ed: IEEE, 2010, pp. 1-10.

[109] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark : Cluster Computing with Working Sets," *HotCloud'10 Proceedings of the 2nd USENIX conference on Hot topics in cloud computing,* p. 10, 2010.

[110] N. Marz, "Apache Storm". [Online]. Available: https://storm.apache.org/. [Accessed: January 2020]

[111] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: distributed data-parallel programs from sequential building blocks," in *ACM SIGOPS Operating Systems Review*, 2007, pp. 59-72.

[112] V. K. Vavilapalli, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia*, et al.*, "Apache Hadoop YARN," in *Proceedings of the 4th annual Symposium on Cloud Computing - SOCC '13*, ed. New York, New York, USA: ACM Press, 2013, pp. 1-16.

[113] P. Zikopoulos, C. Eaton, and D. DeRoos, "Understanding big data," *New York et al: McGraw ...,* p. 166, 2012.

[114] F. Morales, M. Ruiz, L. Gifre, L. M. Contreras, V. López, and L. Velasco, "Virtual network topology adaptability based on data analytics for traffic prediction," *Journal of Optical Communications and Networking,* vol. 9, pp. A35-A45, 2017.

[115] A. Imran and A. Zoha, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network,* vol. 28, pp. 27-33, 2014.

[116] H. Daki, A. El Hannani, A. Aqqal, A. Haidine, A. Dahbi, and H. Ouahmane, "Towards adopting Big Data technologies by mobile networks operators: A Moroccan case study," in *Cloud Computing Technologies and Applications (CloudTech), 2016 2nd International Conference on*, 2016, pp. 154-161.

[117] D. S. Terzi, R. Terzi, and S. Sagiroglu, "Big data analytics for network anomaly detection from netflow data," in *International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 592-597.

[118] Ö. F. Çelebi, E. Zeydan, Ö. F. Kurt, Ö. Dedeoglu, Ö. Iieri, B. A. Sungur*, et al.*, "On use of big data for enhancing network coverage analysis," in *2013 20th International Conference on Telecommunications, ICT 2013*, ed: IEEE, 2013, pp. 1-5.

[119] I. A. Karatepe and E. Zeydan, "Anomaly Detection In Cellular Network Data Using Big Data Analytics," in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, ed, 2014, pp. 1-5.

[120] E. J. Khatib, R. Barco, P. Munoz, I. D. La Bandera, and I. Serrano, "Self-healing in mobile networks with big data," *IEEE Communications Magazine,* vol. 54, pp. 114-120, 2016.

[121] E. J. Khatib, R. Barco, I. Serrano, and P. Munoz, "LTE performance data reduction for knowledge acquisition," in *Globecom Workshops (GC Wkshps), 2014*, 2014, pp. 270-274.

[122] I. de la Bandera, R. Barco, P. Munoz, and I. Serrano, "Cell Outage Detection Based on Handover Statistics," *Communications Letters, IEEE,* vol. 19, pp. 1189-1192, 2015.

[123] A. Sahni, D. Marwah, and R. Chadha, "Real time monitoring and analysis of available bandwidth in cellular network-using big data analytics," *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on,* pp. 1743-1747, 2015.

[124] J. Liu, F. Liu, and N. Ansari, "Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop," *IEEE Network,* vol. 28, pp. 32-39, 2014.

[125] W. Huang, Z. Chen, W. Dong, H. Li, B. Cao, and J. Cao, "Mobile Internet big data platform in {China} Unicom," *Tsinghua Science and Technology,* vol. 19, pp. 95-101, 2014.

[126] "Wi-Fi direct | Wi-Fi Alliance". [Online]. Available: ttp://www.wi-fi.org/discover-wi-fi/wi-fi-direct. [Accessed: January 2020]

[127] A. Omar, "Improving Data Extraction Efficiency of Cache Nodes in Cognitive Radio Networks Using Big Data Analysis," in *2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies*, ed: IEEE, 2015, pp. 305-310.

[128] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger*, et al.*, "Cloud RAN for mobile networks—a technology overview," *Communications Surveys & Tutorials, IEEE,* vol. 17, pp. 405-426, 2015.

[129] K. I. Pedersen, Y. Wang, S. Strzyz, and F. Frederiksen, "Enhanced inter-cell interference coordination in co-channel multi-layer LTE-advanced networks," *Wireless Communications, IEEE,* vol. 20, pp. 120-127, 2013.

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

[131] M. Cayrol, H. Farreny, and H. Prade, "Fuzzy pattern matching," *Kybernetes,* vol. 11, pp. 103-116, 1982.

[132] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning,* vol. 23, pp. 69-101, 1996.

[133] M. Molina, I. Paredes-Oliva, W. Routly, and P. Barlet-Ros, "Operational experiences with anomaly detection in backbone networks," *Computers & Security,* vol. 31, pp. 273-285, 2012.

[134] F. Ricciato, "Traffic monitoring and analysis for the optimization of a 3G network," *IEEE Wireless Communications,* vol. 13, pp. 42-49, 2006.

[135] M. S. Parwez, D. Rawat, and M. Garuba, "Big Data Analytics for User Activity Analysis and User Anomaly Detection in Mobile Wireless Network," *IEEE Transactions on Industrial Informatics,* vol. 13, pp. 2058 - 2065, 2017.

[136] J. Spiess, Y. T'Joens, R. Dragnea, P. Spencer, and L. Philippart, "Using big data to improve customer experience and business performance," *Bell Labs Technical Journal,* vol. 18, pp. 3-17, 2014.

[137] J. Zhong, W. Guo, and Z. Wang, "Study on network failure prediction based on alarm logs," in *Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference on*, 2016, pp. 1-7.

[138] L. H. Shuan, T. Y. Fei, S. W. King, G. Xiaoning, and L. Z. Mein, "Network Equipment Failure Prediction with Big Data Analytics," *International Journal of Advances in Soft Computing & Its Applications,* vol. 8, pp. 59-69, 2016.

[139] K. Yang, R. Liu, Y. Sun, J. Yang, and X. Chen, "Deep Network Analyzer (DNA): A Big Data Analytics Platform for Cellular Networks," *IEEE Internet of Things Journal,* vol. 4, pp. 2019-2027, 2017.

[140] Y. Qiao, Z. Lei, J. Yang, and G. Cheng, "FLAS: Traffic analysis of emerging applications on Mobile Internet using cloud computing tools," in *Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on*, 2013, pp. 1-6.

[141] G. Qi, W.-T. Tsai, W. Li, Z. Zhu, and Y. Luo, "A cloud-based triage log analysis and recovery framework," *Simulation Modelling Practice and Theory,* vol. 77, pp. 292-316, 2017.

[142] B. H. Park, S. Hukerikar, R. Adamson, and C. Engelmann, "Big Data Meets HPC Log Analytics: Scalable Approach to Understanding Systems at Extreme Scale," in *Cluster Computing (CLUSTER), 2017 IEEE International Conference on*, 2017, pp. 758-765.

[143] C. Jardak, P. Mähönen, and J. Riihijärvi, "Spatial big data and wireless networks: experiences, applications, and research challenges," *IEEE Network,* vol. 28, pp. 26-31, 2014.

[144] L. Da Xu, W. He, and S. J. I. T. o. i. i. Li, "Internet of things in industries: A survey," vol. 10, pp. 2233-2243, 2014.

[145] E. Bertino, "Big Data - Security and Privacy," *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy,* pp. 757-761, 2015.

[146] K. Crawford, "Six provocations for big data," pp. 1-17, 2011.

[147] a. Labrinidis and H. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment,* pp. 2032-2033, 2012.

[148] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns.," *Nature,* vol. 453, pp. 779-82, 2008.

[149] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *Journal of Big Data,* vol. 2, pp. 1-32, 2015.

[150] A. Rajasekar, R. Moore, C.-Y. Hou, C. a. Lee, R. Marciano, A. de Torcy*, et al.*, "iRODS Primer: Integrated Rule-Oriented Data System," *Synthesis Lectures on Information Concepts, Retrieval, and Services,* vol. 2, pp. 1-143, 2010.

[151] M. Jensen, "Challenges of Privacy Protection in Big Data Analytics," in *2013 IEEE International Congress on Big Data*, ed: IEEE, 2013, pp. 235-238.

[152] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in big data analytics," *Journal of Parallel and Distributed Computing,* vol. 74, pp. 2561-2573, 2014.

[153] E. A. Brewer, "Towards robust distributed systems," in *PODC*, 2000.

[154] R. Iyer, "Datacenter-on-Chip Architectures Terascale Opportunities and Challenges," *Intel Technology Journal,* vol. 11, 2007.

[155] J. Gemson Andrew Ebenezer, S. Durga, A. A. a. Belle, R. R. c. Thiagarajan, S. M. R. a. M. R. Soroushmehr, F. d. F. Navidi*, et al.*, "Big Data Analytics in Healthcare," *ARPN Journal of Engineering and Applied Sciences,* vol. 2015, pp. 1-16, 2015.

[156] Z. Zheng, "Naive Bayesian classifier committees," in *European Conference on Machine Learning*, 1998, pp. 196-207.

[157] M. R. Mia, S. A. Hossain, A. C. Chhoton, and N. R. Chakraborty, "A Comprehensive Study of Data Mining Techniques in Health-care, Medical, and Bioinformatics," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, 2018, pp. 1-4.

[158] E. Miranda, E. Irwansyah, A. Y. Amelga, M. M. Maribondang, and M. Salim, "Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier," *Healthcare Informatics Research,* vol. 22, pp. 196-205, 2016.

[159] M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," in *Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on*, 2015, pp. 520-525.

[160] L. A. Muhammed, "Using data mining technique to diagnosis heart disease," in *Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on*, 2012, pp. 1-3.

[161] V. Chaurasia and S. Pal, "Data mining approach to detect heart diseases," 2014.

[162] K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *Computer Science and Education (ICCSE), 2010 5th International Conference on*, 2010, pp. 1344-1349.

[163] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Systems with Applications,* vol. 40, pp. 96-104, 2013.

[164] N. Cheung, "Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering," B. Sc. Thesis, University of Queenland, 2001.

[165] B. Šter and A. Dobnikar, "Neural networks in medical diagnosis: Comparison with other methods," in *International conference on engineering applications of neural networks*, 1996, pp. 427-30.

[166] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications,* vol. 17, pp. 43-48, 2011.

[167] T. J. Peter and K. Somasundaram, "An empirical study on prediction of heart disease using classification data mining techniques," in *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on*, 2012, pp. 514-518.

[168] D. Sitar-Taut, D. Pop, D. Zdrenghea, and A. Sitar-Taut, "Using machine learning algorithms in cardiovascular disease risk evaluation," *Journal of Applied Computer Science & Mathematics,* vol. 3, pp. 29-32, 2009.

[169] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, 2008, pp. 108-115.

[170] "Heart Disease Data Set ", ed. Center for Machine Learning and Intelligent Systems, Bren School of Information and Computer Science, University of California, Irvine: https://archive.ics.uci.edu/ml/datasets/Heart+Disease, [Accessed: January 2020]

[171] W. W. LaMorte, "Using Spreadsheets in Public Health," *handout, School of Public Health, Boston University,* [Accessed: January 2020]

[172] Framingham Heart Study, "History of Framingham Heart Study". [Online]. Available: http://www.framinghamheartstudy.org/about-fhs/history.php. [Accessed: 26 April]

[173] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials,* vol. 18, pp. 1153-1176, 2016.

[174] L. Soibelman and H. Kim, "Data preparation process for construction knowledge generation through knowledge discovery in databases," *Journal of Computing in Civil Engineering,* vol. 16, pp. 39-48, 2002.

[175] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*: Springer, 2015.

[176] E. M. Karabulut, S. A. Özel, and T. Ibrikci, "A comparative study on the effect of feature selection on classification accuracy," *Procedia Technology,* vol. 1, pp. 323-327, 2012.

[177] A. Lewis and A. Segal, "Hyperlipidemia and primary prevention of stroke: does risk factor identification and reduction really work?," *Current atherosclerosis reports,* vol. 12, pp. 225-229, 2010.

[178] M. L. Dyken, "Stroke risk factors," in *Prevention of stroke*, ed: Springer, 1991, pp. 83-101.

[179] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing,* vol. 237, pp. 350-361, 2017.

[180] J. L. Flores, I. Inza, and P. Larrañaga, "Wrapper discretization by means of estimation of distribution algorithms," *Intelligent Data Analysis,* vol. 11, pp. 525-545, 2007.

[181] C.-H. Lee, "A Hellinger-based discretization method for numeric attributes in classification learning," *Knowledge-Based Systems,* vol. 20, pp. 419-425, 2007.

[182] S. Ramírez-Gallego, S. García, H. Mouriño-Talín, D. Martínez-Rego, V. Bolón-Canedo, A. Alonso-Betanzos*, et al.*, "Data discretization: taxonomy and big data challenge," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 6, pp. 5-21, 2016.

[183] N. I. o. Health, "Your guide to lowering your blood pressure with DASH," ed: Smashbooks, 2006.

[184] S. Association, "Blood pressure information pack", 2017. [Online]. Available: https://www.stroke.org.uk/sites/default/files/take_a_moment_blood_pressure_info_pack.pdf. [Accessed: January 2020]

[185] U. D. o. Health, H. Services, N. I. o. Health, L. National Heart, and B. Institute, "Your guide to lowering your cholesterol with TLC," *NIH Publication,* 2005.

[186] S. H. Jee, I. Suh, I. S. Kim, and L. J. Appel, "Smoking and atherosclerotic cardiovascular disease in men with low levels of serum cholesterol: the Korea Medical Insurance Corporation Study," *Jama,* vol. 282, pp. 2149-2155, 1999.

[187] C. C. Petersen, "A note on transforming the product of variables to linear form in linear programs," *Diskussionspapier, Purdue University,* 1971.

[188] A. B. Bishop, T. Hughes, and M. McKee, "Water Rourses Systems Analysis-Course Notes," 1999.

[189] J. P. Hart and A. W. Shogan, "Semi-greedy heuristics: An empirical study," *Operations Research Letters,* vol. 6, pp. 107-114, 1987.

[190] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, "Introduction to algorithms second edition," ed: The MIT Press, 2001.

[191] V. Chandrasekhar and J. G. J. I. T. o. C. Andrews, "Spectrum allocation in tiered cellular networks," vol. 57, 2009.

[192] M. Hadi, A. Lawey, T. El-Gorashi, and J. Elmirghani, "Using Machine Learning and Big Data Analytics to Prioritize Outpatients in HetNets," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 726-731.

[193] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," in *Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on*, 2012, pp. 173-177.

[194] J. Butler and A. Kalogeropoulos, "Hospital strategies to reduce heart failure readmissions: where is the evidence?," ed: Journal of the American College of Cardiology, 2012.

[195] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, R. J. N. C. Manikandan*, et al.*, "Classification of stroke disease using machine learning algorithms," pp. 1-12, 2019.

[196] N. N. Alotaibi and S. Sasi, "Stroke in-patients' transfer to the ICU using ensemble based model," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 2004-2010.

[197] S. N. Min, S. J. Park, D. J. Kim, M. Subramaniyam, and K.-S. J. E. n. Lee, "Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study in Korea," vol. 79, pp. 214-220, 2018.

[198] L. Regression, "A Self-Learning Text," *Statistics for Biolology and Health, Third Edition, David Kleinbaum, Mitchel Klein,* 1994.

[199] L. Breiman, *Classification and regression trees*: Routledge, 1984.

[200] V. Kotu and B. Deshpande, *Data Science: Concepts and Practice*: Morgan Kaufmann, 2018.

[201] L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash*, et al.*, "Prediction and control of stroke by data mining," vol. 4, p. S245, 2013.

[202] T. Kansadub, S. Thammaboosadee, S. Kiattisin, and C. Jalayondeja, "Stroke risk prediction model based on demographic data," in *2015 8th Biomedical Engineering International Conference (BMEiCON)*, 2015, pp. 1-3.

[203] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*: Elsevier, 2011.

[204] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*: Chapman and Hall/CRC, 2012.

[205] C. Zhang and Y. Ma, *Ensemble machine learning: methods and applications*: Springer, 2012.

[206] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. J. a. p. c. Stamatopoulos, "Stacking classifiers for anti-spam filtering of e-mail," 2001.

[207] Y. Sun, R. P. Jover, and X. Wang, "Uplink interference mitigation for OFDMA femtocell networks," *IEEE Transactions on Wireless Communications,* vol. 11, pp. 614-625, 2011.

[208] A. Zanchetti, A. Dominiczak, A. Coca, C. Tsioufis, D. L. Clement, E. Agabiti Rosei*, et al.*, "2018 ESC/ESH Guidelines for the management of arterial hypertension," *European Heart Journal,* vol. 39, pp. 3021-3104, 2018.

[209] Y. Yang and G. I. Webb, "Discretization for naive-Bayes learning: managing discretization bias and variance," *Machine learning,* vol. 74, pp. 39-74, 2009.

[210] I. Kononenko, "Semi-naive Bayesian classifier," in *European Working Session on Learning*, 1991, pp. 206-219.

[211] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive bayes," in *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, 2002, pp. 249-256.

[212] A. Goldsmith, *Wireless communications*: Cambridge university press, 2005.