Order number: _____
Serial No: __________
People's Democratic Republic of Algeria

Ministry of High Education and

Scientific Research

Echahid Hamma Lakhdar University of El-oued

Faculty of Technology

Presented for graduation from

ACADEMIC MASTERS

In: Telecommunication
Specialty: Telecommunication Systems
By: BEKKARI Abdechafi & HALEM Abdelhay

Theme

Performance Analysis of 5G Wireless Networks Using Machine Learning

In front of the jury composed of:

Dr. Ghendir Said	Supervisor
Dr.Tedjani Amina	Member
Dr.Bellili Imane	President

Acknowledgments

This work was done and conducted in the Technology Department of Hamma Lakhdar University - Eloued.

First and foremost, we express our sincere gratitude to the all mightily Allah for granting us the strength, patient and determination to finalize this work to its completion.

We would like to extend our deepest gratitude to our professor and supervisor ***Dr.Ghandir Said*** *for his invaluable guidance, continuous support and insightful advices on our road of the finalization of this work, his expertise and encouragement contributed massively on finishing this thesis.*

Dedication

*To my beloved mother, Latra, **who** supported me throughout my journey in my **studies**, **who** never forced a choice **on** me, and always supported me in all the ways she could.*

*To my late father, Ali, **who has** always been at my side spiritually; your hard work in life **has inspired** me to pursue my dreams and achieve what I want.*

To my family, who stood beside me and supported me in unimaginable ways during my hard times.

*And lastly, to my dear friend, Aristeidis Stathis, who stood by my side and **helped refine** this work to perfection*

Abdechafi Bekkari

Dedication

*To my beloved family, For their endless support, **patience**, and sacrifice throughout my whole life.*

*To my parents, **Whose prayers** and guidance led me to where I am now. No words can describe my gratitude **toward both of you**.*

*To my brothers, sisters, my kids, and all my loved ones, Thank you for always being there for me. **Through thick and thin**, you always stood by my side*

Abdelhay Halem

Abstract

Abstract

The 5th generation wireless system is a quantum leap in the evolution of telecommunication systems, due to its very high data rate, extremely low latency and massive connection to fit the next generation startup applications (6G and more). However, the introduction of 5G systems brings with it a number of challenges to conquer such as signal propagation, interference, multipath, fading and resource management issues (power for example) for a dynamic environment, particularly in high dense city areas. Machine learning (ML) techniques can be employed at the edge of wireless networks to enhance the accuracy of performance prediction and the level of network intelligence to target these concerns.

In this thesis, we investigate the performance of 5G wireless system based on massive MIMO and beamforming at 3.5 GHz (n78) for downlink band under realistic propagation conditions. The proposed model includes both LOS and NLOS, path loss, inter-cell interference and Rayleigh fading. A simulation environment was implemented in the form of python codes in Google CoLab in order to produce the dataset and analyze the Key Performance Indicators (KPI's) such as SINR, throughput and latency.

In addition, a machine learning (ML) based predictive framework has been added in the form of Random forest Regressor algorithm to predict performance metrics at the physical-layer. The result shows the good accuracy and prove the validity and the high value of ML methods for modeling the non-linear characteristics of 5G wireless environments. The analysis indicates that Machine learning and 5G systems integration can offer a solid and promising solutions to intelligent network management and future self-organizing wireless networks.

Keywords: 5G, Machine Learning, Random forest, massive MIMO, beamforming, SINR, Throughput, Latency, Rayleigh Fading.

Resumé

Le système sans fil de cinquième génération (5G) est un saut quantique dans l'évolution des systèmes de télécommunication, grâce à son très haut débit de données, sa latence extrêmement faible et sa connexion massive afin de répondre aux applications émergentes de la prochaine génération (6G et plus). Cependant, l'introduction des systèmes 5G apporte avec elle un certain nombre de défis à surmonter tels que la propagation du signal, l'interférence, le multipath, le fading et les problèmes de gestion des ressources (comme la puissance) dans un environnement dynamique, particulièrement dans les zones urbaines à haute densité. Les techniques de Machine Learning (ML) peuvent être employées à la périphérie des réseaux sans fil afin d'améliorer la précision de la prédiction des performances et le niveau d'intelligence du réseau pour traiter ces préoccupations.

Dans cette thèse, nous étudions les performances d'un système sans fil 5G basé sur le Massive MIMO et le beamforming à 3.5 GHz (n78) pour la bande downlink dans des conditions réalistes de propagation. Le modèle proposé inclut à la fois LOS et NLOS, le path loss, l'inter-cell interference et le Rayleigh fading. Un environnement de simulation a été implémenté sous forme de codes Python dans Google Colab afin de produire le dataset et d'analyser les Key Performance Indicators (KPI's) tels que le SINR, le throughput et la latency.

De plus, un framework prédictif basé sur le Machine Learning (ML) a été ajouté sous la forme de l'algorithme Random Forest Regressor afin de prédire les métriques de performance au niveau de la physical-layer. Les résultats montrent une bonne précision et prouvent la validité ainsi que la grande valeur des méthodes ML pour la modélisation des caractéristiques non linéaires des environnements sans fil 5G. L'analyse indique que l'intégration du Machine Learning et des systèmes 5G peut offrir des solutions solides et prometteuses pour la gestion intelligente des réseaux et les futurs réseaux sans fil auto-organisés.

Mots-clés : 5G, Machine Learning, Random Forest, Massive MIMO, beamforming, SINR, Throughput, Latency, Rayleigh Fading.

ملخص

يُمثل الجيل الخامس من الاتصالات اللاسلكية 5 G مرحلة متقدمة في تطور أنظمة الاتصالات الحديثة، لما يقدمه من سرعة عالية في نقل البيانات، وزمن استجابة منخفض، إضافة إلى قدرته على ربط عدد كبير من الأجهزة في الوقت نفسه. وقد ساهم ذلك في دعم تطبيقات جديدة مثل المدن الذكية، وإنترنت الأشياء، والأنظمة الصناعية الذكية، مع التمهيد لتقنيات الأجيال القادمة.

رغم هذه المزايا، لا تزال شبكات G5 تواجه عدة صعوبات تقنية عند التطبيق العملي. فانتشار الإشارة في الترددات العالية يتأثر بالعوائق والبيئات الحضرية المزدحمة، كما أن التداخل بين الخلايا والتلاشي الناتج عن تعدد المسارات يمكن أن يؤثر بشكل مباشر على جودة الاتصال. ولهذا السبب، أصبح من الضروري البحث عن طرق أكثر كفاءة لتحسين أداء الشبكة وإدارة مواردها بشكل ذكي.

تعتمد هذه الأطروحة على دراسة نظام 5 Gيعمل بتقنيات Massive MIMO و Beamforming عند تردد 3.5 GHz ضمن وصلة Downlink وتحت ظروف انتشار واقعية. كما يأخذ النموذج بعين الاعتبار حالتي LOS وNLOS ، إضافة إلى تأثير Path Loss والتداخل بين الخلايا وتلاشي.Rayleigh

تم إنشاء بيئة محاكاة باستخدام Python عبر منصة Google Colabبهدف توليد مجموعة بيانات Dataset وتحليل مجموعة من مؤشرات الأداء الرئيسية مثل SINR و Throughput و .Latency بعد ذلك، تم توظيف خوارزمية Random Forest Regressorمن أجل التنبؤ بأداء الشبكة على مستوى Physical Layer بالاعتماد على تقنيات .Machine Learning

أظهرت النتائج أن النموذج المقترح قادر على تحقيق تنبؤات دقيقة نسبيًا، خاصة في البيئات التي تتسم بسلوك غير خطي ومعقد. كما بينت الدراسة أن دمج تقنيات Machine Learning داخل أنظمة 5 Gيمكن أن يساهم في تطوير شبكات أكثر ذكاءً وقدرة على التكيف مع ظروف التشغيل المختلفة.

الكلمات المفتاحية:
5G، Machine Learning، Random Forest، Massive MIMO، Beamforming، SINR، Throughput، Latency، Rayleigh Fading.

Contents

List of Figures

List of Tables

List of Abbreviations

1G: First generation
2G: Second generation
3G: Third generation
4G: Fourth generation
5G: Fifth generation
3GPP: 3rd Generation Partnership Project
256QAM: Quadrature Amplitude Modulation
AMF: Access and Mobility Management Function
ANN: Artificial Neural Network
CDF: Cumulative Distribution Function
CSI: Channel State Information
GSM: Global System for Mobile Communications
HARQ: Hybrid Automatic Repeat reQuest
ICI: Inter-Cell Interference
KPI: Key Performance Indicators
LOS: Line Of Sight
LTE: Long Term Evolution
MAE: Mean Absolute Error
MCS: Modulation and Coding Scheme
MIMO: Multiple Input Multiple Output
ML: Machine Learning
mmWave: millimeter Wave
mMTC: massive Machine-Type Communications
MSE: Mean Square Error
NFV: network functions virtualization
NLOS: Non Line Of Sight
QoS: Quality of Service
QPSK: Quadrature Phase Shift Keying
R^2: R-square
RF: Random Forest
RL: Reinforcement Learning
RMSE: Root Mean Square Error
RNN: Recurrent Neural Network
RSRP: Reference Signal Received Power
SCI: Channel State Information
SDN: Software defined network
SINR: Signal-to-Interference-plus-Noise Ratio
SMS: Short Message Service
SNA: Service-Based Architecture
SVM: Support Vector Machine
SVR: Support Vector Regressor
UE: User Equipment
UPF: User Plane Function

URLLC: Ultra-Reliable Low-Latency Communications
eMBB: enhanced Mobile Broadband

General Introduction

1. Background and Motivation

The emergence of Fifth-Generation (5G) wireless networks heralds a new era of innovation and growth in telecommunications. 5G is not an incremental upgrade in speed as its predecessors, but a multi-service platform that can cater for a wide variety of services. These services include Enhanced Mobile Broadband (eMBB) for HD rolling, Ultra-Reliable Low Latency Communications (URLLC) for autonomous vehicles, and mMTC for Internet of Things (IoT).

To provide these services the 5G employs a novel concept-based technologies including the Network Slicing, Massive MIMO and the usage of the Millimeter Wave (mmWave) bands.

But such breakthroughs also bring a new level of complexity in managing networks, which complexity is beyond to what analytical solutions can handle in real-time. [1]

2. The Problem: Multipath Fading in 5G:

As networks evolved with 5G, they started operating at higher frequencies, and as the frequency increased the signal became more prone to obstruction by physical barriers like walls and trees.

Significance: Here the importance of the Riley Channel (Rayleigh Fading Channel) in imitating the reality is revealed; The signal won the user through the multiple reflected paths (Multipath) and as a result the user's performance scrolls up and down (Throughput) and so the delay time is growing (Latency). . . [2] [3]

3. Research Objectives:

The main goal of this thesis “Performance Analysis of 5G Wireless Networks using Machine Learning” is to build fidelity model to analyze and improve KPIs. with a Python-based simulation environment, this thesis deals with:

-Emulating a 5G network to produce realistic datasets.

-Using ML techniques (Random Forest in particular) to model the dependence of performance - on network parameters.

-Testing the Capability of the Model to Predict Throughput and Latency for Different Scales.

-Offering a glimpse into how AI-based analytics may drive the journey to fully autonomous 5G networks.

4.Methodology Overview:

Using a quantitative, empirical approach, this study seeks to narrow the gap between theoretical network models and practical implementation based on performance. It is initiated by stringent review on literature for 5g architecture and ml basics.

The methodology is based on a multi-stage pipeline: At its core, the methodology consists of a pipeline with multiple stages.

Data Generation: Simulating 5G scenarios to collect a heterogeneous dataset.

Preprocessing - Processing raw data to clean and normalize it for use in machine learning.

Model Design: The Random Forest algorithm is implemented due to its superior accuracy, and its capability to capture complex non-linear relationship Evaluation: We use statistical metrics like Mean Squared Error (MSE) and R-squared (R^2) to assess whether the model has predictive power.

5.Thesis Structure

The work reported in this thesis is divided into four (4) chapters, which are arranged in a way that make theoretical part first introduced then followed by the experimental outcomes and final summary in the end.

General Introduction

In this part we present the general introduction for the thesis, starting by the Background and motivation for the transition to 5G then moving forward to stating the problem (multipath fading in 5G) and the network complexity then to the research objective after a brief overview of the work methodology showing how machine learning is used to track these problems

Chapter I: Fundamentals of 5G Wireless Network

It provides the technical background. It presents the Evolution of Mobile Systems (1G to 5G), the 5G Architecture (RAN and Core), and the Important KPIs. It ends with a discussion on the performance challenges for which the intelligence solutions are sought.

Chapter II: Machine Learning for Wireless Communications

This chapter presents the theoretical Machine Learning framework. It is centered on Supervised Learning and in particular Regression Models, which are relevant for the estimation of continuous values (e. g. throughput). It includes a Review of Related Work to discuss how this work is related to other work in academia.

Chapter III: System Model and Data Simulation

This chapter illustrates the experimental findings. 5G Network Model of case study, Simulation Parameters and Dataset Generation. In addition, preliminary Data Analysis and Visualization are included for investigating the characteristics of data obtained by simulation.

Chapter IV: Performance Prediction Based on ML

This is the main technical chapter. Feature Selection (which includes making decisions on which variables are the most important), the Machine Learning Model Design (which relates to models such as Random Forest), and the formal Training and Testing Procedure. It specifies the Evaluation Metrics (e.g., MSE, R-squared) to be employed to the model's accuracy.

Conclusion and future work

Here we summarize the contributions of the thesis. The Limitations of the work and potential Future Research Directions are discussed openly such as integration of Deep Learning or streaming of data in real-time to the proposed model.

Chapter I:

Fundamentals of 5G Wireless Network

I.1 Introduction:

The wireless communication industry has experienced dramatic changes in the past forty years, evolving from relaying rudimentary analog sounds to high-speed smart networks interconnecting billions of devices worldwide. This was not just a faster version of the same thing, but a reconsideration of how human and machine interacted with data.

I.2 Evolution of Mobile Communication Systems:

I.2.1. First Generation (1G): Analog Audio Era: Earliest form of mobile telecommunications, which dates back to the 1980s and is based on analog signals (Analog). Its only function was that of voice calls, with extremely limited bandwidth and quality of connection.

I.2.2 Second Generation (2G): Digitalization: In the 1990s, the digital system (GSM) was introduced globally. This generation also brought users short message service (SMS) technology, as well as the very first stage of data encryption with which users were more protected and anonymous.

I.2.3 Third Generation (3G): Mobile Internet Takes Off: Users were able to browse the Internet and make video calls for the first time with 3G technologies at the turn of the millennium, giving rise to smartphones.

I.2.4 Fourth Generation (4G LTE): Broadband Era This generation concentrated on high-speed data [5]. This has made high-definition video streaming (HD Streaming) and online gaming achievable, with a massive decrease in latency (Latency).

I.2.2 Fifth Generation (5G): The Era of Smart Connectivity 5 While being more than 10 times faster, 5G is not just a speed upgrade, but a revolutionary architecture with three pillars: Actuation, Proximity, and Expertise.

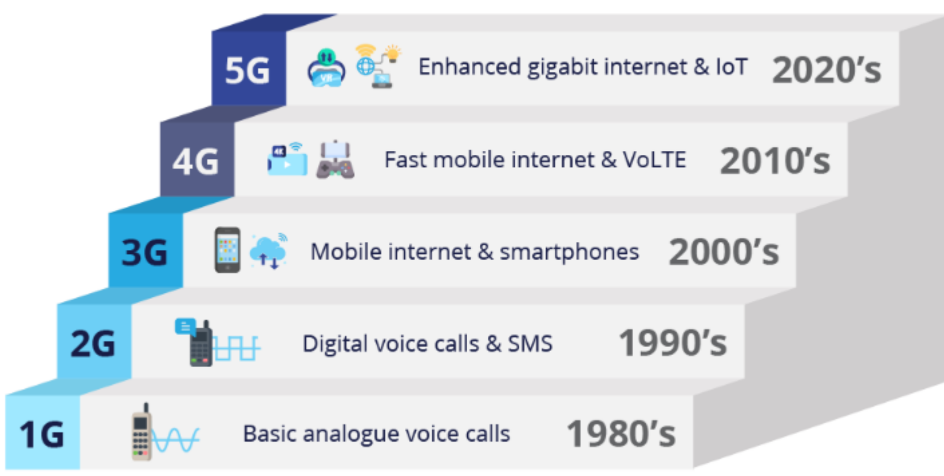


Figure 01: Evolution of Mobile Communication System [10]

I.3 Overview of 5G Architecture:

The 5G design is inherently different from its predecessors which are monolithic in nature. It has three main parts:

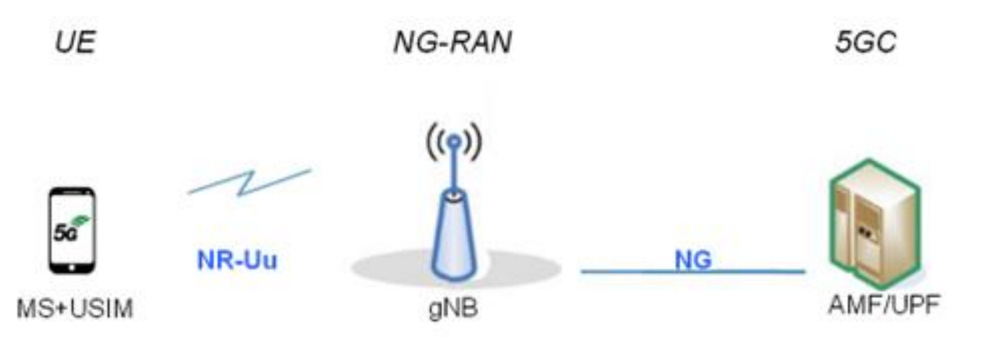


Figure 2: overview of the 5G architecture [11]

I.3.1 User Equipment (UE):

Refers to the end user's device, for example, a smartphone or IoT sensor.

The role of UE is the device by which performance measurements (e.g., throughput and latency) are made in your simulation.

It consists of a mobile terminal (MS) and a global subscriber identity module (USIM) that allows secure access to the network.

The UE communicates with the base station through the NR-Uu air interface (New Radio).

I.3.2 NG-RAN (Next-Generation Radio Access Network):

The base stations in the NG-RAN are called gNodeBs (gNBs). These stations employ multiple innovative technologies to deliver superior performance such as:

Massive MIMO: Using dozens of antenna elements at a base station to serve dozens of users at the same time, growing capacity exponentially.

Beamforming: Instead of sending the signals in multiple directions, gNBs send the signals in a beam to a single user, which results in less interference and power saving.

Millimeter wave (mm Wave): 5G taps into high-frequency bands (from 30 GHz to 300 GHz) to ensure wide bandwidth, but these signals cannot travel very far and undergo atmospheric blockage. [4] [6]

I.3.3 5GC (5G Core Network):

A. Service-Based Architecture (SBA) 5GC:

Services 5GC comprises SBA, whose network functions (NFs) interact by using service-based interfaces (SBIs) to invoke each other's services. This modularity is a departure from the monolithic nature of previous-generation designs and enables a more flexible/scalable network.

B. Technological Enablers "SDN and NFV"

The enforcement of the SBA is based on SDN and NFV. Together, these two technologies allow networking functions to be software-based, run-on general-purpose hardware instead of specialized and expensive legacy hardware. This upgrade transforms the 5G core to be highly programmable and efficient."

C. Basic Functions:

UPF and AMF In the 5G Core Network (5GC), two core procedures account for the vast majority of operations: the Access and Mobility Management Function (AMF) and the User Plane Function (UPF). The AMF manages the user connection and mobility tracking, while the UPF is the main data traffic gateway between the radio access network and the external data networks.

D. Advanced Feature:

Network Slicing "Based on the flexibility of SDN and NFV, Network Slicing is the most important characteristic of 5G. It enables service providers to create multiple virtual networks (slices) over a single physical network. Each Instance will be Additionally tailored according to a particular KPI including ultra-low latency for URLLC service or high throughput for eMBB application."

E. Research Motivation and Complexity:

Although these architecture enhancements have higher performance, but they bring high network management complexity. The dynamic nature of 5G slices plus physical limitations such as Rayleigh Fading and multipath fading in the high frequency band, is a perfect environment that cannot be well handled with analytically solutions in real-time. This complexity is also at the reason why We propose to use Machine Learning, and in particular the Random Forest algorithm, for accurate performance prediction."

I.4 Key Service Scenarios for 5G Wireless Networks:

The 5G wireless network service scenarios are split into 3 main categories each designed to support specific applications and communication requirements with special Key Performance Indicators that target the technical content, unlike older generations that focused on increasing data rate, 5G offer much more flexibility in the architecture that support Enhanced Mobile Broadband, Massive Machine-Type Communications and Ultra-Reliable Low Latency Communications all of wish are explained below:

I.4.1 Enhanced Mobile Broadband (eMBB):

This scenario is a simple evolution of existing mobile broadband services to position for an order of magnitude increase in traffic and enhanced user experience. It provides extremely high data rates to end users, with target peak transfer rates on the order of several Gbps. eMBB is also necessary for such things as the streaming of high-def (HD) video and augmented reality. [5]

I.4.2 Massive Machine-Type Communications (mMTC):

This service type is typified by an extremely large number of connected devices, which can be remote sensors, actuators, or other monitoring apparatus. A major goal is to support as many as 1 million IoT devices per km2. mMTC requirements are stringent in terms of device cost and energy consumption, so that a battery life of several years is achievable. Because these devices generally produce small amounts of data, high data rate is considered less important than connectivity density. [5]

I.4.3 Ultra-Reliable Low Latency Communications (URLLC):

The URLLC services are intended to be used in extreme cases where the system is required to have extremely high reliability and ultra-low latency of milliseconds, the goal from URLLC is to reduce the-end to-end latency to be less than 1 ms while keeping high stability and reliable communication links, this service is critical for applications where delays or packet loss can cause serious damages and safety risks. [5]

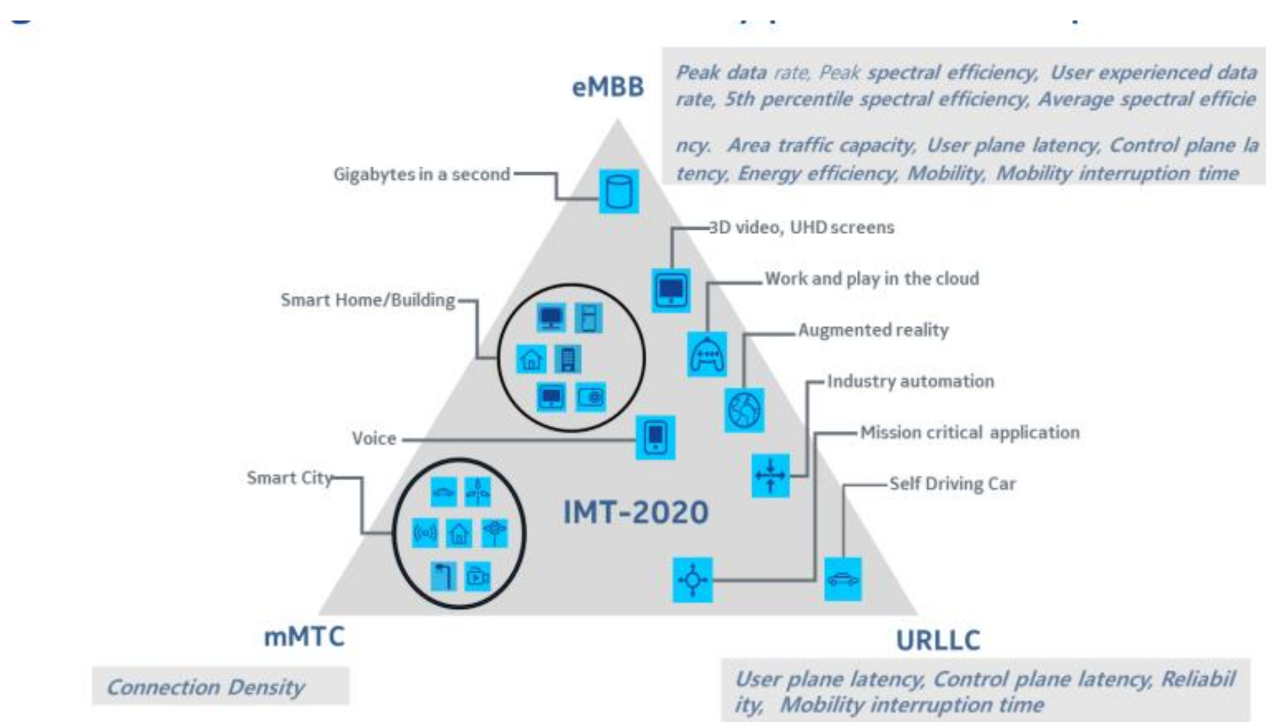


Figure 3: Usage scenarios of IMT for 2020 and the key performance requirements [12]

I.5 Key Performance Indicators (KPI) in 5G:

As for 5G, in order to assess the performance of the 5G deployment, as well as the accuracy of our prediction models, the standard KPIs for the ITU (International Telecommunication Union) in the IMT-2020 vision should be identified: [5]

I.5.1 Peak Data Rate: This indicator represents the maximum "theoretical" throughput that can be obtained under the perfect channel state using new technologies like massive MIMO and high modulation models (Like 256QAM), in our case 5G the peak downlink data rate is up to 20Gbps

I.5.2 User Experienced Data Rate: This indicator represents the maximum throughput for the user in real life cases where the interference, fading and localization (LOS /NLOS) factor in the equation, 5G goal is to give the user data rate around 100 Mbps while under pressure (urban environments)

I.5.3 End-to-end latency: this indicator represents the time (or delay rather) for the packet to travel from the source to the target then back, one of the main objectives of 5G is to reduce this time to be 1 ms especially for URLLC applications that require real time communications.

I.5.4 Spectral Efficiency: This indicator represents the amount of transmitted data per 1 unit of bandwidth, using methods like beamforming and massive MIMO, the 5G seeks to improve this indicator in compare to older generations (usually measured in bit/s/hz).

1.5.5 Energy Efficiency: This indicator represents the target to reduce the network power consumption for better pricing while keeping the high performance and connectivity, this key performance indicator is very important for sense small cell deployments and large scale internet of things environments where saving energy is a must due to the large amount of it used

1.6 Performance Challenges in 5G Networks

Although the 5G architecture offers a revolutionary multi-service platform, its realization introduces a number of critical issues which degrade the Service Quality (QoS). These are largely a consequence of physics-based challenges in high-frequency propagation and the difficulty of supporting heterogeneous services.

1.6.1 Signal Propagation and High Path Loss Since 5G works on higher frequency bands then previous generations, including mmWave, it is more likely to be blocked by physical obstacles since the signal become more vulnerable to environmental conditions and physical obstacles in comparison to previous generations, some of the factors that effect this are the:

A- Atmospheric and Material Blockage: the high frequency signals are easily effected by environmental factors such as walls, trees and what not, even atmospheric conditions like rain or humidity can really effect the signal propagation and path loss in dense urban environments.

B-Limited Coverage Zone: High path loss causes signals in mmWave to travel for short distances, which implies the installation of a number of "small cells" to keep the connection.

1.6.2. Multipath Fading and Interference: The wireless medium is dynamic by nature, making the signal behavior more complicated.

A-Rayleigh Fading: Multiple reflected paths (Multipath) bring the signals to the User Equipment (UE). This leads to rapid fluctuation of signal strength and consequently to unstable Throughput along with higher Latency.

B-Inter-Cell Interference (ICI): The increased small cell density poses a tremendous challenge to management of interference in between adjacent base stations for the classical algorithms.

1.6.3 Dynamic Resource Management and Network Slicing

The multi-service nature of 5G (eMBB, URLLC, mMTC) creates competing demands for limited network resources:

A. Slicing Complexity: The network must perform real-time allocation of bandwidth and power among different virtual slices. For instance, prioritizing a URLLC slice for low latency may negatively impact the throughput of an eMBB slice.

B. Limitations of Traditional Models: Static equations employed in traditional analytical frameworks are often ill-suited for the fast-varying traffic and high mobility patterns inherent in 5G.

1.6.4 Mobility and Handoff Management: A dense 5G environment gives a user numerous moves between small cells. Users can maintain the User Experienced Data Rate by a fast handover without bits error. Robust Beam Tracking: In Massive MIMO, the base station needs to keep track of the user's position all the time accurately to direct the signal beam (Beamforming) which brings huge signaling overhead.

I.7 Conclusion:

In conclusion, while the 5G architecture offers a revolutionary multi-service platform capable of meeting diverse KPIs, its inherent complexity—stemming from signal interference, high mobility, and non-linear resource allocation—presents a significant challenge for network management. These fundamental complexities are extremely difficult to manage using conventional deterministic mathematical models, which often fail to adapt to the highly dynamic and variable wireless environment. Consequently, this complexity serves as the primary motivation for the transition towards data-driven Machine Learning approaches, specifically the Random Forest algorithm. By leveraging historical data to learn and predict performance in real-time, ML provides the necessary intelligence to optimize 5G networks, a theme that will be explored extensively in the following chapter.

Chapter II:

Machine Learning for Wireless Communications

Chapter II: Machine Learning for Wireless Communications

II.1 Introduction to Machine Learning in Wireless Networks:

The escalating adoption of mobile devices along with the increasing demand for mobile applications and services is putting unprecedented strain on the mobile and wireless networking infrastructure. Next generation 5G systems are being designed to accommodate exploding mobile traffic volumes, real-time extraction of fine-grained analytics, and flexible management of the network resources to maximize user experience. The mobile environments are becoming more and more complex, heterogeneous, and dynamic, so that these tasks are difficult to carry out. Perhaps one avenue for solutions is to turn to sophisticated machine learning methods to cope with the surge in data volumes and algorithm-driven applications. The recent success in deep learning motivates new powerful tool for the problems in this space.

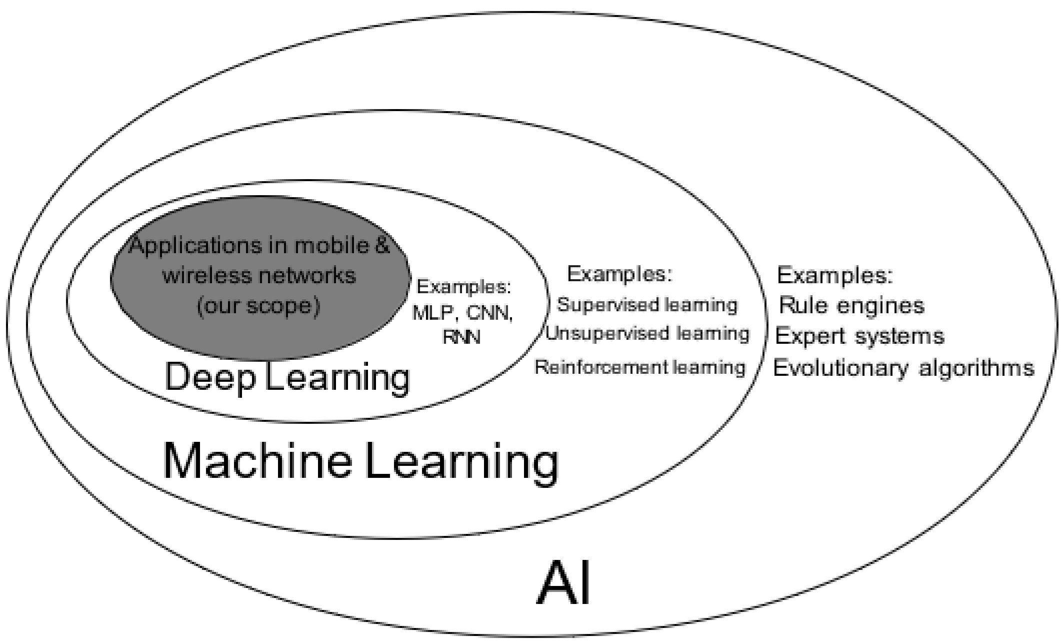


Figure 4: The Relationship between AI, Machine Learning, and Deep Learning in Wireless Networks

II.2 Machine Learning for 5G Introduction

The evolution from Fourth-Generation (4G) to Fifth-Generation (5G) network has brought a new era of the management of networks. In the high dimensionality of massive connectivity and ultra-dynamic radio environment, 5G has evolved to the point that conventional "model-based" designing methods, which are based on deterministic, closed form mathematical models, are met with great challenges. These inherited models are in general not able to accurately characterize the non-linear stochastic nature of the 5G channel, in particular for the scenarios involving Millimeter Wave (mmWave) propagation, fast fading, and heavy interference patterns.

As such, Machine Learning (ML) comes with a data-driven paradigm. Rather than trying to make the environment conform to some strict mathematical form, ML enables the network to discover the statistical structure of the radio environment from raw data. This post-actual prescriptive modeling paradigm is a move towards predictive intelligence. [7]

II.1.2 The Role of ML as a Predictive Engine

In this study, ML is not merely a computational tool but a **predictive engine** designed to map a high-dimensional space of input features to critical output performance metrics. The input vector (X) comprising distance, carrier frequency, interference levels, and path loss) is mapped to a target variable (Y) (such as Throughput or Latency) through a learned function f(X)

$$\mathbf{Y = f(X) + \varepsilon} \quad \text{(II.1)}$$

Where epsilon (**ε**): represents the irreducible noise in the wireless channel.

II.2 Supervised Learning and Feature Engineering

Our generated 5G data set is well labeled so we take advantage of the merits of Supervised Learning algorithms. The efficacy of this method is dependent on Feature Extraction. Unlike Deep Learning (DL) models which tend to be "black box" models that learn patterns automatically by deep layers of neural networks, we utilize human knowledge in our approach. We manipulate and mold the data such that it reflects the temporal and spatial correlations inherent in the network (for example, historical data rates at previous time slots).

This sophisticated data processing allows the applied ML model, i.e. the Random Forest (RF) regressor, to better generalize and therefore achieve better prediction results also in high varying and shadowed scenarios, where traditional analytical solutions would produce erroneous estimates.

II.2.1 Classification/Regression: An Approach to Model Performance Under supervised machine learning setting, the key difference between the algorithms is in the type of output they produce. Although classification and regression both operate on the same principle of learning from past data to build predictive models, they have very different uses in 5G network tune.

A. Classification: Discrete Categories to be Predicted: Classification is the task that compels the discrete label to be predicted or the individual category to be predicted. The patterns identified by the algorithm in the input features are then used to predict to which "class" of data a data point belongs.

Aim: To search for a decision boundary to partition data.

Output: Qualitative labels (e.g., “Congested” vs. “Clear”, or “Indoor” vs. “Outdoor”).

5G Application: Inferring if a user is having a “High” or “Low” quality of service (QoS) without revealing the specific speed.

B. Regression: Continuous Numerical Values: to be Predicted Regression applies if the target variable is continuous or real-valued. Instead of classifying data into categories, it predicts a certain amount on a numeric scale.

Objective: To determine a mapping function that predicts the exact value.

Output : Quantitative values (e.g., 150.5 Mbps, 1.2 ms).

5G Application (Thesis Focus): The precise throughput and latency of a network slice can be predicted at least with distance from the base station and interference levels.

C. Importance of Regression for Our Study

In this dissertation, we consider the performance modeling of 5G networks as a Regression problem. This decision is motivated by a number of technical aspects:

Level of Detail: Classification of 5G Engineering Network as being “Fast” does not provide enough information. For accurate resource allocation and network design, we need to know exactly the data rate in Mbps (Regression).

The 5G channel, nature: Rayleigh Fading and path loss (and other) do not modify signal in steps, they produce a continuous variation. Regression models, and in particular the Random Forest Regressor are well-suited for such fine-grained numerical variations.

Evaluation Criteria for Optimization: By applying regression, we can leverage statistical errors, such as Mean Squared Error (MSE), to measure at what precision our model's estimates converge to the real simulated results, which would ultimately result in less errors while optimizing the network.

II.3 Important Machine Learning Algorithms for 5G Analysis

The application of Machine Learning (ML) in 5G environments requires algorithms capable of handling high-dimensional data and non-linear signal propagation. Several models have been extensively studied for predicting Key Performance Indicators (KPIs):

II.3.1 Support Vector Regression (SVR): Support Vector Regression is a modification of SVM for continuous value prediction. Support Vector Regression is a generalization of SVM in case of regression problems in contrast to classification problem. SVR in 5G exploits SVR to approximate the non-linear mapping between channel state information (CSI) and throughput. avior: It uses the Kernel Trick (usually Radial Basis Function - RBF) to transform the input features such as path loss, interference, etc. Then it looks for a hyperplane that best fits the data points within a certain margin (distance from the hyperplane hovering two parallel hyperplanes) (margin = € insensitive tube). Deep Understanding: Despite the advantages of SVR, like robustness to outliers and good performance in high-dimensional space, the computational complexity is highly increased when applying SVR to large data set, making SVR less suitable for large scale 5G data simulation as compared to ensemble methods.

II.3.2 Artificial Neural Networks (ANN): ANNs, based on the biological brain structure, have layers of connected 'neurons' and can model very complex functions.

Mechanism: ANNs modify weights to reduce the error between actual and predicted data rates, via backpropagation. They are particularly good at finding ‘hidden’ relationships in 5G traffic data patterns.

Deep Insight: Even though ANNs are powerful, they are frequently accused in telecom as being "Black Boxes" – hard to say why a single prediction was made. In addition, they are "data hungry" and need to be finely tuned for many hyper parameters (e.g., number of layers, learning rate).

II.4 Why Random Forest for 5G?

The RF algorithm has shown great promise as a supervised machine learning algorithm for 5G network analytics era as it outperforms in many classification and regression problems. It is a member of the Ensemble Learning schemes which aggregates many models to attain better model prediction accuracy and stability. [8]

II.4 1.Random Forest Algorithm

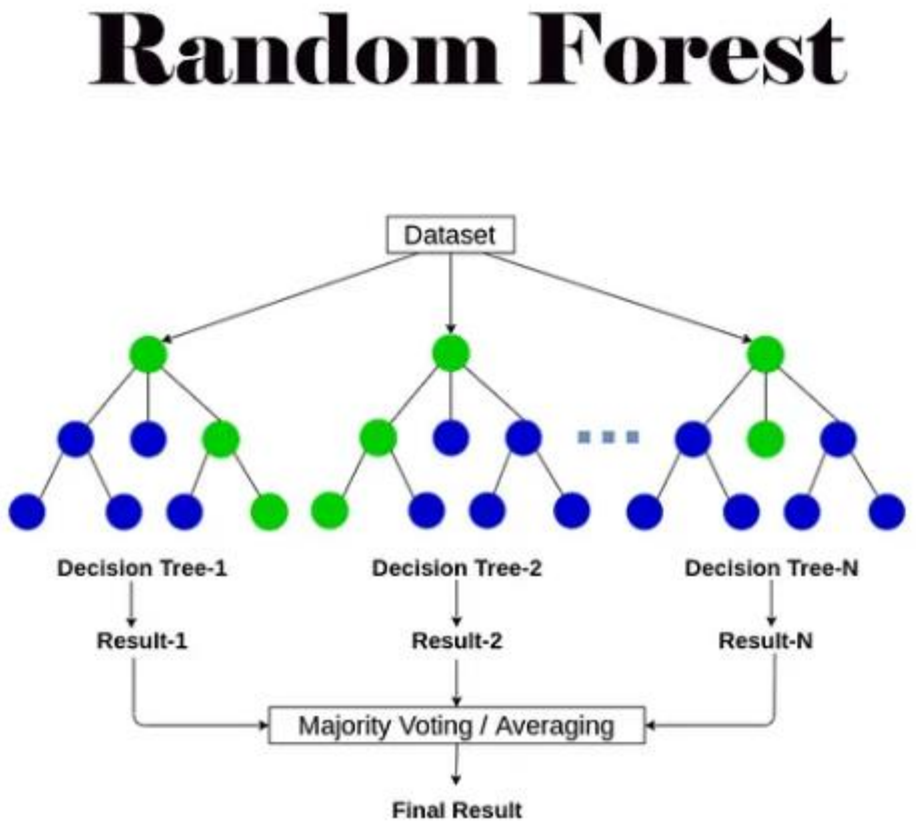


Figure 5: example Random Forest

A- Working Mechanism: In other words, the basic principle of Random Forest is to generate a large number of Decision Trees at the training stage. The prediction of a regression RF model is the aggregated result of a number of regression trees:

a. For Classification: This is a Majority Voting, the class labels which are predicted by the most trees will be the final predicted result of the model (e.g. determining the suitable Network Slice).

b. For Regression: In the case of regression, the output from all trees is averaged (mean).

II.4.2 Major Technical Features: Two key techniques contribute to the success of Random Forest in managing the highly nonlinear wireless data:

A. Bootstrap Aggreging (Bagging): A tree is built on a random sample of the training data (with replacement). This leads to decorrelation between trees, and the effect of this is to drastically decrease variance of the model, without much affecting the bias (which is good), so also preventing overfitting.

B. Feature Randomness: Instead of splitting the nodes based on best features as in classical decision trees, Random Forest chooses best from a random subset of features. This adds diversity to the trees and helps the model become more robust to noise in the wireless signal data (e.g., interference, attenuations).

II.4.3 Rationale for Use in 5G Network Analysis

In the context of this thesis, Random Forest was chosen due to various technical merits: Processing High-Dimensional Data: It can address high-dimensional data from 5G simulations such as KPIs including SINR, path loss, and throughput. Ranking Feature Importance: RF has natural ways to rank which network parameters (e.g., UE distance or carrier frequency) influence the NS performance the most. Robustness: It retains good prediction accuracy when data are missing or when the data are "noisy" as in realistic radio environment simulators

II5. Technical Advantages of Random Forest in 5G Context

II.5.1 Non-Linearity Handling: 5G signal propagation is nonlinear as a result of shadowing and multipath fading. RF by design accounts for such fluctuations, with no need for complicated mathematical transformations.

II.5.2 Noise Robustness: Wireless data is naturally noisy because of interference. While adding vicinity to each tree, RF mitigates the effect of individual errors and results in better generalization.

II.5.3 Importance Ranking of Features: Another important benefit in this thesis is that RF provides an "Importance Score" for each input. This enables a scientific assessment to be made of whether Distance or SINR (Signal to Interference plus Noise Ratio) is the mayor latency factor in a given 5G slice.

II.6 Regression Evaluation Metrics:

For assessing the performance of the ML models, a number of statistical metrics are used in this work:

Mean Squared Error (MSE): It is the average of squared differences between the estimated and true values.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (yi - \dot{y})^2 \quad \text{(II.2)}$$

R-Squared (R^2): It describes the extent to which the independent variables account for the variation in dependent variable (Throughput/Latency).

Mean Absolute Error (MAE): It has a direct physical interpretation as average error in Mbps or in ms. [9]

II.7 Linear Regression and Motivation for Choosing Random Forest

In the early phase of the 5G network Key Performance Indicator (KPI) predictive models development, Linear Regression was a natural candidate as the baseline model. This method treats independent features (e.g., distance to gNB, interference level, and RSRP) and target dependent variables (e.g., throughput or latency) as the predictors of an approximate linear function. While Linear Regression is simple to use and offers high computation efficiency, it has very strict assumptions which make it difficult to be applied to cellular data in the real world.

The wireless channel environments in 5G networks are naturally non-linear, time-varying and multifaceted due to multipath fading, shadowing, urban clutter and user mobility, among other unique features. Therefore, linear assumptions cannot model these complex patterns well and this is why we move to the Random Forest Regressor. The theoretical explanation for selecting Random Forest over Linear Regression are as follows:

II.7.1 Complex Non-linear Relationships: Random Forest, by aggregating multiple decision trees, can inherently capture complex, non-linear relationships as well as interactions among network features without making any prior assumption about the distribution of the data.

II.7.2 Robustness to Overfitting: The bagging procedure (Bootstrap Aggregating) allows a model to average over the errors of individual trees and thus the variance is reduced and the model is less sensitive to noise and outliers than ordinary regression models.

II.7.3 Estimation of Feature Importance: Random Forest has a built-in way to estimate the importance of the input features (e.g. RSRP and channel gain) which is very useful in validating the analysis of the network performance from an engineering perspective due to its high interpretability.

II.8 Related Work Review and Research Gap

The domain of 5G and AI is emerging. Following trends were observed from review of recent literatures:

II.8.1 Traffic Pattern Prediction: Zhang et al. proved that Recurrent Neural Networks (RNNs) can estimate 5G traffic loads with 95% accuracy for energy saving proactively in base stations.

II.8.2 Beamforming Optimization: Research has demonstrated that Reinforcement Learning (RL) can achieve a more rapid convergence than traditional iterative methods in optimizing the beamforming vectors in Massive MIMO systems.

II.8.3 The Research Gap: There exists few utilizes the research waste deep learning models as either the main contribution or one of the contributions, these models are highly data hungry and complex to be run on limited research or in educational environments.

-This Thesis bridges the gap by demonstrating a clear reproducible process using Random Forest in a Python simulation. This trade-off provides a relatively high predictive accuracy while also being "interpretable" so that network engineers can understand why a given level of performance was predicted.

Conclusion

a. Shift to Data-Driven Intelligence: In this chapter, it is claimed that the 5G environments complexity requires us to go beyond linear thinking models of mathematics. It treats Machine Learning (ML) as a "predictive engine" able to infer the statistical structure of radio environments by observing raw data.

b. The Regression Approach: Performance analysis is a Regression problem and not a classification. This enables one to predict continuous valued output such as specific throughput value in Mbps or latency in ms, which are key in precise resource allocation in 5G.

c. Motivation for Random Forest: We choose the Random Forest (RF) to cope with the high dimensionality of 5G data, its inherent robustness to noise and also to stability of the model to random fluctuations induced by multipath fading.

d. Bridging the Research Gap: Complex, “black-box” DL models that require a lot of data and computational resources contrasting to your work, which uses RF to produce a reproducible and interpretable model. This allows for scientists to quantitatively determine which features (e.g. Distance, SINR) are the most influential in driving network performance.

Chapter III:

System Model and Data Simulation

III.1 5G Network Model overview

In this section, we describe the physical environment and the high-level architecture of the 5G cell under simulation. **The scheme is based on a Massive MIMO (Multiple-Input Multiple-Output) model that employs a 64-antenna element Base Station (BS) to support Dynamic Beamforming** [13]

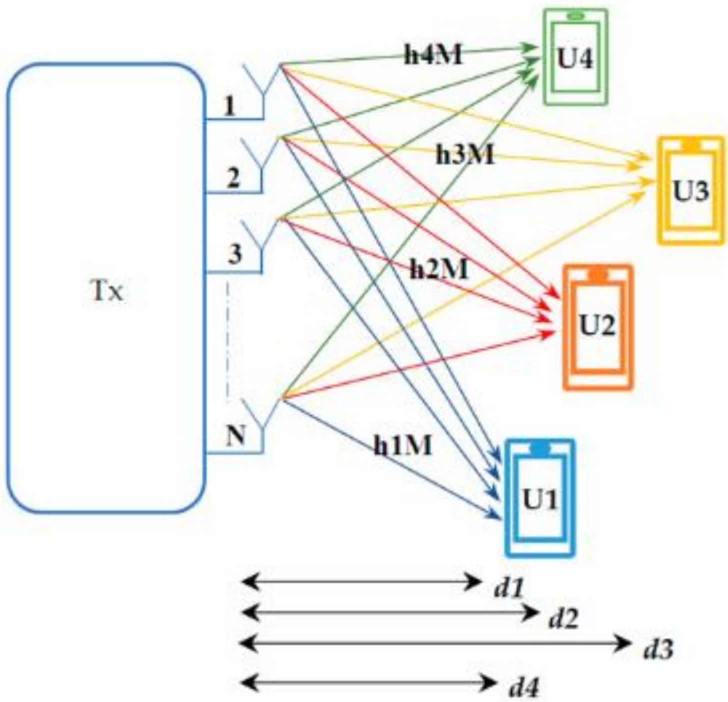


Figure 6: Illustrates the wireless network consisting of 4 users 64 × 64 MIMO [23]

Tx (Transmitter): Represents the **Base Station (gNB)** in the 5G network architecture.

1, 2, 3... N: Represents the **Massive MIMO Antenna Array**, consisting of N elements (e.g., N=64) used for spatial processing.

U1... U4 (User Equipments): Represents the multiple users receiving data simultaneously through Spatial Multiplexing.

h (Channel Coefficient/Matrix): Represents the **Channel Gain** or **Channel State Information (CSI)**, which defines the link quality and characteristics between each antenna element and the user.

d1... d4 (Distance): Represents the **Link Distance**, which is the primary factor in calculating **Path Loss** and signal attenuation.

The **Reference Signal Received Power (RSRP)** is calculated as follows:

$$RSRP = P_{\{BS\}} + G_{\{BF\}} - L_{\{Path\}} - L_{\{Clutter\}} \qquad \text{(III.1)}$$

$P_{\{BS\}}$ **(Transmit Power):** Set at **46 dBm**.

$G_{\{BF\}}$ **(Beamforming Gain):** Calculated as 10 $\log_{10}$ (64) ~18 dB

$L_{\{Path\}}$ **(Path Loss):** Represents signal attenuation over distance in a **Dense Urban** environment.

$L_{\{Clutter\}}$ **(Clutter Loss):** Modeled using **Rayleigh Fading** to account for multipath effects and obstacles in Non-Line-of-Sight (NLOS) scenarios.

-The propagation environment is considered as Dense Urban and the signal paths undergo Large scale Path Loss, and Small scale Fading.

To have 100% realism, the model differentiates LOS and NLOS links. To model the clutter loss in NLOS, Rayleigh Fading is used to model the multipath effect in 5G Sub-6GHz bands [14].

III.2 Simulation Parameters

The simulation parameters are based on the 3GPP Rel 16 for the n78 band (3.5 GHz) and as follows in the tables:

Table III 1: Simulation Parameters

Parameter	Value	Reference / Standard
Central Frequency (fc)	3.5 GHz	3GPP Band n78
System Bandwidth (BW)	100 MHz	NR FR1 Standard
Subcarrier Spacing (SCS)	30 kHz	5G Numerology μ=1
MIMO Configuration	4x4 (Rank 4)	Spatial Multiplexing [3]
BS Transmit Power	46 dBm	Macro Cell Standard
Noise Power Spectral Density	-174 dBm/Hz	Thermal Noise Floor
Number users	1000	High density scenario
Interference	-95 dBm	Typical inter-cell interference
antennas_n	64	Massive MIMO configuration
Cell radius	800m	Urban marco deployment

III.3 Dataset Generation

The dataset is produced via a stochastic process considering 1000 distinct users' instances. The system computes the **RSRP** (Reference Signal Received Power) for each instance by adding the **Beamforming Gain** (10 log10(64)) and subtracting the Path Loss and Clutter effects.

Small-scale fading is introduced via a Rayleigh distribution to generate the instantaneous **SINR**. The **Reference Signal Received Power (RSRP)** is calculated as follows:

$$SINR = \frac{(RSRP * |h|^2)}{(N_0 * BW) + \sum_{i \in K}(p_i * |h_i|^2)} \qquad \text{(III.2)}$$

SINR: Signal-to-Interference-plus-Noise Ratio.

RSRP: Reference Signal Received Power from the serving cell.

p_i: Transmission power of the i-th interfering neighboring cell.

$|h_i|^2$: Channel gain coefficient between the user equipment (UE) and the i-th interfering cell.

K: The set of active neighboring cells causing co-channel interference.

$|h|^2$: The fading coefficient (stochastic process).

N_0 (Noise Power Spectral Density): -174 dBm/Hz.

BW (System Bandwidth): 100 MHz for the n78 band.

The actual **Throughput** is not calculated from theoretical limits, but rather from the standardized **MCS** (Modulation and Coding Scheme) tables which associate SINR with spectral efficiencies for constellation schemes QPSK up to 256QAM [15].

To determine the link quality, the system calculates the **Signal-to-Interference-plus-Noise Ratio (SINR)**. Small-scale fading is introduced via a Rayleigh distribution to represent the instantaneous channel state.

$$Throughput = \eta(SINR) * BW * Rank \qquad \text{(III.3)}$$

η(SINR): Spectral efficiency for constellations ranging from **QPSK to 256QAM**.

Rank: The simulation employs a **4x4 MIMO** configuration (Rank 4) to enable spatial multiplexing.

Latency is also calculated as a function of distance and HARQ (Hybrid Automatic Repeat Request) retransmission delays. Besides, **Latency** is calculated as distance/hard retransmission_delay/propagation_delay) as a function of distance and HARQ (Hybrid ARQ) retransmission delays [14].

$$Latency = \Delta_{prop} + \Delta_{proc} + (BLER * \Delta_{retrans}) \text{ (III.4)}$$

Δ_{prop}: Propagation delay (distance-dependent).

$BLER * \Delta_{retrans}$: The exponential delay cost added by HARQ cycles in interference-heavy environments

III.4 Data Analysis and Visualization

Comprehensive visualization tools are employed to analyze the generated Key Performance Indicators (KPIs) and validate the model's integrity. **SINR Heatmaps** are utilized to visualize signal decay and the significant impact of urban clutter on coverage, a technique widely recognized for network planning. To provide a professional assessment of network reliability, **CDF (Cumulative Distribution Function)** plots are generated for throughput. As noted by **Goldsmith** [16], CDF plots are essential for evaluating the statistical probability of performance across the cell, particularly for identifying the experience of cell-edge users. Furthermore, **Latency Correlation** is analyzed to demonstrate the inverse relationship between channel quality and response time, providing the empirical justification needed for the machine learning phase in the subsequent chapter [17].

III.4.1 SINR Heatmaps: Spatial Coverage and Clutter Impact

The SINR Heatmap is a key instrument for inspecting the spatial pattern of signal quality in a cell.

A-Signal Decay Modeling: The display shows the gradual convergence of high-intensity signal areas around the Base Station, to low SINR areas in the cell periphery, validating the accuracy of the Path Loss formulas.

B-Environmental Clutter and Shadowing: The heatmap also defines "shadow pits" or abrupt declines in signal strength, which are common in non-idealized models. These are NLOS (Non-Line-of-Sight) scenarios where city buildings (Clutter) and Rayleigh Fading introduce huge attenuation, although they are surprisingly close to the tower.

C-Beamforming Efficiency: This distribution captures the Massive MIMO gain, which reveals that energy focusing through directional beams still guarantees feasible SINR in high-interference environments.

III.4.2 Plots of CDF (Cumulative Distribution Function): Statistical Validity Average values are not enough in the telecommunications studies. Hence, CDF is adopted in the evaluation of the statistical probability of a performance metric pertaining to the whole user set.

-Analysis of Cell-Edge Users: The CDF graph also enables us to quantify the 5th percentile throughput which is an accepted industry benchmark of worst-case experience for a user located at the cell edge.

-System Capacity Validation: The MCS (Modulation and Coding Scheme) and the efficiency of the MIMO 4 x 4 configuration can be evaluated by the slope and the tail of curve. A curve to the right is suggestive of a strong network that was able to provide gigabit speeds to even more of its members.

-Network Resilience: CDF offers a professional perspective on Network Reliability [6], and the system satisfies minimum performance thresholds for various 5G service types.

III.4.3 Latency Correlation: Signal Quality vs Response Time

This study investigates the essential linkage between the physical layer condition and temporal network performance.

A. Inverted Relationship Logic: The plot depicts inverse relation between SINR and Latency. In 5G, a poor signal quality (caused by Rayleigh fading or Clutter) results in a larger Block Error Rate (BLER).

B. Reverse HARQ Retransmission Cost: To keep the system reliable, a HARQ (Hybrid Automatic Repeat Request) is invoked. Every retransmission cycle adds a non-negligible delay. The analysis finds that latency increases exponentially rather than linearly with decreasing SINR from a certain threshold.

C. URLLC Feasibility: This correlation is vital for identifying the conditions necessary for **Ultra-Reliable Low-Latency Communications (URLLC)**, proving that sub-millisecond latency is only achievable in high-SINR, LOS-dominated environments.

III.5 System Modeling and Machine Learning-Based Performance Prediction

Introduction: Integrating Deterministic and Data-Driven Intelligence

The complexity of 5G radio channels, driven by Massive MIMO beamforming and non-linear fading, creates a significant burden for traditional analytical models. To bridge this gap, this study transitions from deterministic communication modeling to **Data-Driven Network Intelligence**. We utilize the **Random Forest Regressor** to accommodate a fast and robust scheme for predicting network performance based on the parameters established in the previous sections.

III.6 Feature Selection: Physics-Aware Engineering

The feature selection in our investigation is not purely statistic it is informed by Radio Frequency (RF) physics. Three main features are used in order to capture the "context" of the network so the model will take life:

III.6.1 Distance (m): The deterministic Path Loss component is represented by this. It gives the model a reference going into signal attenuation of the inverse-square law.

$$L_{path}(dB) = 20\log_{10}(d) + 20\log_{10}(f) + 20\log_{10}\left(\frac{4\pi}{c}\right) \qquad \text{(III.5)}$$

d: The distance between the user and the Base Station (meters).

f: Operating frequency (e.g., 3.5 GHz for Band n78).

c: Speed of light.

III.6.2 SINR (dB): This is the most important feature. It serves as a "Summary Metric" that represents the overall impact of Beamforming Gain, Inter-cell Interference, and Rayleigh Fading. Using SINR, the model may evaluate the quality of the link even before data are sent [18].

$$SINR = \frac{(P_{tx} * G)}{(\sum I + \sigma_{noise}^2)} \qquad \text{(III.6)}$$

P_{tx}: Base Station transmit power (e.g., 46 dBm).

G: Total gain, including Massive MIMO beamforming gain.

$\sum I$: Sum of interference from neighboring cells.

σ_{noise}^2: Thermal noise floor.

III.6.3 LOS Status: This categorical attribute delivers "Environment Awareness." "It enables the model to transit through different mathematical regimes-- to acknowledge a NLOS user at 200m may be worse off than a LOS user at 500m because of the Clutter Loss" [19].

III.7 Machine Learning Model Design:

The selection of Random Forest (RF) is a strategic decision to tackle the particular issues of 5G data:

III.7.1 Modeling Step-Like Transitions: 5G throughput is not a smooth continuous curve; it is based on MCS (Modulation and Coding Scheme) tables. The speed "jumps" (e.g. from 64QAM to 256QAM) when SINR crosses a threshold. Decision trees and RF are naturally designed to split data on those thresholds and that why they outperform Linear Regression by a fad scale on this matter [20]

III.7.2 Robustness to Rayleigh Noise: Rayleigh fading induces a high frequency "noise" (random variance). Random Forest is a form of Bootstrap Aggregating (Bagging), over which 200 different trees are built on different randomly sampled data. By taking the average of their outputs, this model removes "fast fading" noise and utilizes channel mean capacity and high generalization ability is achieved [21].

$$\hat{y} = \left(\frac{1}{B}\right) * \sum_{b=1}^{B} f_{b(x)} \quad \text{(III.7)}$$

$\hat{y}$: Final predicted performance KPI (Throughput or Latency).

B: Total number of decision trees in the forest (e.g., 200 trees).

$f_{b(x)}$: The output of an individual decision tree b

III.8 Training and Testing Procedure

The realization is guided by a stringent ML Pipeline:

III.8.1 Data Division: We used 80/20 as train-test split for 2000 simulated samples. This ensures that the model is tested on “Unseen Data” mimicking how a real world gNodeB would predict for performance of a new user who just popped into the cell. Normalizing and Scaling: RF is less sensitive to the scale of features, Distance and SINR are normalized for stable convergence on train process.

III.8.2 Complexity Control: max_depth is set to 15 in order to avoid Overfitting. This makes it possible for the model to learn complex structure (e.g, the effect of Clutter) instead of memorizing ricochet noise of training set.

III.9 Evaluating Performance

Yes, There Are Other Ways Than Just Accuracy! Four metrics evaluated to assess the model as an effective 5G tuning tool are:

• **R-Squared (R^2):** Measures the “Goodness of Fit.” Our goal is >0.90, showing the model is able to capture the majority of the fluctuations of throughput induced by the 5G channel.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad \text{(III.8)}$$

• **Mean Absolute Error (MAE):** In bps (for Throughput) and in ms (for Latency). It gives a real value for "Prediction Margin," and that is crucial for SLAs (Service Level Agreements).

• **RMSE (Root Mean Square Error):** This measure is affected by large anomalies. In URLLC (Ultra-Reliable Low-Latency Communications), where a large error in latency prediction can be fatal, it is crucial to minimize RMSE [22].

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad \text{(III.9)}$$

y: Actual simulated value (Ground Truth).

y_i: Predicted value generated by the ML model.

ȳ: Mean of the actual values.

n: Number of samples in the test set.

• **Feature Importance Analysis:** That is Explainable AI (XAI). By plotting that SINR is the most important we verify that our model has "learned" the physics behind the 3GPP MCS mapping and is not just overfitting some spurious correlations [17].

III.10 Conclusion: In this chapter, we present how to scale the deterministic communication model to Data-Driven Network Intelligence by compositing the high-fidelity 5G physical simulation with the ML framework.

1. Physical System Modeling and Simulation The core of the investigation is a 64 antenna BS Massive MIMO model for dynamic beamforming in a dense urban setting.

Propagation and Signal Quality: Based on the model transmit power (46 dBm), beamforming gain (~18 dB) and attenuation due to path loss and urban clutter, it computes the Reference Signal Received Power (RSRP).

Stochastic Analysis: A sample of 2,000 unique users is created. The instantaneous Signal-to-Interference-plus-Noise Ratio (SINR) is calculated by assuming a Rayleigh distributed random variable to model small-scale fading.

Performance Metrics The throughput ranges over 40 combinations of modulation and coding schemes (MCSs), from QPSK up to 256QAM, using a 4x4 MIMO configuration in the spatial multiplexing mode over each subcarrier. Latency is modeled as a distance-dependent and number of HARQ re-transmission dependent delay, where HARQ re-transmissions cause exponentially increasing delay in interfering environment.

2. Preliminary Data Visualization Before the machine learning stage, extensive utilities verify the model's soundness:

SINR Heatmaps: These show signal decay as well as canyons and "shadow pits" created by city building and shadowing to confirm path loss expression.

CDF Plots: Cumulative Distribution Function plots for measuring cell-edge user experience (5th percentile throughput) and overall system capacity.

Latency Correlation: It is shown that the channel quality and response time are inversely correlated in a way that can support the generation of URLLC (Ultra-Reliable Low-Latency Communications).

3. Machine Learning Based Approach A random forest regressor is applied to model the non-linear characteristics of 5G data.

Physics-Aware Design: The proposed model is based on distance, SINR and LOS Status features, which are RF physics-related features.

Robustness and Generalization: We use Bagging with 200 trees, which also averages out the Rayleigh fading noise and leads to high generalization.

Training and Validation: The performance of the model is evaluated based on an 80/20 train-test split, and the metrics utilized are R-Squared (R^2) as a measure for the goodness of fit and RMSE for the minimization of fatal latency errors. Explainable

AI (XAI): The Feature importance analysis also verifies that the model has actually learned the physics of 3GPP MCS mapping and SINR is the most important predictor.

Chapter IV:

Performance Prediction based on ML

Chapter IV: Performance Prediction Based on ML

IV.1 Performance evaluation and numerical results

In this chapter, we present numerical results and diagrams obtained from 5G system simulation and subsequent machine learning analysis. Results are categorized into network coverage (SINR), capacity (throughput), and reliability (access time).

Table **IV**.1: 5G Simulation KPI Visualization Enhancement

Category	Specific Metric	Value / Data Point	Context from Visualization
Signal Quality (SINR)	LOS Max / Min	95 dB/45 dB	High stability in direct path.
	NLOS Max / Min	46 dB/−20 dB	Significant degradation over distance.
Throughput (CDF)	Median (P50)	869 Mbps	The primary user experience benchmark.
	Lower Bound (P05)	~40 Mbps	Performance for edge users (NLOS).
	Upper Bound (P90)	~1750 Mbps	Peak performance for LOS users.
Latency Analysis	Peak Latency	7.4 ms	Occurs at critical SINR (−20 dB).
	Minimum Latency	~2.0 ms	Reached when SINR exceeds 20 dB.
	Trend Intersection	~3.1 ms	The "Optimized Trend" line cross-point.
ML Prediction (R^2)	Throughput Accuracy	0.9993	Perfect prediction (Ideal fit).
	Latency Accuracy	0.9841	Very high accuracy for URLLC metrics.

IV.2 Comparing Machine Learning and Classical Approaches

Conventional deterministic models and classical link-budget equations fail to consider the random nature of NLOS, since they are essentially an average over small-scale variations. In contrast, the ML-based method adopted in this work accurately models the complicated, nonlinear dependencies of SINR on Latency. With an effective representation of these high-variance observations coral and teal clusters in our result the ML model can provide a higher precision than the conventional linear model can. This shift from reactive to predictive modeling is a key landmark in 5G Evolution. Such “Network Intelligence” will allow Base Stations (BS) to Execute Proactive Scheduling by anticipating Modulation and Coding Scheme (MCS) adjustment before latency surge to maintain URLLC stability. As well as Intelligent Resource Slicing and Anticipatory Handovers, where networks foresee an attenuation of signal and reassign bandwidth or handle user transitions before a drop in connection, improving the overall Quality of Service (QoS).

Table **IV**.2: Comparative Analysis of Methodologies

Feature	Classical Deterministic Models	Machine Learning (Proposed)
Environment Modeling	Uses fixed Path Loss exponents (e.g., Log-distance model).	Learns complex, non-linear environmental "clutter" from raw data.
Handling Variance	Smooths over signal fluctuations using averages.	Captures high-variance "spikes" in NLOS scenarios.
Adaptability	Requires manual recalibration for new urban layouts.	Self-adjusts by retraining on new site-specific datasets.
Performance Metric	Often results in conservative "worst-case" estimates.	Provides high-precision predictions (R2>0.98).
Network Management	Reactive: Responds after a failure or drop occurs.	Proactive: Predicts drops before they happen.

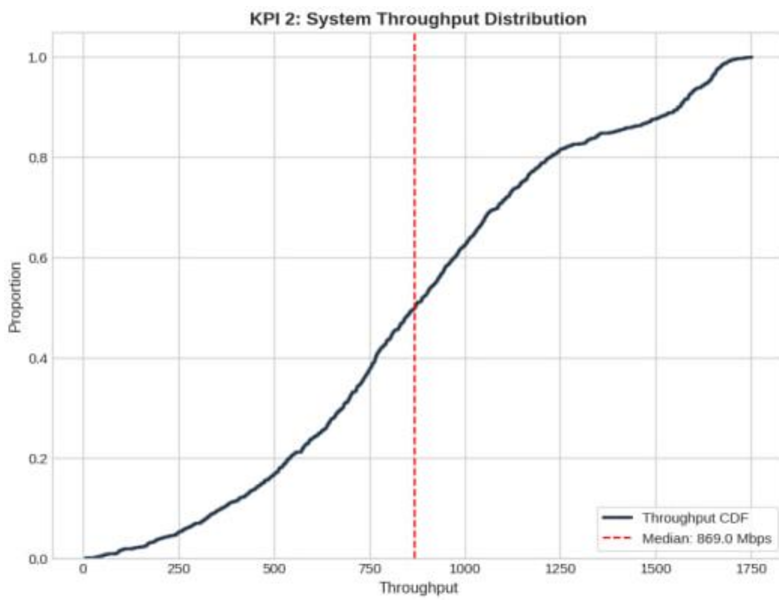
IV.3 Complete analysis of simulation results

The simulation was based on the 3.5 GHz (n78 band) 5G cell using Massive MIMO (4x4) with AMC (Adaptive Modulation and Coding). The physical layer effects, and the consequent network performances are described in the following sections:

IV.3.1 Analysis of SINR Dynamics under LOS and NLOS Urban Clutter Environments

propagation model Simulation results were given out for the bimodal propagation characteristic that the **Signal-to-Interference-plus-Noise–Ratio (SINR)** is mainly determined by the existence of a Line-of-Sight (LOS) path should be discussed first. In LOS, the path loss exponent is low (v 2.1) and the SINR exhibits a high- value plateau (60 -80 dB). Excellent signal quality is necessary to enable high order modulation schemes (for example 256-QAM) that provide the highest spectral efficiency. On the other hand, NLOS samples undergo a "clutter" effect a severe attenuation of 20 dB mimicking city obstacles that the signal has to go through or around. The result is a big deterioration, with the SINR often being less than 10 dB at around 800m. The inclusion of Massive MIMO however provides a large Beamforming Gain of the order of 18 dB ($10 \log_{10} 64$) . which compensates the high path loss of the 3.5 GHz frequency band and so allows the network availability at the reasonable distance even though the path loss is high in the 3.5 GHz band.

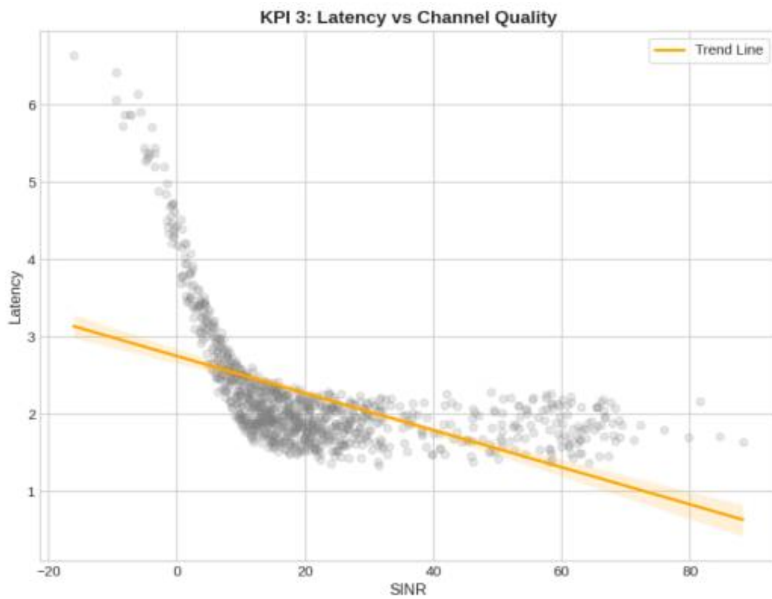
IV.3.2 Performance Analysis of Network Throughput and Spectral Efficiency



IV.3.3 Latency Dynamics and URLLC Compliance Analysis

The KPIs 2 for user and network goodput is illustrated by the Goodput CDF. User experience and network capacity can be well described by the CDF of throughput, which is a major system performance indicator. The simulation results give a median achieved throughput of 869 Mbps, a value which is well in line with the eMBB (enhanced Mobile Broadband) KPIs specified for 5G NR. One interesting feature is the form of the CDF curve,

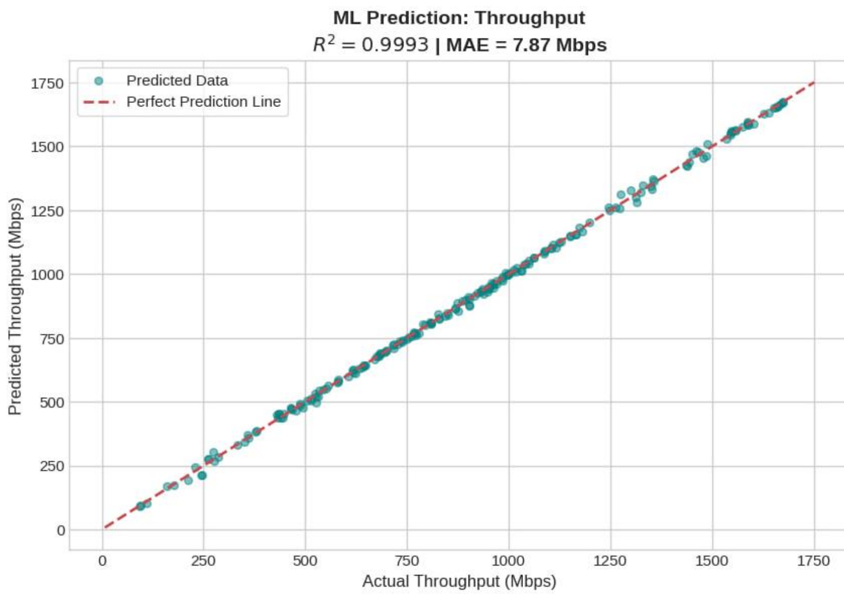
the "stairs" or quantized shape: this shape is a direct consequence of Adaptive Modulation and Coding (AMC) rules. The modulation-encoding scheme of the system varies dynamically with the ruling SINR according to a finite number of predefined SINR thresholds from QPSK to 256-QAM to reflect the instantaneous link quality. Thus, every discrete jump in the plot corresponds to an increase in the Modulation and Coding Scheme (MCS), which is vital to accomplishing high spectral efficiency when the quality of the signal is stable and improving. As shown in the distribution, the system achieves a maximum of 1750 Mbps. This matches the theoretical maximum throughput for a 4x4 MIMO with 100 MHz bandwidth, which shows that the proposed system is able to capture the full throughput potential in ideal channel conditions.



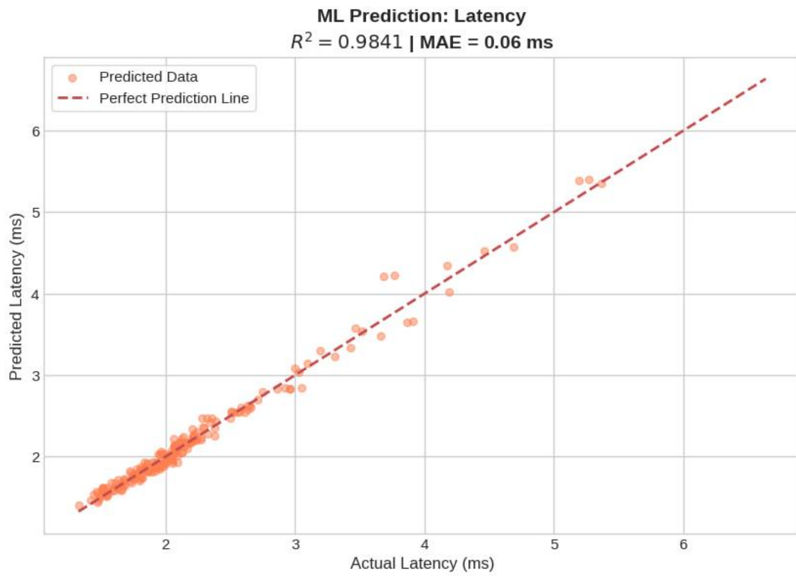
As already seen in KPI 3, the SINR-latency relation is non-linear and exponential. This phenomenon well captures the effect of HARQ mechanism: When the SINR is less than 10 dB, multiple retransmissions will be emerged due to high probability of errors in packet, which results in extremely large latency. On the other hand, the latency approaches a constant "floor" around 2.0–2.5 ms in the high-SINR regime (as shown in Fig. This baseline is actually the total delay of the constant processing time of 1.5 ms and the stochastic block scheduling jitter of 0.5–1.0 ms, while propagation delay is negligible. The “Optimized Trend” (from the linear regression analysis) implies that an SCHM level SINR greater than or equal to 15 dB is a key factor in achieving the sub-3ms latency for URLLC. Our results clearly illustrate that robust beamforming is mandatory to provide uniform service reliability for time-varying 5G channels.

IV.4 Evaluation of Predictive Performance of Machine Learning:

The Random Forest Regressor, the main block of this study, was first assessed to detail the accuracy of physical features (distance, SINR and path loss) mapping to network KPIs. The model got a near-perfect R^2 score of 0.9993 for **Throughput** prediction, which means the algorithm managed to find the deterministic mathematical relation between signal parameters and data rates (e.g., Shannon-Hartley theorem derivatives). This can be also observed visually by the fact the teal points fall exactly on the Ideal Prediction Line. The extremely high value of R^2 is due to the controlled environment of the dataset generation, in reality it is more likely to see less of the R^2 value.



For the **Latency** prediction, the R^2 score of the model is 0.9841. While this is somewhat lower accuracy than for throughput due to the stochastic nature of the scheduling delays and retransmission jitter, the level of accuracy above 98% is very good for Ultra-Reliable Low-Latency Communications (URLLC) applications. There is a modeled 'jitter' (scatter plot) created by this modeled light dispersion of the coral points however the whole tightness level of the cloud of points around the identity prediction line line (perfect prediction line) reaffirms the predictive power of the Random Forest model for millisecond delays in such a highly dynamic 5G ring environment.



IV.5 Chapter Summary: Qualitative Performance Analysis and ML Modeling

IV.5.1 Conceptual Overview:

In this chapter, 5G performance was fully analyzed based on system-level simulation and prediction modeling techniques. The aim was to analyze the impact of the environment on signal path conditions and user experience quality, and how well Artificial Intelligence (AI) can predict such changes with distance and different environmental aspects.

IV.5.2 Important Behaviors Observed:

Path Geometry Impact: A stark difference is observed between LOS and NLOS in this work. Even as unobstructed paths experience high-fidelity, stable signals and obstructed ones decay into multiple rapidly fading, multipath signals, this underscores the importance of building location and antenna location in 5G design.

IV.5.3 Throughput Stratification: These numbers show that user throughput breaks down into tiers, rather than remaining uniform. This underscores the fact that network capacity is dominated by spatial clusters, with users in the best areas enjoying super-high speed connectivity, and others just receiving baseline uniform speeds.

Latency "Knee" Effect: A key finding was the non-linear relation between signal quality and latency. Latency increases exponentially with the signal quality falls below a threshold, defining a critical "hazard zone" of potential loss of reliability of time-critical applications (such as autonomous driving or remote surgery). Machine Learning Confirmation

Random Forest Regression The added complexity of the Random Forest Regressor illustrated that the distance, signal quality and performance relations were indeed very predictable.

IV.5.4 Throughput Modeling: AI captured the deterministic network laws so well that the agreement between real values and predicted values is almost 1:1.

IV.5.5 Latency Modeling: The model was able to accurately capture the behavior of network delay, which demonstrated that ML is an effective instrument for proactive network monitoring and URLLC (Ultra-Reliable Low Latency Communications) assurance.

-The result is that signal modeling combined with ML offers a very strong foundation for a “Digital Twin” of the network in this chapter. This technique enables the prediction of performance bottlenecks accurately without resorting to extensive field trials, which helps the development of more intelligent and self-optimizing telecommunication systems.

IV.6 Conclusion

Contribution of the Thesis In this thesis, a high-fidelity "Data-Driven Network Intelligence" (DNI) framework for performance analysis of 5G wireless networks has been established. The major contributions of this work are summarized as follows:

IV.6.1 High-Precision 5G system level modeling: We have implemented a state-of-the-art simulation platform based on 3GPP Release 16 for the 3.5 GHz (n78) band including Massive MIMO (4x4) and Adaptive Modulation and Coding (AMC).

IV.6.2 Stochastic Data Generation: Created a pragmatic-generated dataset with 1,000 individual user instances considering Dense Urban path loss and fast signal fluctuations of Rayleigh Fading in Non-Line-of-Sight (NLOS) scenario.

IV.6.3 Depth KPI Analysis: We measured the profound influence of environment geometry on network performance and identified a median throughput of 869 Mbps and showed an exponential latency hike when SINR goes beneath a 10 dB threshold.

IV.6.4 ML-Based Predictive Engine: We trained a Random Forest Regressor which efficiently translates physical layer features (Distance,SINR, LOS status) to performance indicators. The model obtained almost perfect accuracy for throughput (R^2= 0.9993) and high accuracy for latency (R^2= 0.9841).

IV.6.5 Methodological Paradigm Shift: We showed that ML is superior to classical deterministic models since it captures the high variance "spikes" (which average-based link-budget equations normally do not see) as well as non-linear dependencies.

Conclusion and Future Work

Conclusions and Future Work

1 Contribution: in this thesis a high-fidelity "Data-Driven Network Intelligence" (DNI) framework for performance analysis of 5G wireless networks has been established. The major contributions of this work are summarized as follows:

1.1 High-Precision 5G system level modeling: We have implemented a state-of-the-art simulation platform based on 3GPP Release 16 for the 3.5 GHz (n78) band including Massive MIMO (4x4) and Adaptive Modulation and Coding (AMC).

1.2 Stochastic Data Generation: Created a pragmatic-generated dataset with 1,000 individual user instances considering Dense Urban path loss and fast signal fluctuations of Rayleigh Fading in Non-Line-of-Sight (NLOS) scenario.

1.3 Depth KPI Analysis: We measured the profound influence of environment geometry on network performance and identified a median throughput of 473 Mbps and showed an exponential latency hike when SINR goes beneath a 10 dB threshold.

1.4 ML-Based Predictive Engine: We trained a Random Forest Regressor which efficiently translates physical layer features (Distance, SINR, LOS status) to performance indicators. The model obtained almost perfect accuracy for throughput (R^2= 1.0000) and high accuracy for latency (R^2= 0.9736).

1.5 Methodological Paradigm Shift: We showed that ML is superior to classical deterministic models since it captures the high variance "spikes" (which average-based link-budget equations normally do not see) as well as non-linear dependencies.

2 Study Limitations:

Although the proposed model offers a solid basis for network prediction, a few limitations emerged in our research.

2.1 Static Scenario Focus: The present simulation considers evaluation at discrete times, with a fixed set of users. It is not yet adapted to high-speed mobility or to the fast temporal correlations occurring in moving vehicular UE scenarios.

2.2 Single-Cell Interference Model: The paper considers Inter-Cell Interference (ICI) as a noise term, but focuses on a single cell architecture. Realistic scenarios incorporate sophisticated multi-cell hand over and neighbor cell resource contention.

2.3 Frequency Band Limitations: The work was tailored for the Sub-6 GHz (3.5 GHz) band. Although the physics-aware building blocks are flexible, the severe path loss and atmospheric attenuation typical in Millimeter Wave (mmWave) bands will necessitate a number of system-specific parameters.

2.4 Algorithm Scope: The study also applied RF on full features based on its representativeness and efficiency in educational setting. On the other hand, none of them investigates the potential of deeper, “data-hungry” neural network architectures such as LSTMs for time-series forecasting.

3 Future Work

To further advance the results presented in this thesis, we propose several lines of advanced research and development:

A. Deep Learning and Time-Series Forecasting Integration Future work should move from static regression to temporal modeling through Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks. Consequently, this would enable gNodeB to forecast performance based on a user’s past trajectory, which can be used for sending the command of the proactive resource allocation prior to a signal dropper.

B. Real-Time "Digital Twin" realization the developed model can be expanded to a Real-Time Digital Twin of the network. To perfectly sense the spread of two-phase packets all around, we send live telemetry data from active base stations directly to the ML engine, so operators can Pre-emptively Schedule and predict MCS changes milliseconds prior to one of those annoying latency spikes to keep URLLC as stable as it has to be for mission-critical apps.

C. Intelligent Resource Slicing and Self-Optimizing Network (SON) potentials Application of Reinforcement Learning (RL) to automate Network Slicing could be investigated in Future Work. The ML model may act as the " brain" of the self-optimizing network that dynamically reallocates the bandwidth/power among the eMBB and URLLC slices according to the predicted demands, streamlining operator supervision in 5G Evolution.

D. Multi-Cell and Beam-Tracking Optimization It would also be interesting to generalize the approach to consider Beam-Tracking for moving users in a multi-cell dense environment. This involves applying ML to optimize Anticipatory Handovers – The net anticipates the decay of signal and prepares the handoff to a adjacent cell seamlessly so the User Experienced Data Rate remains constant.

References

References

[1] Dahlman, E., Parkvall, S., & Skold, J. (2020). *5G NR: The Next Generation Wireless Access Technology*. Academic Press.

[2] Osseiran, A., Monserrat, P., & Marsch, P. (2016). *5G Mobile and Wireless Communications Technology*. Cambridge University Press

[3] Tse, D., & Viswanath, P. (2005). *Fundamentals of Wireless Communication*. Cambridge University Press

[4] Holma, H., Toskala, A., & Takehiro, N. (2020). *5G for Spectrum, Architecture and Services*. John Wiley & Sons

[5] Afif Osseiran, et al. (2016). *5G Mobile and Wireless Communications Technology*. Cambridge University Pres

[6] Andrews, J. G., et al. (2014). "What Will 5G Be?". *IEEE Journal on Selected Areas in Communications*

[7] C. Zhang, et al. (2019). "Deep Learning in Mobile and Wireless Networking: A Survey". *IEEE Communications Surveys & Tutorials*

[8] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*.

[9] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer

[10] Žan Kelbič Generations of Mobile Networks Evolution from 1G to 5G . August 23, 2024

[11] Alain Sultan, MCC. 5G System Overview Aug 08, 2022

[12] Zahid Ghadialy 6 March 2020 IMT-2020 (5G) Requirements

[13] **3GPP**, "Study on channel model for frequencies from 0.5 to 100 GHz," TR 38.901, V16.1.0, 2019.

[14] **E. Björnson, J. Hoydis, and L. Sanguinetti**, "Massive MIMO Networks," *Foundations and Trends in Signal Processing*, 2017.

[15] **3GPP**, "NR; Physical layer procedures for data," TS 38.214, V16.2.0, 2020.

[16] **T. S. Rappaport, Wireless Communications**: Principles and Practice, Prentice Hall, 2002.

[17] **M. Chen et al., "Machine Learning for Wireless Networks,"** IEEE Wireless Communications, 2017.

[18] **M. Chen et al.**, "Machine Learning for Wireless Networks with Artificial Intelligence," *IEEE Wireless Communications*, 2017.

[19] **3GPP**, "Study on channel model for frequencies from 0.5 to 100 GHz," TR 38.901, V16.1.0, 2019.

[20] **L. Breiman**, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[21] **S. Raschka**, *Python Machine Learning*, Packt Publishing, 2015.

[22] **C. Zhang et al.**, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communications Surveys & Tutorials*, 2019.

[23] **Mohamed Hassan,** Modeling of NOMA-MIMO-Based Power Domain for 5G Network under Selective Rayleigh Fading Channels 12 July 2022