الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة الشهيد الشيخ العربي التبسي – تبسة
Echahid Cheikh Larbi Tebessi University-Tebessa
Faculty of Science and Technology
Department of Electronics and Telecommunications

Master's Thesis

Submitted for the degree of Academic Master

Field: Telecommunications

Specialty : Networks and Telecommunications

Presented by:

- Gaidi Ahmed
- Hamdi Safi Eddine

Title

Improving The Performance of Next-Generation Network Communication using Machine Learning (ML)

Defended and evaluated on 10-11 / 06 / 2025 , by the jury composed of:

First and last name		Rank	Position
Mr.	Meraoumia Abdellah	Prof.	Chair
Mrs.	Aouiche Chaima	MCB	Supervisor
Mr.	Houam Lotfi	MCB	Reviewer

Graduating Class: 2024/2025

"It is a capital mistake to theorize before one has data."

— SIR ARTHUR CONAN DOYLE
(WRITING AS SHERLOCK HOLMES)

Abstract

The rapid development of wireless networks with the massive increase in the number of users and demands for high-speed data have pose real challenges in managing high frequency bands such as millimeter waves, which offer high data rates, but their limited coverage is a drawback which calls for the need for advanced technologies such as machine learning (ML). In this study, we explore the use of various ML models to predict throughput performance in 6G networks. We used the Lumos5G dataset, a widely recognized real-world dataset after rigorous processing. We trained and tested these models, including Random Forest, XGBoost, GradientBoosting, XTREE and advanced stacking with multiple meta-models, optimizing their performance through the Optuna hyperparameter tuning framework. Furthermore, we generated synthetic samples using generative adversarial networks (GANs) to address sample shortages and select key features to improve model‘s accuracy. The results indicated significant improvements in throughput prediction accuracy, which was quantified using metrics such as the root mean square error (RMSE) and coefficient of determination (R2), confirming the effectiveness of the proposed technique.

Keywords: Wireless Networks, Throughput, Machine Learning, Ensemble learning, Synthetic learning, Terahertz, mmWave, GAN, Stacking GBM.

الملخص

يشكل التطور السريع للشبكات اللاسلكية مع الزيادة الهائلة في عددها، لقد فرض المستخدمون والطلب على البيانات عالية السرعة تحديات حقيقية في إدارة الترددات العالية النطاقات مثل الموجات المليمترية، والتي توفر معدلات بيانات عالية، ولكن تغطيتها المحدودة هي العيب الذي يستدعي الحاجة إلى تقنيات متقدمة مثل التعلم الآلي (ML).

في هذه الدراسة، نستكشف استخدام نماذج التعلم الآلي المختلفة للتنبؤ بأداء الإنتاجية في شبكات 6G. لقد استخدمنا مجموعة بيانات Lumos5G، وهي مجموعة بيانات واقعية معترف بها على نطاق واسع بعد إجراء اختبارات صارمة للمعالجة. لقد قمنا بتدريب واختبار هذه النماذج، بما في ذلك Random Forest وXGBoost وGradient Boosting وXTREE والتكديس المتقدم مع نماذج وصفية متعددة، مما أدى إلى تحسين الأداء من خلال إطار ضبط المعلمات الفائقة Optuna.

علاوة على ذلك، تم إنشاء عينات اصطناعية باستخدام Generative Adversarial Networks (GANs) لمعالجة نقص العينة واختيار الميزات الرئيسية لتحسين دقة النموذج. وأشارت النتائج إلى تحسينات ملحوظة في دقة التنبؤ بالإنتاجية، والتي تم قياسها باستخدام مقاييس مثل Root Mean Square Error (RMSE) ومعامل التحديد (R^2)، مما يؤكد فعالية التقنية المقترحة.

الكلمات المفتاحية: الشبكات اللاسلكية، الإنتاجية، التعلم الآلي، التعلم الجماعي، التعلم الاصطناعي، تيراهرتز، mmWave، GAN، تكديس GBM

Résumé

Le développement rapide des réseaux sans fil avec l'augmentation massive du nombre de les utilisateurs et la demande de données à haute vitesse posent des défis réels dans la gestion de la haute fréquence bandes telles que les ondes millimétriques, qui offrent des taux de données élevés, mais leur couverture limitée est un Le désavantage de l'apprentissage automatique (ML), qui nécessite des technologies de pointe. En cette étude, nous explorons l'utilisation de divers modèles ML pour prédire la performance du débit en 6G réseaux. Nous avons utilisé l'ensemble de données Lumos5G, un ensemble de données traitement. Nous avons formé et testé ces modèles, y compris la forêt aléatoire, XGBoost, GradientBoosting, XTREE et stacking avancé avec plusieurs méta-modèles, performance à travers le cadre de réglage des hyperparamètres Optuna. En outre, nous des échantillons synthétiques générés à l'aide de réseaux antagonistes génératifs (GAN) pour cibler l'échantillon les pénuries et de sélectionner des caractéristiques clés pour améliorer la précision du modèle. Les résultats ont indiqué amélioration de la précision des prédictions de débit, qui a été quantifiée à l'aide de mesures telles que l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R2), confirmant la l'efficacité de la technique proposée.

Mots-clés : Réseaux sans fil, Débit, Apprentissage machine, Apprentissage d'ensemble, Synthétique apprentissage, Terahertz, mmWave, GAN, empilement de GBM.

Acknowledgments

I would like to express my sincerest thanks to my beloved families, particularly my **lovely mother**, wise **father** and my **Brothers**, for their strong moral encouragement, perseverance, and constant inspiration in pursuing my studies.

A particular mention of thanks is for my **grandmother** who has always been a tranquil refuge and a well of peace. Since it was where I lived, it was very important in providing the ideal setting that enabled me to concentrate and stay strong in preparing this work.

I also have a great debt of thanks to our friends who happened to be with me and provided me with encouragement and motivation at all the stages of this work. I especially want to quote my close friend **Achref Graidia**, whose encouragement and friendship mattered the most.

I extend my deepest gratitude and appreciation to my dear parents, especially my caring mother who has always been my source of support and strength. I also thank my dear sister who saved my life with her donation, which I don't know how to repay. Words cannot express how grateful I am to her. I thank everyone who supported me on this journey.

We would also like to acknowledge our sincerest gratitude to the **Aissat-Idir Telecommunication Center of Algérie Télécom** Team, and Miss **Souhila Baghiri** for her generous supervision and practical counseling and assistance she extended to us during our internship.

We also want to express our heartfelt gratitude to our academic supervisor, **Chaima Aouiche**, for her constant direction, incisive comments, and wise suggestions which were vital to the completion of this task. Our gratitude is also directed to all the staff members of Jury : **Pr. Meraoumia Abdellah** as chair and **Dr. Houam Lotfi** as Reviewer we also directed to all the staff members of **Electrical Department** of **Echahid Cheikh Larbi Tebessi University** for providing a lively academic climate and for sharing the ideas and inspiration they imparted to us while studying.

Lastly, we would like to thank all such people **directly or indirectly** who helped us in finishing this work. We wish you all prosperity for your assistance.

Ahmed Gaidi
Safi Edine Hamdi

Table of Contents

Abstract ..... **I**
Résumé ..... **II**
Acknowledgements ..... **III**
Table of Contents ..... **IV**
List of Figures ..... **V**
List of Tables ..... **VI**
List of Abbreviations ..... **VII**

General Introduction 1

Chapter 1: Background .....
1.1 Introduction 4
1.2 Background and Evolution of Wireless Networks 4
1.3 Challenges in 5G Network Communication 8
1.4 From 5G To Six Generation Transition 13
1.5 Key Problems and Motivation 14
1.6 Conclusion 15

Chapter 2: Six Genrtation and Machine Learning – A Perfect Synergy .....
2.1 Introduction 16
2.2 Six Generation Wireless Network Overview 16
2.3 Machine Learning Overview and Insights 24
2.4 Six Generation and ML Synergy 31
2.5 Conclusion 35

Chapter 3: Methodology .....
3.1 Introduction 36
3.2 Dataset Collection and Features Description 36
3.3 Data Preprocessing 39
3.4 Experimentation Settings and Tools 43
3.5 Identifying the Problem and Labelling 44
3.6 Conclusion 48

Chapter 4: Results and Discussions .....

4.1 Introduction 49
4.2 Performance Metrics in ML 50
4.3 Model Training & Hyperparameter Optimization Strategy 52
4.4 Models Validation 53
4.5 Results and Discussion 56
4.6 Comparison Between Models with and Without Synthetic Data 65
4.7 Ensemble Model Techniquesn without Synthetic Data 68
4.8 Conclusion 72

Chapter 5: Conclusion and Future Direction .....
5.1 Challenges and Limitations 73
5.2 Future Directions and 7G 74

General Conclusion 75

Bibliography 76-

LIST OF FIGURES

Chapter 1

Fig. 1.1 3G architecture 5
Fig. 1.2 4G architecture 6
Fig. 1.3 5G architecture 7
Fig. 1.4 Wireless technology speed comparison 7
Fig. 1.5 Uplink/downlink speed trends 8
Fig. 1.6 Bandwidth allocation across generations 6
Fig. 1.7 Overview of low band (Sub-1 GHz) spectrum .. 9
Fig. 1.8 Overview of mid-band (Sub-6 GHz) spectrum . 10
Fig. 1.9 Overview of millimeter wave (mmWave) spectrum 10
Fig. 1.10 Distance impact on signal in mmWave 11
Fig. 1.11 Number of user impact on capacity 12
Fig. 1.12 Paving the Way to 6G Through Innovation .. 13

Chapter 2

Fig. 2.1 6G visions and requirements 17
Fig. 2.2 VIVO proposal for 6G architecture 18
Fig. 2.3 Significant technologies of 6G 19
Fig. 2.4 Overview of terahertz (THz) spectrum 20
Fig. 2.5 Massive MIMO and beamforming 21
Fig. 2.6 Cell massive MIMO 21
Fig. 2.7 Significant technologies of 6G 23
Fig. 2.8 Unsupervised learning significant type 25
Fig. 2.9 Types of regression models and descriptions .. 30
Fig. 2.10 GANs working flow 31

Chapter 3

Fig. 3.1 Percentage of NaN values before cleaning ... 40
Fig. 3.2 Percentage of NaN values after cleaning 41
Fig. 3.3 Key features in cleaned Lumos5G dataset ... 42
Fig. 3.4 Impact of various factors on throughput (%) .. 45
Fig. 3.5 nr-ssRsrp effect on throughput 46
Fig. 3.6 lte-rsrp effect on throughput 46
Fig. 3.7 Moving speed effect on throughput 46

Chapter 4

Fig. 4.1 Pipeline diagram of strategies 53
Fig. 4.2 Flowchart of grouped features configuration .. 54
Fig. 4.3 Comparison R^2 MAE RMSE between methods . 55
Fig. 4.5 Performance comparison of ML models 57
Fig. 4.6 True vs predicted throughput (LightGBM) ... 58
Fig. 4.7 True vs predicted throughput (Random Forest) . 58
Fig. 4.8 True vs predicted throughput (XGBoost) ... 58
Fig. 4.9 True vs predicted throughput (Xtree) 58
Fig. 4.10 True vs predicted throughput (Gradient Boosting) 58
Fig. 4.11 Throughput distribution: original vs generated . 60
Fig. 4.12 Distribution of 3 features comparison 60
Fig. 4.13 Comparison of Means and Standard Deviations . 61
Fig. 4.14 Model performance after GANs 63
Fig. 4.15 True vs predicted (Gradient Boosting) GANs .. 63
Fig. 4.16 True vs predicted (LightGBM) after GANs .. 64
Fig. 4.17 True vs predicted (Random Forest) after GANs . 64
Fig. 4.18 True vs predicted (XGBoost) after GANs ... 64
Fig. 4.19 True vs predicted (Xtree) after GANs 64
Fig. 4.20 Model comparison based on R^2 65
Fig. 4.21 Model comparison based on MAE 66
Fig. 4.22 Model comparison based on RMSE 66
Fig. 4.23 Performance after ensemble techniques 70
Fig. 4.24 Performance comparison after ensemble 71

LIST OF TABLES

Chapter 2

Tab. 2.1 Key Performance Indicators of 5G and 6G Communication Systems 16
Tab. 2.2 Network Architectures and ML Algorithms 34

Chapter 3

Tab. 3.1 Lumos5G dataset overview in a tabular format 39

Chapter 4

Tab. 4.1 Benchmarking Regression Performance Using R^2, MAE, and RMSE Thresholds 51
Tab. 4.2 Hyperparameter of the models 53
Tab. 4.3 R^2, MAE, and RMSE evaluation 56
Tab. 4.4 Hyperparameters of models after GANs 62
Tab. 4.5 Benchmarking Regression Performance Using R^2, MAE, and RMSE after GANs 62
Tab. 4.6 Model Performance after synthetic data description 67
Tab. 4.7 Hyperparameters of models after GANs 69
Tab. 4.8 Benchmarking Regression Performance Using R^2, MAE, and RMSE after GANs 69

List Of Abbreviations

6G	Sixth Generation Mobile Network.
LTE	Long-Term Evolution .
THz	Terahertz (High-frequency electromagnetic waves).
MIMO	Multiple Input Multiple Output.
RIS	Reconfigurable Intelligent Surface.
UAV	Unmanned Aerial Vehicle.
VLC	Visible Light Communication.
FSO	Free Space Optics.
NOMA	Non-Orthogonal Multiple Access.
AI	Artificial Intelligence.
ML	Machine Learning.
RF	Random Forest .
GBM	Gradient Boosting Machine.
XGBoost	eXtreme Gradient Boosting.
LightGBM	Light Gradient Boosting Machine.
MAE	Mean Absolute Error.
RMSE	Root Mean Squared Error.
MSE	Mean Squared Error.
R^2	Coefficient of Determination.
mmWave	Millimeter Wave .
Sub-6 GHz	Frequencies below 6 GHz used in 5G.
WCDMA	Wideband Code Division Multiple Access .
Mbps	Megabits per second .
Gbps	Gigabits per second.

GENERAL INTRODUCTION

The explosive increase in mobile data traffic has led to the design of even more sophisticated wireless networks. Despite the fact that fifth-generation (5G) technology has greatly elevated the speed of data transmission as well as reliability, its constraints in handling the demands of future technologies and automation have facilitated the evolution towards establishing sixth-generation (6G) networks. The future deployment of 6G is expected to revolutionize mobile communication by integrating advanced technologies, such as artificial intelligence (AI), machine learning (ML), intelligent reflecting surfaces, terahertz communication, and ultra-massive MIMO, to achieve extremely fast data transmission, ultra-reliable low-latency communication, and smart network management.

Owing to the complexity and the high computational needs involved in optimizing these emerging technologies, traditional mathematical models are often inadequate. Thus, AI and ML are increasingly being utilized to enhance network efficiency, maximize resource utilization, and enable self-optimizing systems. This provides a comprehensive overview of 6G networks, their technological underpinnings, predicted features, and the key role played by AI-based solutions in addressing the significant challenges. The research also delves into the deployment timelines, possible applications in different industries, and regulatory and technical challenges to be surmounted to tap the complete potential of 6G.

Thesis outline

The outline of this Thesis is as follows:

• **Chapter 1**: Offers a historical perspective of wireless communications networks development, ranging from initial mobile technologies to the emergence of 5G. The chapter emphasizes 5G's fundamental technology advancements such as improved mobile broadband, lower latency communication and greater network capacity. It also brings to mind 5G's significant challenges to performance and deployment, for instance spectrum scarcity, expense of infrastructure, energy efficiency and the challenge of providing ultra reliable links to multiple use cases

• **Chapter 2**: Its overview of 6G technologies, the most prominent innovations such as Terahertz communication, UltraMassive MIMO and Reconfigurable Intelligent Surfaces and how they are crucial for enabling next generation connectivity. The chapter goes on to describe some machine learning methods, such as ensemble learning and synthetic data usage, that improve performance and flexibility in 6G networks. It ends with the discussion of AI applications within 6G and how they are pivotal to constructing intelligent, efficient and smart communication systems.

• **Chapter 3:** The research methodology framework commencing with an explicit introduction to the Lumos5G dataset, i.e., its setup, origins and why it was employed. It sets the characteristics and nature of the data gathered. The chapter then briefly discusses data preprocessing operations like column filtering, missing and categorical data handling, and normalization methods applied to standardize data. It then explains the experimental setting, such as the programming environments, core libraries, and hardware specifications employed. It then sets the core analytical problem and how meaningful labels are established through a process that gives rise to model building later on

• **Chapter 4**: It gives a detailed review of the result of the experiments highlighting model performance metrics. It offers an explanation of the model choice and the process of hyperparameter tuning and subsequent validation procedures for attaining credible results. Two experiment phases are compared in the chapter utilizing real data in isolation and the combination of real and synthetic data to measure the data augmentation effect. Also it investigates utilizing ensemble learning methods as a new

strategy to improve model performance and prediction accuracy without relying on synthetic data.The results are fully described with a critical discussion of interest to the interpretation of the observations extracted from the experimental model performance based on the strengths and weaknesses developed throughout the study.

• **Chapter 5**: This last chapter summarizes the research in the sense of providing an overview of the key contributions, albeit conceding some of the overall problems in the research area. It also provides directions for future research, especially concerning future 7G networks and ends with a general overview of the study's main findings.

Chapter I

Background

1.1 Introduction

Wireless communications have evolved significantly from the basic analog voice services of 1G to the high-speed, broadband capability of 5G each a noteworthy technological leap, from voice calls in 1G and digital text in 2G, to mobile internet and multimedia in 3G, high speed data in 4G, and ultra-fast, low-latency connectivity in 5G. As we are approaching the 6G era, this new generation can transform the industry again but also brings along new challenges: operating at higher frequency bands can lead to spectrum shortage and increased interference, and the ultra-dense characteristic of future networks raises issues in terms of energy efficiency, security, privacy, and hardware limitations. This study analyzes these challenges in the context of the overall development of wireless networks, thereby laying the foundation for the innovation required to fully exploit the potential of 6G.

1.2 Background of wireless networks

We provide a concise summary of the progression of mobile wireless technology from 1G to 6G , emphasizing significant breakthroughs in each generation :

- 1G came into being in the early 1980s using analog signals to facilitate basic voice calls with low speed and bad sound quality. AMPS, TACS, and NMT were some of the technologies that built the platform for mobile communication [1] .

- 2G: The introduction of second-generation (2G) mobile networks in the early 1990s marked a shift from analog to digital technology. Along with traditional voice services, 2G networks introduced new capabilities such as short message service (SMS) and basic email functionality. 2G networks also improved audio quality and enhanced security. Some of the well-known 2G (second generation) mobile networks include GSM (Global System for Mobile Communications), IS-95 (Interim Standard-95), PDC (Personal Digital Cellular), and CDMAone (Code Division Multiple Access)[1][11].

- 3G:In the early 2000s, 3G supported both voice and data services,

paving the way for mobile internet, multimedia, and video streaming. Developments such as CDMA2000, WCDMA, and EDGE continued to advance the capabilities of mobile communications. the 3G mobile network architecture combines circuit-switched and packet-switched domains to deliver voice-messaging and broadband data services. At the line's edge, the radio access network (RAN) comprises Node B base stations, which establish wireless communication with user equipment and Radio Network Controllers (RNCs), that manage radio resources allocation handovers and signal processing. The RAN interfaces with the Core Network (CN) which bifurcate into two functional layers: the circuit-switched domain, governed by Mobile Switching Centers (MSC), for routing voice calls and SMS through legacy networks (e.g The Gateway MSC (GMSC) acts as a bridge between the 3G network and external circuit-switched carriers while the IP Multimedia Subsystem IMS allows advanced IP-based services. Data traffic traverses the IP-based network backbone, while voice calls utilize dedicated circuit-switched infrastructure. The simultaneous support allows real-time and best-effort services to be delivered efficiently This hybrid architecture reflects the dual functionality of 3G technology blending legacy telephony frameworks with modern IP-centric data delivery [1] [12]. The described architecture is represented in Figure 1.1 :

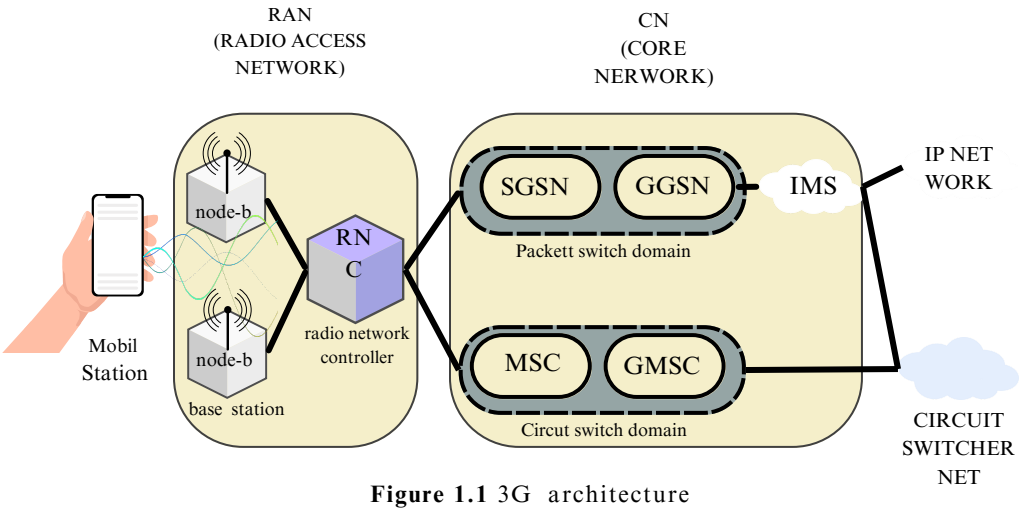


Figure 1.1 3G architecture

- 4G: In the early 2010s, 4G delivered high-speed data transfer and broader coverage, facilitating HD video streaming, mobile conferencing, and gaming. With LTE and WiMAX, 4G set a new benchmark for mobile broadband.The 4G LTE network architecture replaces the 3.5G hybrid circuit with a fully IP-based flat structure optimized for high-speed data and low latency.The Radio Access Network (E-UTRAN) consists of eNBs (Evolved Node B) that integrate

The Radio Access Network (E-UTRAN) consists of eNBs (Evolved Node B) that integrate the roles and functions of 3G's Node B and RNC managing communication, mobility and resource allocation directly with user equipment (UE). The Evolved Packet Core streamlines the core system into key components: the MME (Mobility Management Entity) handles UE authentication and session management, the SGW (Serving Gateway) routes data traffic between eNBs and the core, The PGW connects to external IP networks while enabling the transition of voice services to VoLTE (Voice over LTE)[1] [13], delivered entirely via IP via the IMS (IP Multimedia Subsystem) by eliminating legacy circuit-switched infrastructure. This simplified, all-IP architecture reduces latency improve scalability and unifies voice and data traffic through a single packet-switched framework as is illustrated in Figure 1.2 :

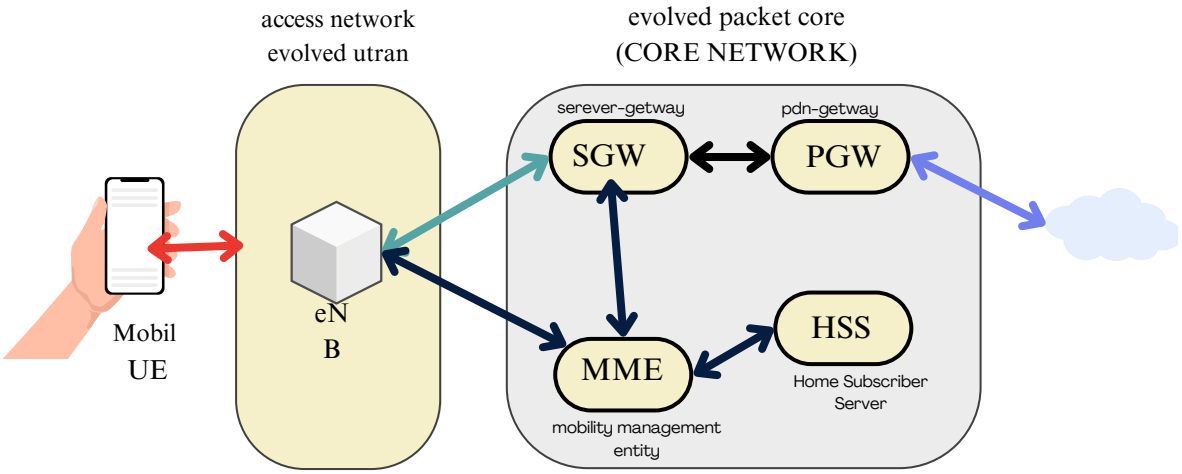


Figure 1.2 4G architecture

- 5G: Since around 2018, 5G delivers ultra fast speeds, ultra low latency, and massive connectivity, and is the cornerstone of IoT, smart cities, and industrial use cases. It is transforming the way we interact with technology. The 5G network adopts a Service-based Architecture (SBA) with a clear separation between control plane and user plane. The Radio Access Network (RAN) is made up of gNBs (Next-Generation Node Bs), which connect User Equipment to the network, handling radio transmission and beamforming. The 5G Core FiveGC introduces modular functions: the AMF management of access and mobility is responsible for UE registration, mobility and security.

The SMF management is responsible for setting up modifications and terminated sessions; and the user plan transport function (UFP) acts as the anchor for high-speed traffic. Supporting functions for QoS and security include the authentication server function AUSF for secure authentication, the UDM (Unified Data Management) for subscriber data storage (replace HSS), and the policy control function PCF (Pol It enables dynamic network scaling, edge computing and ultra-reliable low latency communications (URLLC) by fully decoupling hardware from software for scalability [1] [14] [15]. The design is summarized in Figure 1.3 :

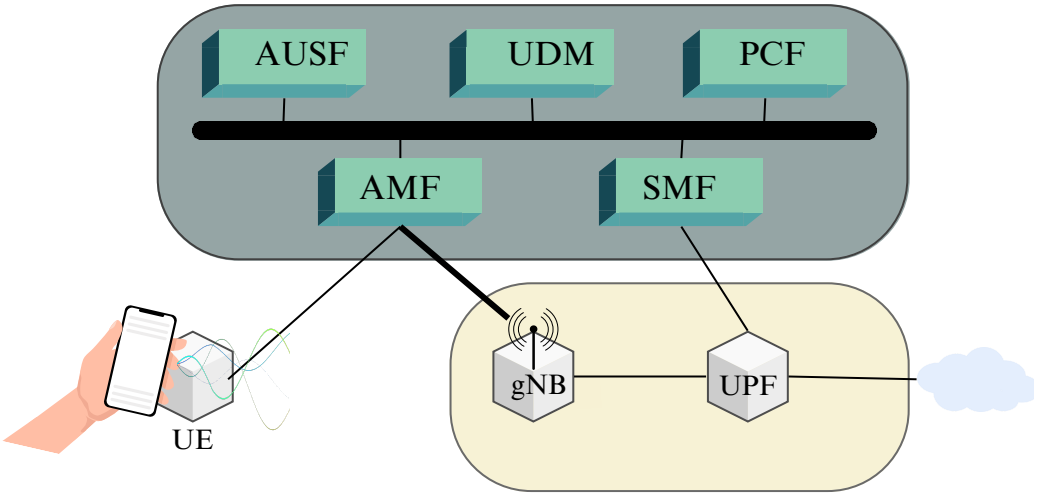


Figure 1.3 5G architecture

- The evolution from 3G to 5g reflects a monumental shift in mobile connectivity driven by advancements in speed bandwidth and latency. While 3G pioneered mobile internet with basic data and voice services 4G activates high-speed broadband for streamingand apps, and 5g now redefine connectivity with ultra-fast speeds and near-instant responsiveness. Below in figure 1.4 is a general comparison of the capabilities they have:

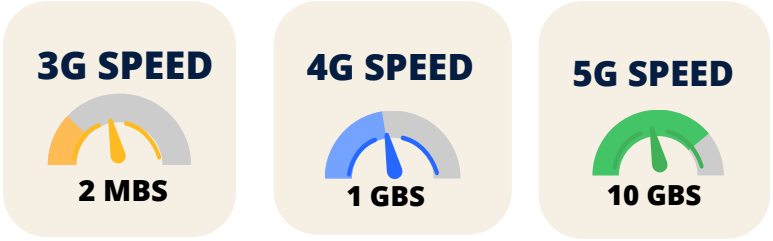
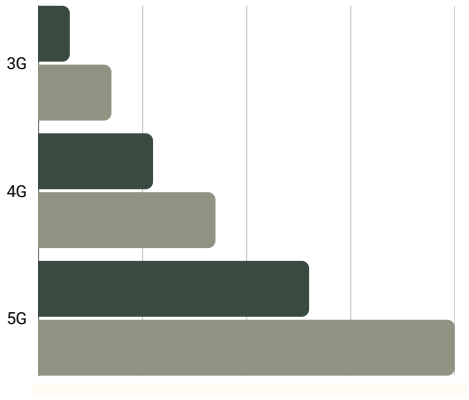
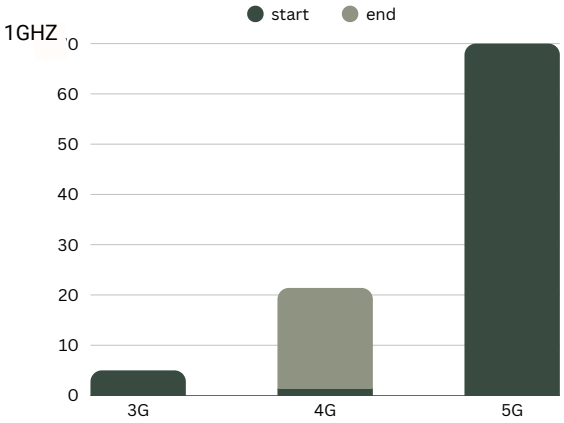


Figure 1.4 wireless technology speed comparison



highlights the progression of uplink and downlink speeds across 3G, 4G, and 5G technologies. As mobile networks have evolved, each generation has introduced significant improvements in data transfer capabilities



the progression of bandwidth from 3G to 4G and 5G technologies. Each generation brings an increase in bandwidth, supporting faster speeds and more connected devices

general explanation of Uplink and Downlink

3G	Uplink max:5.76 Mbps	Downlink max:42 Mbps
4G	Uplink max: 50 Mbps	Downlink max:300 Mbps
5G	Uplink max: to 2 Gbps	Downlink max: to 10 Gbps

Figure 1.5 Uplink/Downlink Speed Trends

general explanation of bandwith range

3G	5 MHz (WCDMA)
4G	1.4 MHz to 20 MHz
5G	100 MHz (Sub-6 GHz)/1 GHz (mmWave)

Figure 1.6 Bandwidth Allocation Across Generations

1.3 Challenges in 5G Network Communication

Fifth generation of 5G technology pioneers a monumental shift in global connectivity that revolutionize how we interact with the digital world and improve our lives. 5G delivers unprecedented data speeds of up to **10 Gbps**, almost instantaneous latency and huge network capacity to support billions without interruption. Mobile 5G not only enhances mobile experiences but also fuels transformative advances in intelligent cities, autonomous systems and augmented reality etc. Its reliance on the high frequency band mmWave limits signal penetration while the need for dense infrastructure drives up costs and energy needs. Security risks, fragmented global standards and public scepticism further challenge its seamless expansion. Interconnected e-computing infrastructure that is balancing its groundbreaking potential with these complexes will define the future of connectivity in 21St Century [2] [5]. However, we now recite the most important problems of the fifth generation such as:

1.3.1 Bandwidth (Millimeter Waves)

The performance of 5G networks depends heavily on the bandwidth available in the network. It is based on the use of various frequency bands to balance performance between speed and coverage, where these frequencies are divided into three basic bands :

- **Low Band (Sub-1 GHz) :** Provides excellent coverage up to **10 km** and good penetration of obstacles (trees, walls or even natural terrain , .ect). Although its connection is stable, it has the disadvantage of low data transfer speeds (100–250 Mbps) necessary to meet the growing needs for high-speed applications [16]

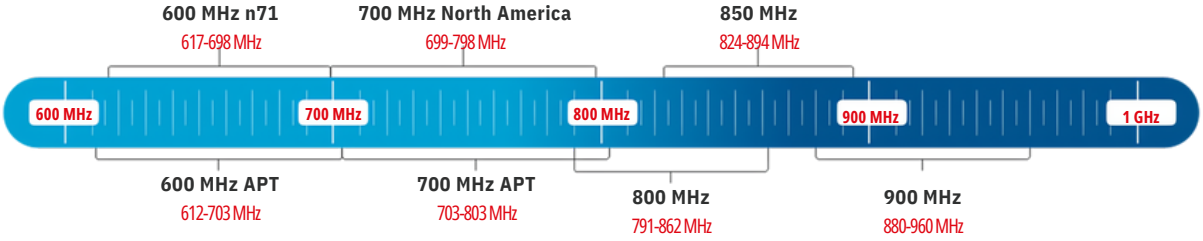


Figure 1.7 An ourview of theLow Band (Sub-1 GHz) Spectrum [41]

- **Mid-Band (Sub-6 GHz) :** This band is characterized by a balance between data transfer speed and coverage, operating at frequencies not exceeding 6 GHz, and is best suited for general use in high-density urban areas, providing speeds between 500 Mbps and 1 GHz per second, with latency periods between 20 and 30 milliseconds. This is what gives balance to this band. Regardless of this, it faces a real challenge of congestion due to the continuous increase in the number of users, which requires the development of technologies to keep up with this increase [2] [5] [16].

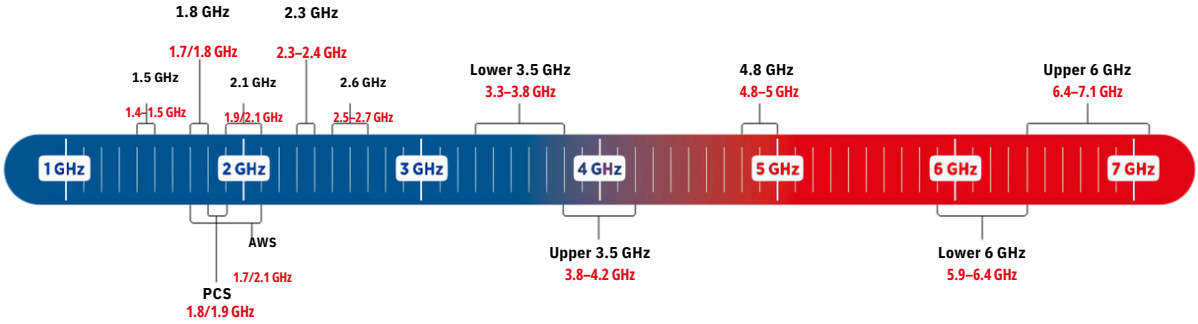


Figure 1.8 An Overview of the Mid-Band (Sub-6 GHz) Spectrum in 5G Networks [42]

- **High Band (mmWave – Up to 24 GHz and above) :** These are the high frequencies that give very high transmission speeds, often exceeding **3 Gbps** and up to **10 Gbps**, with a relatively low latency of **10 ms**, are ideal for applications that require ultra-high speeds such as virtual reality and real-time applications that do not tolerate delays [2].

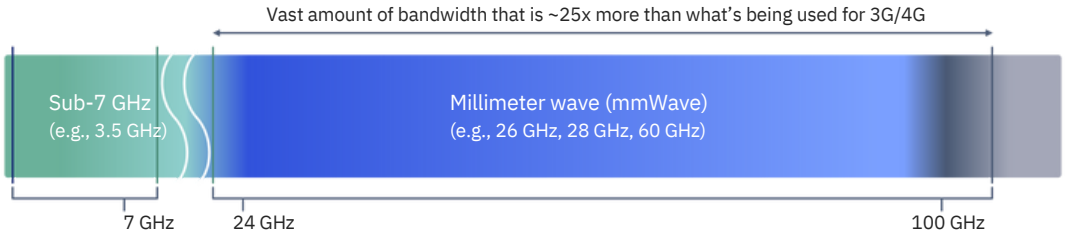


Figure 1.9 An Overview of the Millimeter wave (mmWave) Spectrum [43]

- Although this band is efficient and useful for real-time applications, it has very poor coverage (**200-300 meters**) and is sensitive to obstacles, requiring a high density of communication towers and costly infrastructure. It was also found that the 5G network suffers from great challenges in accommodating the huge number of connected devices and their various requirements of high speeds and low interference, which calls for finding innovative solutions to optimize performance in the future [16].

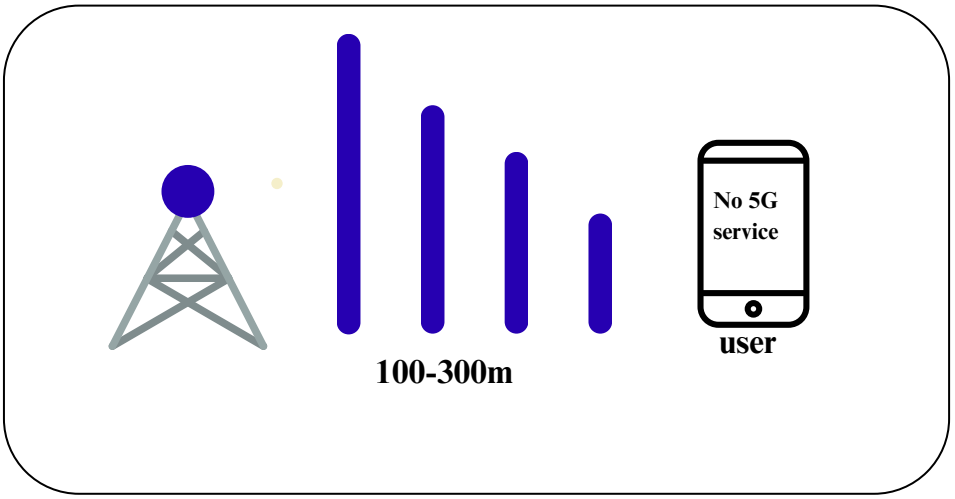


Figure 1.10 distance impact on signal in mmwave

1.3.2 Capacity and Security

Network capacity refers to the ability of the network to efficiently provide data services and meet user demands. Although 5G networks provide good performance and speed, traffic congestion resulting from the interconnection of numerous devices at the same time can negatively impact service quality [2][5][16]. Despite the implementation of advanced technologies such as Massive MIMO, beamforming, and network slicing, maintaining performance in the face of excessive load remains difficult. The most important capacity challenges are:

- **Data Demand :** Due to the increase in the quality of the content (e.g. 4k/8k video),which exceeds the network's capacity to transmit the large amount of data, which requires more capacity.
- **Increasing number of devices :** The network accommodates about 1 million users/km^2. More than that increases the pressure on the network
- **High frequencies :** we refer to millimeter frequencies, which offer very-high speed but suffer from absorption and interference, which makes the network has limited capacity in high-density areas.
- **Unequal distribution :** This is an uncontrollable issue. The number of users is unevenly distributed, limiting capacity in some areas.

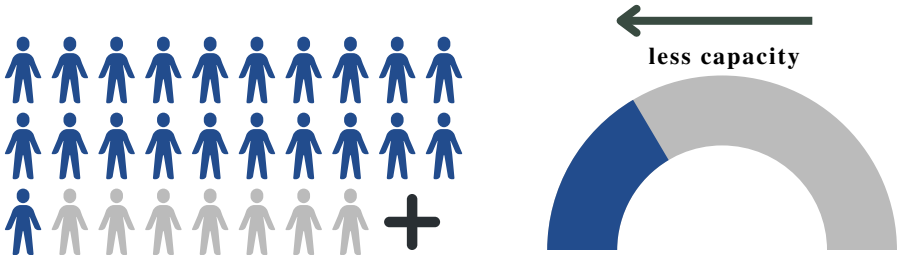


Figure 1.11 number of user impact on capacity

On the other hand, 5G network security threats are a serious issue as the intricate topology of the network and the support for massive device connectivity expose it to increased vulnerabilities to cyberattacks and espionage [2]. The key threats are:

- **Cyberattacks:** The advanced wireless nature of 5G enables potential data interception on route. Common examples include:

Man-in-the-Middle (MITM) attacks: Whereby an attacker takes control of communication between two parties without their knowledge, typically in public insecure Wi-Fi or IoT networks.

Distributed Denial of Service (DDoS) attacks: These flood servers with fake traffic, resulting in service outage and financial loss.

- **Infrastructure complexity:** 5G's adoption of virtualized networks and edge computing de-centralizes data processing, which expands the attack surface and raises issues about confidentiality and integrity of data being exchanged.

Balancing security and capacity in 5G networks is a serious challenge because boosting capacity to support more devices and higher data rates increases the potential for security threat due to the high density, decentralized state of the network. While technologies like mm-wave boost capacity, they also open avenues for threats like cyberattacks and make network security more difficult.

1.4 From 5G To Six Generation Transition

The ever-increasing demand for mobile broadband and the expansion into vertical industries (e.g., automotive, satellite, IoT) continue to drive the evolution of standards within bodies such as the 3rd Generation Partnership Project (3GPP). This progression is evident through the successful standardization of both the fourth generation (4G) Long-Term Evolution (LTE) and the fifth generation (5G) New Radio (NR). In particular, 5G NR has undergone three significant releases, providing a versatile framework capable of supporting a comprehensive range of use cases across various spectrums, adaptable to multiple deployment scenarios whether indoor or outdoor, macro or small cells.

The next pivotal phase in this evolution is 5G-Advanced, a term introduced by 3GPP in April 2021, representing the bridge between 5G and the sixth generation (6G) mobile networks. Building on the strong foundation laid by 5G, 5G-Advanced is designed to introduce enhanced capabilities that will not only boost performance but also expand 5G's reach into new verticals and use cases. Simultaneously, it will incorporate forward-looking innovations that will pave the way toward 6G development, 5G's capabilities will significantly expand, transforming connectivity experiences and accommodating a wide variety of new applications and services. This evolution serves as a critical step on the path to 6G, where even more advanced technologies will redefine mobile communications, unlocking new possibilities for industries and users alike. As the industry prepares for 6G, the innovations introduced through 5G-Advanced will be instrumental in shaping the future of global connectivity [3].



Figure 1.12 : Paving the Way to 6G Through Innovation [3]

1.5 key problems and motivation

Despite 5G advancements, key challenges still limit network performance. Tackling these issues is essential for progress. Our study key problems are:

- **Unstable throughput** in today's networks, which adversely affects user experience quality.
- **Limitations and disadvantages of mmWave** frequency usage, i.e., lower penetrability, enormous path loss, and short coverage distance, making stable connectivity challenging particularly in cluttered environments.
- **Limited availability of correct and realistic data** characterizing such intricate events in mmWave and other wireless networks, restricting efficient modeling, simulation, and optimization of network behavior.
- **Limited body of current research and studies that employed AI methods** specifically in mmWave communications, creating a knowledge gap and less practical AI solutions for such frequencies.
- **Requirement for strong evaluation methodologies** and techniques to compare synthetic data generated with actual data, making the AI models and network simulations trustworthy.

This study is motivated by the following key reasons :

- Enhance the performance of current networks by understanding jittery throughput reasons more profoundly and how these affect users' experience.
- Push the frequency boundaries for mmWave by creating models and data reflecting an authentic picture of the complexity of these frequencies to offer best coverage and performance.
- Fill the gap between having realistic and accurate data for researchers and developers to develop and test more efficient network solutions.
- Provide for more accurate and robust training of AI models with high quality synthetic data simulating real environments, expediting AI development in telecommunications
- Produce trustworthy evaluation tools for measuring generated data accuracy against original data, guaranteeing model and simulation credibility in real network environments.
- Create new areas of applied research in AI and mmWave communications through the supply of datasets and analytical models that contribute to solving existing challenges.

1.6 Conclusion

The history and development of wireless networks underscore the swift progress from initial mobile generations to the present 5G, with each step advancing considerably in terms of speed, capacity, and connectivity. The transition from 4G LTE to 5G New Radio (NR) has brought in strong frameworks that can support a broad set of use cases across various industries and verticals, representing a significant advance in mobile broadband technology.

5G communication continues to be challenged by requirements such as additional spectrum efficiency, energy consumption optimization, infrastructure deployment, and ultra-reliable and low-latency communication assurance. In addition, the evolving demand for increased coverage, deeper inter-sector integration, and shifting security challenges continue to strain the performance potential of existing 5G. As we move on from 5G to 6G, the future will be spearheaded by some of the major breakthroughs such as 5G-Advanced, which adds new features to provide improved performance and more usage scenarios.This will pave the way for 6G, which will bring even more futuristic technologies such as THz spectrum, AI-driven networks, immersive services, and resilient communication systems that are secure and scalable. 6G will revolutionize human communication and connectivity, eventually resulting in a more interconnected, intelligent, and augmented world.

Chapter 2
Six Generation and ML - A Perfect Synergy

2.1 Introduction

6G promises to be the next giant leap of wireless technology with ultra-high speeds, almost zero latency, and a huge increase in connectivity capability. It looks forward to the day when wireless communication is made high-speed-enabled in a seamless manner to cater to future applications like immersive augmented reality, smart cities, and advanced IoT networks. Concurrently, ML has already begun transforming data processing, system optimization, and decision-making processes in millions of industries.By integrating ML with 6G networks, we create a firm synergy that enhances network management and performance as well as provides opportunities for novel, adaptive services. 6G and ML, combined, will create a more intelligent, responsive, and efficient digital world, initiating a new age of connectivity and technological advancement.

2.2 Six Generation Overview

This table compares 5G and 6G performance metrics, showing 6G's significant advancements in speed, latency, and connectivity capabilities....

Key performance Indicator	5G	6G
Technologies	mMIMO, mmWave bands, network densification, polar codes, LDPC codes, cloud/fog/edge computing	Cell-free mMIMO, RIS, AI, Quantum communications & computing
Multiplexing/Multiple Access Techniques	OFDM, Adaptive Time-Frequency Multiplexing, GFDM, FBMC	Spatial multiplexing, OMA, NOMA, OAM multiplexing
End-to-End Latency	1 msec to 10 msec	0.1 msec
Data rates	1 Gbps to 30 Gbps	100 Gbps to 1Tbps
Reliability (Packet Error Rate)	$1\text{-}10^{-5}$	$1\text{-}10^{-9}$
Mobility	500 kmph	1000 kmph
Energy efficiency	--	1 Tb/J
Delay Jitter	--	10 msec
Peak data rates	20 Gbps	1 Tbps
Experienced data rates	0.1Gbps	1 Tbps 100 bits/sec/Hz
Peak spectral efficiency	30 bits/sec/Hz	3bits/sec/Hz
Experienced spectral efficiency	0.3 bits/sec/Hz	100GHz
Maximum channel bandwidth	1 GHz	10^3 Devices/$100m^2$
Connection density	10^2 Devices/$100m^2$	10 nsec
Processing Delay	100 nsec	THz Bands, Non-RF (VLC, optical range),
Frequency bands	Sub-GHz, mmWave	block-chain based spectrum sharing
Use cases	eMBB, URLLC, mMTC	MBRLLC, mURLLC, human-centric services
Device types	Smart phones, sensors, drones	Smart implants, CRAS, sensors and DLT, XR and BCI
Traffic capacity	10 Mbps/m^2	10 Gbps/m^3
Localization precision	10cm in 2-dimensional	1cm in 3-dimensional
Reliability	99.9%	99.999%
Receiver sensitivity	About -120dBm	Less than -130dBm

Table 2.1 Key Performance Indicators of 5G and 6G Communication Systems. [44]

2.2.1 Six Generation Visions and Requirements

6G networks promise to transform wireless communication by leveraging the synergy of a combination of next-generation technologies that together augment connectivity, reliability, efficiency, and user experience. Zero-energy IoT based on energy harvested from the environment to power battery-less devices reduces environmental footprint and costs, while high-speed connectivity with THz frequencies, masive MIMO, beamforming, and full-duplexing offers the maximum data rates of up to 1 Tbps [19]. In parallel, URLLC supports ultra-reliable and low-latency communication that is critical for real-time services, and high availability and reliability are attained through redundancy, load balancing, and failover.

Sophisticated security procedures guard sensitive information in various frequencies, and harmonized integration with present 5G, Wi-Fi, and terrestrial/non-terrestrial networks provides coverage around the globe with uniform performance. In addition, scalability is provided by heterogeneous networks, network slicing, SDN, hybrid cloud, and AI optimizations to accommodate a large number of M2M and IoT devices, while tailored QoE by edge computing and AI/ML adapts the network performance to the needs of different users. As shown in Figure 2.1, such technologies work synergistically to establish the groundwork for the groundbreaking opportunity of 6G [4] [20].

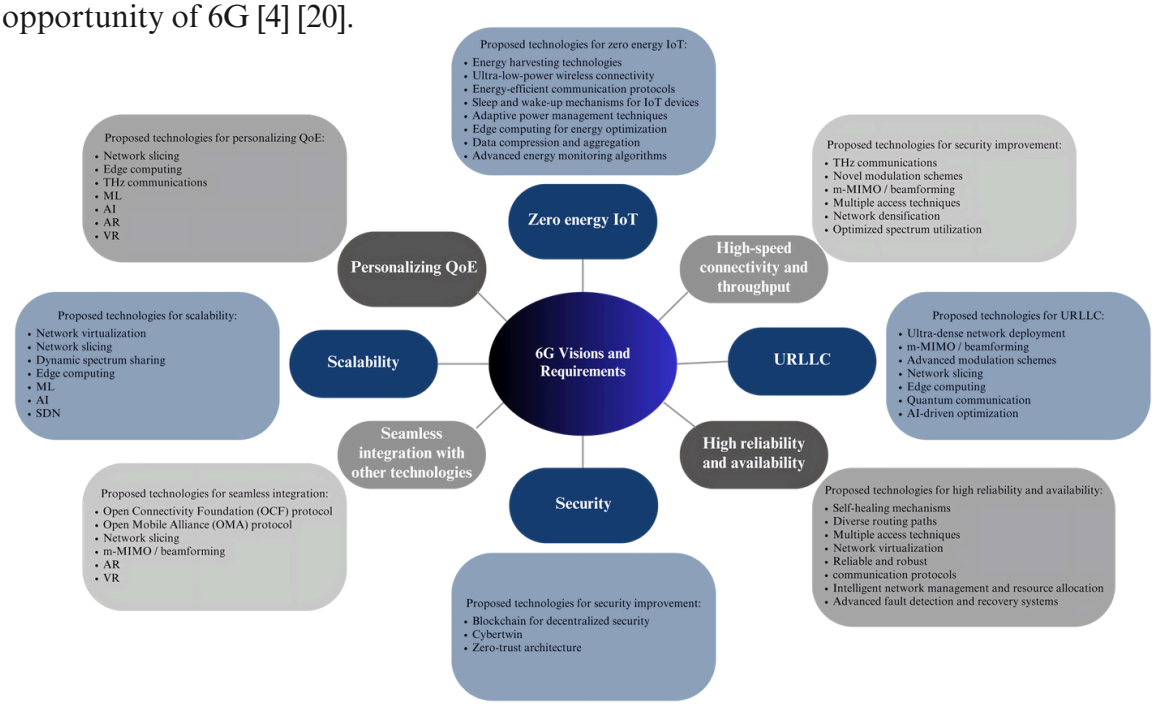


Figure 2.1 6G visions and requirements[4].

2.2.2 Six Generation Network Architecture

The diagram in Figure 2.2, adapted from Vivo's 6G network architecture proposal, illustrates the various functions, with each component working together to perform a specific role that contributes to fast, secure, and intelligent real-time connectivity [6] [21]:

- **Unified Data Management:** Central control function which hosts and manages the user data of identity and sessions and exposes them to subscribers in a synchronized way with support for edge computing in order to speedily retrieve as opposed to retrieve over a suite of databases.
- **Policy Control & Charging:** It is responsible for defining and implementing policies such as speed, priority, and quality, as well as accounting, where the correct fees are charged to the user (billing). It is coupled with the UDM, which it provides with information and is responsible for customizing
- **AI Function:** It relies on AI technologies to analyze actual-time network information and user behavior in order to make the network more interactive, to automatically decide and predict failures ahead of time and thereby make the network self-ruling.
- **Almost Zero Power IoT Function:** Dedicated to empowering and enabling Internet of Things devices that consume little or no power, as they are energized by renewable environmental energy, e.g., various sensors in agriculture, which operate for years directly without any human intervention.
- **Mobility Management:** Responsible for tracking user movement while using the network, as they move from one area to another, which allows ensuring continuity of connection by seamlessly switching between access points. It works in coordination with other functions such as UDM updating the site and transferring policies.
- **Data Control Function:** The intelligent network policy implementer monitors and analyzes data flow in real time to make immediate decisions to avoid network congestion. It also coordinates routes based on traffic flow. It integrates with the policy function to receive appropriate policies to ensure a balance between quality and security.
- **Computing Function:** Responsible for delivering the necessary computing resources, storage, and processing, close to the user. In other words, edge computing enables the network to support time-sensitive, delay-intensive

services, such as autonomous driving or telemedicine. It integrates with AI to facilitate network analysis and enhance AI functionality.

- **Sensing Function:** The network not just transmits information, but it also utilizes and takes advantage of the re-reflected wireless signals to learn about what is happening around it by decoding these signals to detect objects, movement, distance, direction, etc., and these are integrated into the network's infrastructure in the towers. This makes the network a living part that interacts with the environment.
- **Communication Management Function:** Monitoring and managing network traffic by allocating frequencies and data packets based on user needs, such as calls or high-speed data. It also manages seamless network exchange from one tower to another without interruption. It also relies on AI to learn patterns, which provides adaptability and flexibility in the sixth- generation network.
- **Data Forwarding Function**: Transferring data from one point to another inside or outside the network. By identifying the best paths for data and directing packets to the appropriate and safe path to reach the receiving end. It also efficiently deals with network congestion by changing the route to reduce delay. In the sixth generation, this function is smart because it relies on AI and is also distributed at the edges of the network.
- **Data Repository Function:** It stores all types of data that the network handles, such as sensitive data and network performance data, as well as sensor data that the next generation network relies on heavily. It also provides quick access to this data, storing it according to priority, which facilitates other functions.
- **Communication Node**: A go-between agent receives and processes signals from devices to determine priority and response priority, ensuring effective coordination with all available devices within the range.

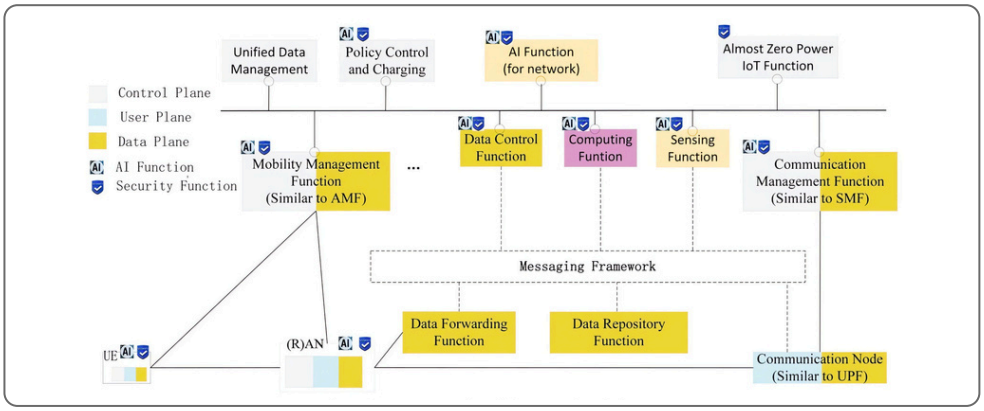


Figure 2.2 VIVO proposal for the 6G architecture [6]

2.2.3 Six Generation : Enablers and Technologies

The sixth generation (6G) will revolutionize connectivity by integrating advanced technologies that provide faster speeds, enhanced reliability, and smarter networks. It will utilize terahertz and millimeter waves, with ultra-massive MIMO, beamforming, and cell-free massive MIMO being the latest antenna technologies ensuring fault-tolerant performance. Additionally, 6G will incorporate reconfigurable intelligent surfaces, quantum communication, UAV and satellite communications, as well as AI/ML-based technologies [21].



Figure 2.3 significant technologies of 6G

2.2.3.1 Terahertz Communication

Terahertz frequencies boast incredible data rates due to their wide bandwidth. Terahertz communications can enable terabit-per-second (Tbps) communications, making them ideal for applications requiring bandwidth-intensive communications, such as high-definition video streaming or augmented reality. Terahertz Ultra Massive MIMO technology is capable of creating highly amplified and highly focused beams using advanced beamforming techniques. Among the many potential solutions, terahertz communications stands out as a highly effective technology for facilitating 6G and subsequent generations [5]. With its potential to enable Tbps transmissions for emerging applications, its importance is undeniable .

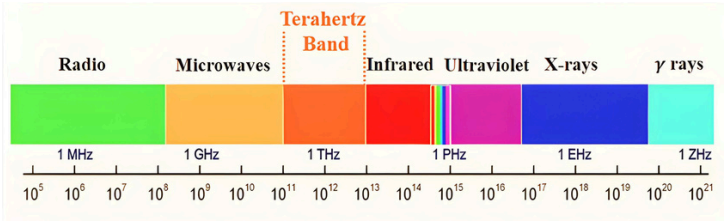


Figure 2.4 An overview of the Terahertz (THz) spectrum, ranging from 100 GHz to 3 THz, which is set to be utilized in 6G networks.[45]

Although it has many advantages that make it a revolution in the field of communications, it suffers from some problems, most notably weak penetration, especially for solid objects such as walls, and the limited number of available channels, which requires the deployment of huge numbers of small transmitters to ensure continuity of communication and coverage. These frequencies are still difficult to use and generate, and pose an engineering challenge due to the limited technologies needed to integrate them into networks and satellites, which are under development [5].

2.2.3.2 Ultra-Massive MIMO & Beamforming

We can define MIMO technology as a wireless communication technology that relies on the use of multiple antennas to send and receive data over multiple paths at the same time to increase user capacity and improve the quality of service. Its biggest goal is to minimize interference. It was developed in 5G networks to massiveMIMO to work better in crowded environments, where it uses 16 to 256 antennas working together to provide better coverage.

Despite the continuous improvement of this technology, the continuous increase in the number of devices connected at the same time, as well as the future 6G network's reliance on Hertzian frequencies, indicates the need for further advances in this technology, which brings us to ultra massiveMIMO.

- **ultra Massive MIMO:** An advanced technology that does not differ in principle from the previous one, aiming to use hundreds to thousands of antennas in base stations, which improves the network capacity significantly.

This advanced technology works by relying on **Smart beamforming** to direct signals with high accuracy to their destination to users. Where AI is used to intelligently direct the beams, which increases the number of devices that can be served on the same channel [5] [7].

It also takes advantage of the use of Hertzian frequencies, which provide high and wide bandwidth and at the same time mutual optimization between the two technologies, as the poor penetration of Hertzian frequencies is compensated by the huge number of antennas.

This technology provides high spectrum and low power consumption and is ideal for smart cities, self-driving cars and all applications that require constant real-time communication. In addition, it supports the presence of millions of simultaneous users.

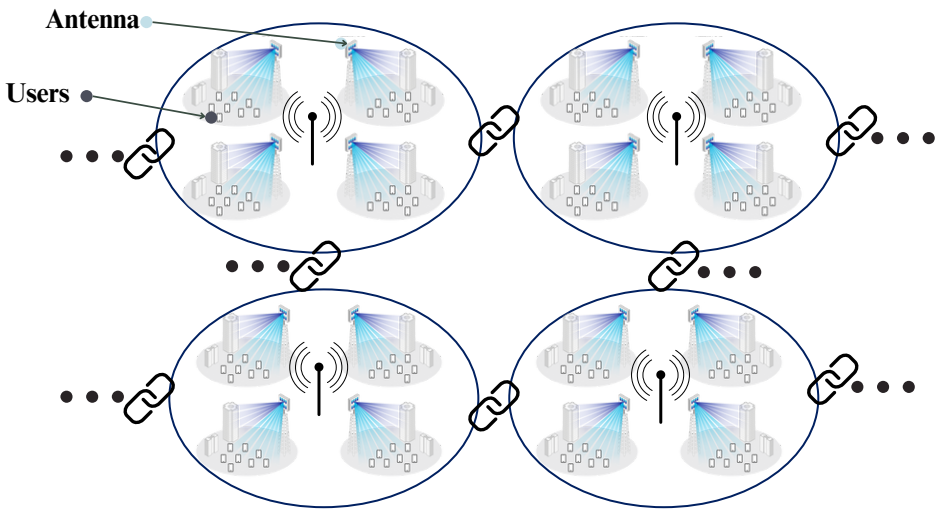


Figure 2.5 massiveMIMO and beamforming

2.2.3.3 Cell Free Massive MIMO

Cell free Massive MIMO is a wireless network architecture wherein a distributed array of a very large number of access points together serve users without the constraint of traditional cell boundaries. The access points, normally with a single antenna and mounted, are linked to a central processing unit CPU that organizes network operation [5] [7].

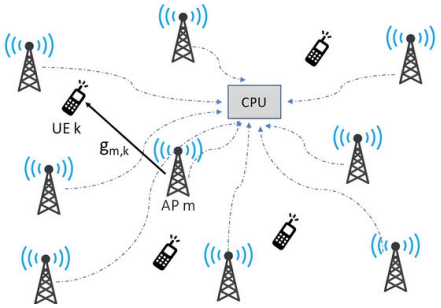


Figure 2.6 Cell massiveMIMO [18]

CPU for a cell free Massive MIMO system forms the core of the functionality, responsible for handling signal processing activities over the dispersed access points. It carries out channel estimation, downlink transmission signal precoding, and detection of user data for uplink transmission. Centralized operation makes it possible to allocate resources and coordinate signals over all access point improving network capacity, coverage, and energy efficiency. However , the central CPU also brings about computation and backhaul requirements, necessitating efficient data handling and coordination to provide the best of the system [5].

2.2.3.4 Millimetre Waves

Waves ranging from 30 to 300 GHz will be essential in 6G networks due to their availability of medium to good speeds up to 100Gbps and their wide bandwidth, making them promising for some future applications [22].

2.2.3.5 Quantum communication

Quantum communication is an advanced technology poised to play a critical role in the evolution of 6G networks. Unlike conventional communication methods that depend on electromagnetic waves for data transmission, quantum communication leverages photons, enabling inherently secure data exchange. This unique characteristic makes it especially suitable for highly sensitive applications such as military and government communications, where maximum security is essential.
Moreover, quantum communication has the potential to support emerging real-time applications, including autonomous vehicles and smart cities. Its ability to facilitate near-instantaneous communication across vast distances can significantly reduce latency and enhance responsiveness. Nevertheless, the implementation of quantum communication faces several obstacles, notably the requirement for specialized hardware and infrastructure, as well as substantial deployment costs.[5]

2.2.3.6 RIS (Reconfigurable Intelligent Surfaces)

A reconfigurable smart surface (RIS) is an emerging wireless communications technology that offers a programmable radio environment to enhance signal propagation. It consists of a planar array of low-cost passive reflective elements commonly referred to as metas-atoms .each capable of dynamically modifying the phase (and sometimes the amplitude) of incident electromagnetic waves, RIS is a smart electronic mirror whose reflection can be directed and whose angular and temporal response can be controlled [23].

The microcontroller integrated within RIS is responsible for adjusting the phases of the meta-atoms based on control commands issued by the network according to the state of the wireless channel. The user receives the reflected signals and transmits signal quality indicators (such as SNR and RSSI) to the base station over a direct uplink channel. Based on this feedback, the network adjusts RIS settings in an adaptive manner via microcontrollers to improve signal reflection and enhance performance in terms of coverage, power, and spectral efficiency

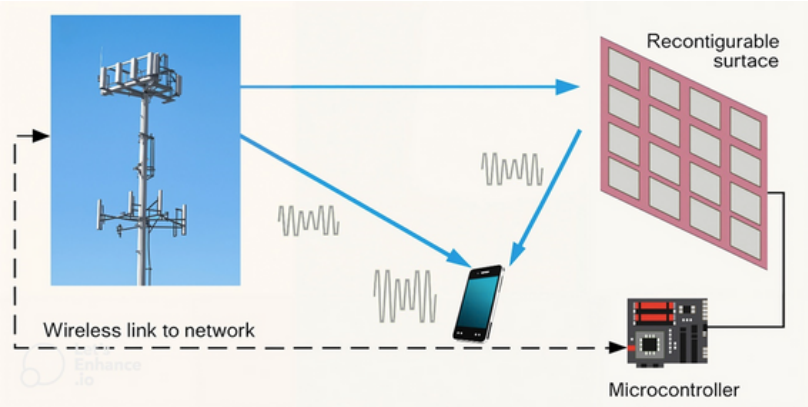


Figure 2.7 Significant technologies of 6G

2.2.3.7 UAV/satellite communication

UAV/satellite technology is the use of Unmanned Aerial Vehicles (UAVs) and satellites to offer wireless communication to remote and far-off places. It offers high-speed data communication and internet connectivity to remote or hard-to-reach locations that are inaccessible via conventional networks, thereby extending coverage to even more isolated areas. It is usable for applications such as real time disaster relief and remote surgery. It has some drawbacks such as extremely high expense of special hardware and problems with regulating usage of airspace [24].

2.2.3.8 AI & ML

AI is a simulation of human intelligence in a form that enables machines to perform tasks that were previously in human intelligence. ML is at the very essence of AI and a subfield that enables systems to learn by experience and get better over a period of time in tasks.

There are a variety of approaches that are a part of ML such as supervised learning where algorithms are trained on labeled data in a bid to make predictions and unsupervised learning that recognizes patterns in unlabeled data.

Additional paradigms such as reinforcement learning and semi supervised learning broaden the field of ML applications. These are elaborated on in later sections with a focus on implications and future possibilities [25].

2.3 Machine Learning Overview and Insights

AI is a computer science subdiscipline that is multi-disciplinary and deals with the study and creation of systems capable of executing tasks like human being pertaining to intelligence. It employs various methodologies and paradigms to make a machine perceive its environment, reason from information, learn from experience, and make good judgments. AI unites scientific theory and empirical science in many areas including ML, natural language processing, computer vision, robotics, and knowledge representation. AI research is not just to mimic human intellectual capability but to extend it, resulting in new solutions in health care, finance, transport, etc... . [8] [9] [25]

2.3.1 ML Functional Description

Its a branch of AI that focuses on developing programs to be able to learn from data without the need to program every computer task. In other words, instead of directly programming the computer to carry out tasks, it is provided with data to analyze and learn from it. This process aims to enable programs to make smart decisions and predict future events. This is done by ML models, each of which has a role, based on mathematical algorithms that give the models the advantage of recognizing patterns in the data.

2.3.2 ML Oriented techniques

ML works by training the models to reach the characteristic patterns of the given data,this is done in stages starting from analyzing and processing the data and understanding the relationship between the data to training in which
experiments are conducted to gain the ability of the model to adapt and improve the accuracy of the predictions .

ML is applied in various applications such as image recognition, medical diagnosis, optimizing network performance and many applications, It is also widely applied in the field of cybersecurity for fraud detection fraud.Building on our previous discussion now we focus on the some broad categories of machine learning [8] [9], we will discuss their distinct contribution and implications :

Unsupervised Learning is used on unlabeled data without a target variable, all variables used in the analysis are invested as inputs, the main goal of unsupervised science is to understand the basic structure of the data and extract useful information from it, without direct human intervention or, more precisely, without external supervision, and to make accurate decisions based on this information [25].

The process of extracting patterns from data is done by unsupervised algorithms that analyze and extract the rules that connect data elements [8] [9].

- Unsupervised learning enforces patterns on unlabeled data through algorithms such as clustering (e.g., K-Means, Hierarchical), association rule mining and dimensionality reduction (e.g., PCA...). Unsupervised learning is commonly applied in anomaly detection fraud detection, and analyzing big data.

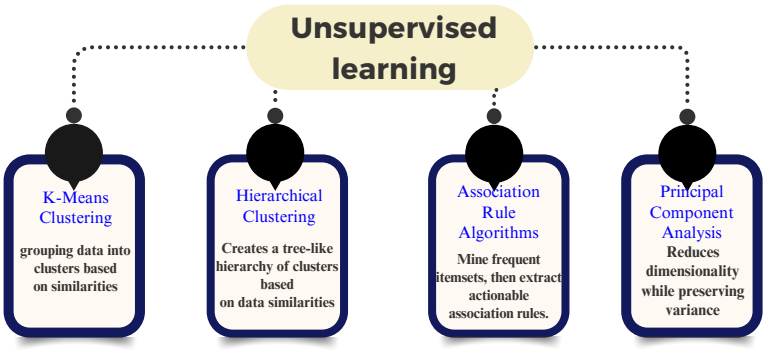


Figure 2.8 Unsupervised learning significant type

2.3.2.1 Supervised Techniques

In supervised the output and input attributes are predefined. Algorithms make predictions and classifications based on the predefined attributes, as well as the accuracy of their results. For example, if the input variable is X and the output variable is Y, the mapping function used in supervised learning would be Y = F(X). The learning process stops when the algorithm achieves an acceptable level of performance. Supervised algorithms perform the first analytical tasks using training data, then construct conditional functions to map the attribute variables. These algorithms require predefined maximum settings to achieve the desired outcome and performance [9] [25].

- **Linear Regression**

Linear regression is one of the widely applied supervised learning methods that makes use of modeling the linear connection between a dependent variable and an independent variable in order to forecast continuous responses

.In essence, linear regression will determine the ideal straight line predicting the observed values, either positive where increases in the independent value result in rises in the dependent value or a negative where increasing one results in decreasing the other. Not just can this be interpreted, it is also at the basis of statistical analysis [8] [9]. The mathematical description of the model is provided by Equation (2.1)

$$y = bx + c \qquad (2.1)$$

Where :

- y : The dependent variable or outcome that the model aims to predict.
- x : The independent variable or predictor used to explain variations in **y**.
- b : The slope, indicating the change in **y** for every one-unit change in **x**.
- c : The intercept, representing the value of **y** when **x** is zero.

- **Random Forest**

Random Forest is a powerful ensemble learning technique which optimizes the predictability through voting of very large numbers of decision trees. Random Forest is one of the extremely popular algorithms for classification as well as regression. In Random Forest, every tree is learned on a bootstrap sample of the training data and at each node of the tree, a random subset of features is used, which promotes tree diversity and avoids overfitting. The last prediction is obtained by taking an average of the prediction of all the trees; in regression cases, it would be simply a case of averaging out the output [8] [9]. Random Forest regression model formula is presented as in Equation (2.2):

$$\hat{Y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \qquad (2.2)$$

Where :

- $\hat{Y}$: The final predicted outcome of the ensemble.
- T: The total number of decision trees in the forest.
- $f_t(X)$: The prediction provided by the $f_i(X)$decision tree for the input **X**.
- X: The set of input features used for prediction.
- $\frac{1}{T}\sum_{t=1}^{T}$: The averaging process that aggregates the predictions from all trees to produce a robust final output.

- **Extremely Randomized Trees**

It is a machine learning model used for classification and regression, which is a modified and improved version of the random forest model, where it came as a solution to two main issues: overfitting, where it chooses partitions completely randomly for the ability to generalize well. It improves training speed by randomly selecting partitions without searching for optimality, which speeds up the handling of big data. $\frac{1}{T}\sum^{T}$ Extra tree model formula is presented as in Equation (2.3):

$$s = X_{\min} + r \cdot (X_{\max} - X_{\min}) \qquad \textbf{(2.3)}$$

Where :

- (s): The randomly chosen split point.
- $(X_{\min})$: The minimum value of the selected feature.
- $(X_{\max})$: The maximum value of the selected feature
- (r): A random number drawn from a uniform distribution in the range $[0, 1]$.

- **Gradient Boosting**

A robust ensemble learning method that constructs models in sequence, and each successive model is trained to correct for the mistakes made by previous models. It can be used in a broad array of classification and regression problems. The prediction is calculated in Gradient Boosting through an aggregation of the weighted predictions of all individual models, normally decision trees [8] [9]. The mathematical formula of a Gradient Boosting regression model is introduced in Equation (2.5):

$$\hat{Y} = \sum_{m=1}^{M} \eta \, f_m(X) \qquad \textbf{(2.5)}$$

Where :

- $\hat{Y}$: The final predicted outcome of the ensemble.
- M: The total number of boosting iterations or trees.
- $f_m(X)$: The prediction provided by the **m-th** model (or tree) for the input X.
- η : The learning rate that scales the contribution of each model.
- X : The set of input features used for prediction.
- $\sum_{m=1}^{M}$: The summation process that aggregates the weighted predictions from all models to produce a robust final output.

- **Extreme Gradient Boosting**

Gradient boosting algorithm that adds up the predictions from multiple decision trees and utilizes powerful regularization techniques to prevent overfitting. Models in XGBoost are sequentially trained, and each subsequent tree is trained to minimize the error of the prior trees [9]. The ultimate prediction is achieved by adding up all the contribution of the individual trees and can be represented in equation (2.6)

$$\hat{Y} = \sum_{m=1}^{M} f_m(X) \qquad \textbf{(2.6.1)}$$

Where :

- $\hat{Y}$: The final predicted outcome of the ensemble.
- M: The total number of boosting iterations or trees.
- $f_m(X)$: The prediction provided by the **m-th** for the input X.

Besides, XGBoost is optimizing a regularized objective to balance model accuracy and complexity. It is typically expressed as:

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(f_m) \qquad \textbf{(2.6.2)}$$

Where :

- $L(y_i, \hat{y}_i)$: is a differentiable loss function that measures the error between the actual and predicted values.
- $\Omega(f_m)$: is a regularization term that penalizes model complexity.

This method enables XGBoost to provide consistent and generalized predictions for different datasets and tasks.

- **Light Gradient Boosting Machine**

A high-speed gradient boosting library for big data processing in regards to fast training speed and reduced memory usage. is extensively used on regression as well as classification problems. LightGBM grows decision trees with leaf-wise tree growth policy, choosing the leaf with the highest loss to split on, resulting in more accurate models and improved performance. The final prediction is obtained by summing over all the trees; for regression tasks, this is accomplished by summing the output of each tree [8] [9]. The equation for the mathematical representation of a LightGBM regression model is presented as Equation (2.7):

$$\hat{Y} = \sum_{m=1}^{M} \eta f_m(X) + \arg\max_{\text{leaf}} G(\text{leaf}) \quad \textbf{(2.7)}$$

Where :

- $\arg\max_{\text{leaf}} G(\text{leaf})$: The process of selecting the leaf where the gain $G(\text{leaf})$ is maximized, improving the model's prediction.

The diagram illustrates supervised learning, dividing it into classification and regression models, with a focus on various regression types used to predict continuous outcomes from labeled data :

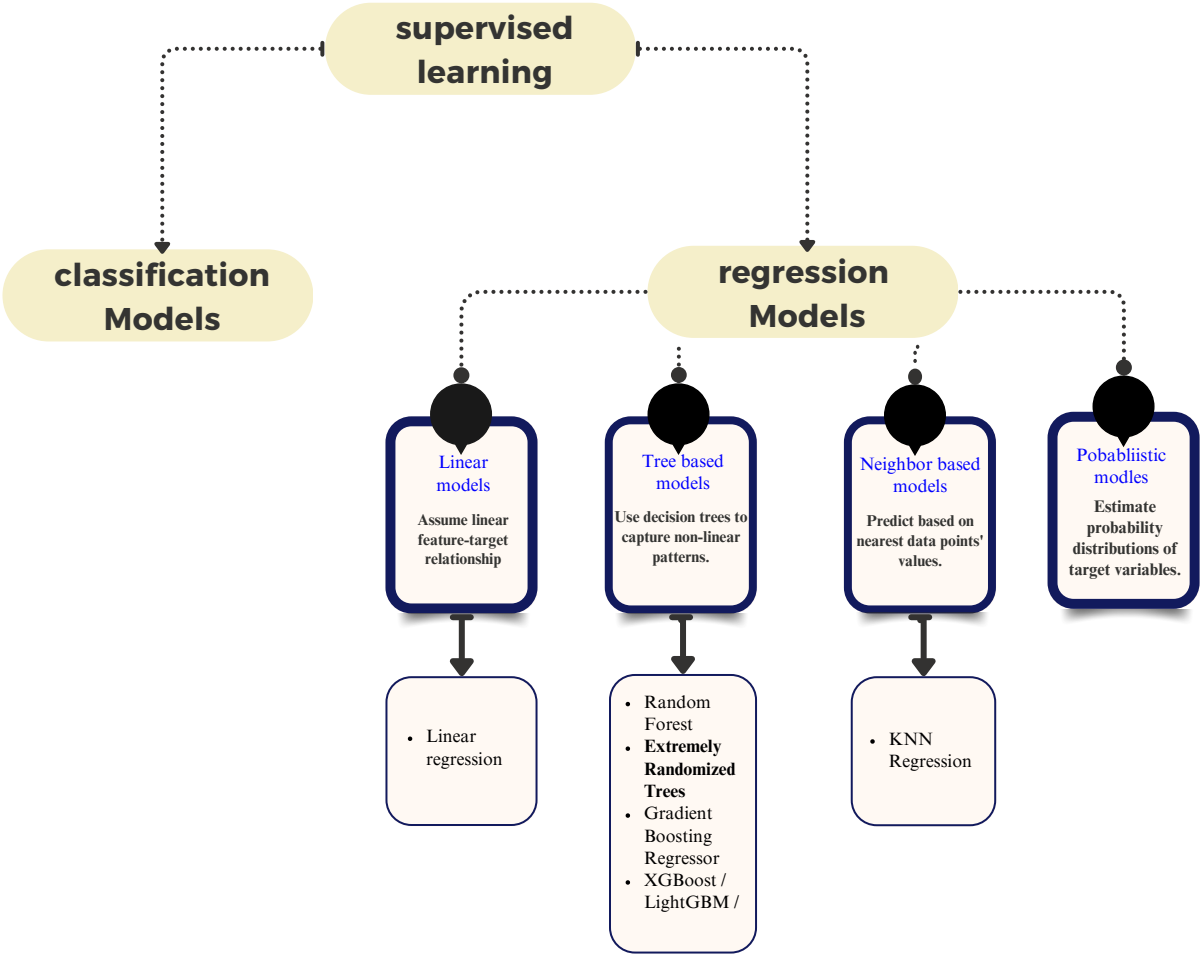


Figure 2.9 Types of Regression Models and Their Descriptions

2.3.2.2 Ensemble Techniques

Ensemble methods operate by aggregating multiple models usually called base learners to maximize overall prediction accuracy stability and robustness. Ensembles are not built from a single model but rather exploit diversity and complementary biases among various algorithms or instances of models. The notion is that single models make mistakes but that their collective decisions minimize variance with bagging bias with boosting or both with stacking .

There are three primary ensemble techniques:

1. **Bagging :** Trains many models on random samples of the data and takes average predictions. Example: Random Forest.
2. **Boosting:** Builds models sequentially each attempting to fit the mistakes of the previous model. Example: Gradient Boosting, XGBoost.
3. **Stacking:** Trains several models and takes their outputs as input for a superlevel "meta model".

- While ensemble methods enhance predictive performance, they may introduce challenges such as increased computational complexity, reduced interpretability, and a higher risk of overfitting if not carefully regularized.

2.3.2.3 Synthetic techniques

Users may create new material, including text, photographs, videos, sounds, code, 3D designs, and other media, by providing a range of instructions to Generative AI, or GenAI. It is taught and "learns" using existing online texts and artifacts [10].

As generative AI continues to learn from additional data, it develops. Large unlabeled data sets are used to train AI models and algorithms, which are created using intricate arithmetic and a lot of processing power. These datasets teach the AI to forecast results in a manner similar to what people might do or produce on their own.

A generative adversarial network (GAN) is a class of ML frameworks whose architecture distinguishes it from traditional deep neural networks. The network consists of two main components: a generator and a discriminator.

- **The generator** creates artificial data from random input (typically noise).
- **The discriminator** strives to differentiate between real data (sampled from the original dataset) and the generated data.

The two models are competitively trained , the generator tries to fool the discriminator by generating data that cannot be differentiated from actual data, and the discriminator tries to correctly distinguish fake samples. The adversarial process constantly enhances the quality of generated data, until the discriminator cannot tell it apart from actual data.

Despite the effectiveness in many applications, they suffer from several technical challenges in achieving stability and reliability. The most prominent of these challenges is the lack of bias during training, as a mismatch between the generator and discriminator can lead to inconsistent results or the failure of the model to achieve the current state. GANs also face the problem of mode collapse, which has a tendency to produce variance or lack of diversity, which weakens the model's ability to represent opinion polls. Furthermore, improving the quality of the generator is problematic, as there are no standardized criteria that allow for accurate and objective performance estimation. Therefore, GANs do not require high intensity regarding parameters and architecture, requiring careful tuning and extensive testing to achieve satisfactory results.

- Despite their growing capabilities, synthetic techniques such as Generative AI and GANs present notable concerns. These include the potential for generating misleading or biased content, dependence on high computational resources, and ethical challenges related to data authenticity, ownership, and misuse. Moreover, their reliance on large-scale datasets may amplify existing biases within the data, posing risks to

2.4 Six Generation and ML Synergy

Analysing the way in which ML can be used most effectively for novel 6G technologies, gaining insights into its revolutionary impact in the network landscape.

- **Autonomous Underwater Vehicles (AUVs)**

Underwater communication is affected by short range transmission and signal loss as a result of water flows, mobile aquatic lives, and turbulence. Autonomous Underwater Vehicles (AUV) help to improve the quality of communication in the Internet of Underwater Things (IoUwT) as mobile relays facilitating the transmission of signals. Sophisticated AI techniques, such as Deep Reinforcement Learning

and Asynchronous Multithreading Proximal Policy Optimization (AMPPO), are being targeted for ongoing research on AUV trajectory optimization and energy efficiency. Such methods have been seen to enhance performance in collision avoidance, energy efficiency and path planning and thus promote underwater communication technology [1] [27].

- **Mobile Edge Computing (MEC)**

ML is one of the principal facilitators of Mobile Edge Computing (MEC) optimization through enhanced task management, resource distribution, and power consumption at the network edge. Techniques such as multi-stack reinforcement learning (RL) enable base stations to maximize spectrum, power, and task allocation, minimize delays, and enhance efficiency. Federated learning FL frameworks augment the multiuser MEC systems to maximize task offloading and resource allocation with low energy consumption and data privacy preservation. DRL algorithms, such as DEETO, maximizes energy efficiency and workloads distribution over multiple edge servers. DDPG based RL also improves security as well as minimizes the cost in MEC networks by maximizing task offloading and resource utilization. These AIdriven advancements enhance MEC systems' effectiveness and responsiveness to meet the demands of continually changing and ever-increasing wireless landscapes [1] [28].

- **NOMA (Non-Orthogonal Multiple Acces)**

With the enormous expansion of telecommunications and the advent of 6G networks, spectrum efficiency has emerged as a greatest challenge. Non-Orthogonal Multiple Access (NOMA) is an effective way to address this challenge, allowing multiple stations to transmit data over the same frequency channel by forming groups with low interference and improved network performance. NOMA makes use of planned interference management rather than traditional frequency or time domain user separation, by allocating multiple transmit power levels to the diverse users in one channel. With this, all users can be decoded based on their respective power levels, optimizing the utilization of spectrum and capacity of the network. Second, NOMA can facilitate more devices and improve quality of connection and therefore is a keystone technology to achieve dense connectivity in future 6G networks. Combining NOMA with AI and Ml technologies, its efficiency could be further increased by optimizing the resource allocation and responding to varied network conditions [1] [29].

- **FSO (Free Space Optics)**

Wireless optical communications are a potential alternative to radio frequency spectrum, especially in 6G, since they have high bandwidth capacity. Optical free communication is a method of communication using light to carry data through the use of free space and not cables, and it is very flexible. Free space optical communication is affected by interference such as multipath fading and atmospheric turbulence that can degrade the quality of signals.

Free optical communication efficiency relies on accurate channel condition prediction since the stability and signal quality are determined by ambient conditions. Efficiency is improved with AI techniques and sophisticated algorithms applied to determine the conditions of a channel and transmission parameter adaptation to ambient conditions. Transmission efficiency and interference and atmospheric turbulence effects are enhanced with the implementation of multilevel technologies for application in multiple-communication systems [30].

- **VLC (Visible Light Communication)**

Visible Light Communication (VLC) technology operates by modulating the intensity of Light Emitting Diodes (LEDs) at high rates to generate digital signals (0s and 1s), which are received by optical sensors such as photodetectors or cameras and decoded into data. The optical signaling method offers high-speed data communication beyond the threshold of human perception, with standard LEDs serving as transmitters and photodetectors or image sensors serving as receivers. Received signals are processed to accurately reconstruct the information transmitted. AI plays a significant role in enhancing VLC by maximizing signal processing, reducing noise, and stabilizing the quality of connections. AI-driven systems control light distribution
dynamically in a way that guarantees signal stability under changing environments
while predicting ambient lighting changes and adjusting system performance in real-time. Furthermore, AI facilitates adaptive routing of signals in order to decrease interference from extraneous light sources for secure and dependable communication. It also keeps track of movement patterns of the user and improves efficiency of lighting as well as data transmission efficiency. Additionally, AI improves the security of VLC by implementing encryption and anomaly detection, identifying irregular patterns of the signal, and preventing possible cyber attacks for more reliable and trustworthy communication [4] [31].

- **Network Architectures and AI Algorithms**

Ml and AI enable 6G networks to be more intelligent and efficient by controlling the network, enhancing security, and minimizing latency. They also assist in intelligently managing resources and providing quality of service (QoS). Below is their function in the Control Plane, User Plane, and Data Plane [21]:

- Control Plane : is at the core of network function management and performance enhancement. With predictive analytics, it can predict traffic congestion and dynamically reallocate resources to ensure intermittent service. It also enhances security by identifying probable cyber attacks and initiating protective measures before any damage is done. It additionally enhances energy efficiency by proactively controlling network devices, turning them on and off based on requirements to enhance sustainability and resourcefulness [3] [6].
- User Plane : To improve the user experience, AI is used to analyze usage patterns and adjust the connection according to actual needs, such as improving the quality of service, allocating bandwidth based on the type of data sent, or prioritizing video calls and high-quality games. It also helps in predicting user movement, facilitating seamless transitions between different networks without interruption, especially in moving environments such as high speed trains or self driving cars [3] [6].
- Data Plane : In Data Plane, AI brings about efficiency and protection to data streaming by scanning the data for viruses, compressing and re-directing it to reduce resource consumption. It makes encryption mechanisms better by scanning vulnerabilities and re-designing the security measures for auto-updating with the newest updates to protect the data during transmission [3] [6].

Network Plane	Descriptions AI Function	ML Algorithms
Control Plane	AI manages network functions, anticipates traffic congestion, improves security, and optimizes energy efficiency.	Predictive models (e.g., Random Forest, Gradient Boosting), Anomaly detection (SVM)
User Plane	AI analyzes user behavior to allocate bandwidth, prioritizes critical data flows, and predicts user movement for seamless handovers.	Behavior analysis (K-Nearest Neighbors) Movement prediction (Time series forecasting models)
Data Plane	AI improves data streaming by identifying threats, compressing data, and enhancing encryption for better security and resource efficiency	Threat detection (Deep Neural Networks) Data compression (Autoencoders), Encryption (LSTM)

Table 2.2 Network Architectures and ML Algorithms.

2.5 Conclusion

Six generation technology unveils a paradigm breaking development in communication networks that ushers in major breakthroughs with unmatched innovations that will determine the future of connectivity. Innovation keynotes such as Terahertz communication, Ultra-Massive MIMO, Millimeter Waves, Reconfigurable Intelligent Surfaces, and Quantum Communication will provide the ultra high speed transmission of data, high capacity, and low latency to address the needs of next generation applications. Use of UAVs and satellite communications will also enhance network coverage and resilience to provide end-to-end global connectivity.

ML is among the most significant facilitators in 6G, with supervised learning tree-based models, neighbor based models, and probabilistic models taking the lead in optimizing and enhancing the 6G networks. These models will take the lead in enhancing communication modes such as Non-Orthogonal Multiple Access, Free Space Optical communication, and Visible Light Communication , along with network architecture and management.

With continued progress toward the deployment of 6G, the integration of next generation communication technology with AI based solutions will provide unprecedented performance, efficiency and responsiveness for introducing an all connected intelligent future.

Chapter 3
Methodology

3.1 Introduction

Data and artificial intelligence have developed to be key elements in various industries in the contemporary world because data has taken the forefront in decision-making. This implies that the process of collecting and processing it turns into an essential step in any machine learning project, thus the quality of outcomes to be achieved depends on data quality itself, thus data collection correctly and cleaning with utmost care are important steps to extract correct information.

The data collection process depends upon the determination of the right sources and their credibility as they are gathered from multiple sources such as open databases and various other sources. However, the raw data obtained is not always directly usable as it may contain missing values or duplicated values, requiring cleaning tasks such as cleaning out outlier values, managing missing values, or even extracting new features that empower the data strength, all for the purpose of producing analysis-ready and use-ready data, to minimize errors affecting the performance of models to be utilized moving forward.

Processed data can be the solution to most of the issues because it can give precise information to help in decision-making and process optimization, but the question is how one can best use the data to provide precise solutions to the problems to be addressed in the future.

3.2 Dataset collection and features description

3.2.1 Overview of Lumos5G

Lumos5G 1.0 is a data set for the IMC'20 paper "Lumos5G: Mapping and Predicting Commercial mmWave 5G Throughput". It captures the Loop area of the Minneapolis city center, close to U.S. Bank Stadium. It samples a 1300-meter loop of roads, railroad crossings, and other city infrastructure. Lumos5G offers data measurements of data gathered from commercial mmWave 5G networks, and thus it is a valuable tool to analyze and forecast mmWave 5G throughput in real World settings. It is being released to advance research in 5G technology, network performance and machine learning based throughput prediction, and help for next generation technology.

3.2.2 Purpose of Lumos5G Dataset

The main objective of Lumos5G [32] is to enable researchers and developers to get insights into mmwave 5G performance across different mobility conditions, e.g., walking or driving. Measurements of network signal strength, patterns of UE mobility, and measurements of throughput performance are included in the dataset. Researchers can explore throughput prediction models, examine trends in signal strength, or determine the impact of mobility on the network performance through Lumos5G.

3.2.3 Data Collection Methodology & Source

The measurements were made with commercial UE equipment by a series of controlled experiments conducted in drive and walk modes in the Minneapolis Loop region. The experiments included measurement of important network parameters like LTE/5G signal strength, throughput, and mobility factors every one second in both clockwise (CW) and counter-clockwise (CCW) directions. The information source is real mmWave 5G network measurements in central Minneapolis, as extracted via Android APIs to connectivity, mobility, and signal quality. For additional information the dataset is publicly accessible through the IEEE DataPort: Lumos5G Dataset.

3.2.4 Feature Definitions and Description

In Lumos5G 1.0 we have 18 features The following are the critical features of the dataset and their definitions :

- **run_num :** Represents the run number or test session identifier. Each run could correspond to a specific test or measurement session.
- **seq_num :** Represents the sequence number within a run. It indicates the order of measurements taken during a test session.
- **abstract SignalStr :** Likely represents the abstract signal strength or a normalized value of the signal strength.
- **latitude and longitude :** Represen the geographical coordinates (latitude and longitude) where the measurements were taken. This indicates the location of the device during the test.
- **movingSpeed :** Represents the speed of the device (e.g., a mobile phone or vehicle) during the test, likely in meters per second (m/s) or kilometers per hour (km/h).
- **compassDirection :** Represents the direction of movement of the device, measured in degrees (0° to 360°).

- **nrStatus :** Likely represents the status of the 5G NR (New Radio) connection. For example, it could indicate whether the device is connected to a 5G network or not.
- **lte_rssi :** Represents the Received Signal Strength Indicator (RSSI) for the LTE (4G) network. It measures the power level of the received signal.
- **lte_rsrp :** Represents the Reference Signal Received Power (RSRP) for the LTE network. It is a measure of the signal strength of the LTE network.
- **lte_rsrq :** Represents the Reference Signal Received Quality (RSRQ) for the LTE network. It measures the quality of the received signal.
- **lte_rssnr :** Represents the Reference Signal Signal-to-Noise Ratio (RSSNR) for the LTE network. It indicates the quality of the signal relative to noise.
- **nr_ssRsrp :** Represents the Synchronization Signal Reference Signal Received Power (SS-RSRP) for the 5G NR network. It measures the signal strength of the 5G network.
- **nr_ssRsrq :** Represents the Synchronization Signal Reference Signal Received Quality (SS-RSRQ) for the 5G NR network. It measures the quality of the 5G signal.
- **nr_ssSinr :** Represents the Synchronization Signal Signal-to-Interference-plus-Noise Ratio (SS-SINR) for the 5G NR network. It indicates the quality of the 5G signal relative to interference and noise.
- **Throughput :** Represents the data throughput of the network, likely measured in megabits per second (Mbps) or kilobits per second (Kbps). It indicates how much data is being transferred over the network.
- **mobility_mode :** Represents the mobility mode of the device during the test. For example, it could indicate whether the device is stationary, walking, or driving.
- **trajectory_direction :** Represents the direction of the trajectory of the device. For example, it could indicate whether the device is moving clockwise (CW) or counterclockwise (CCW).
- **tower_id :** Represents the ID of the cellular tower to which the device is connected. It identifies the specific tower providing the network connection.

The following table **Tab 3.1** shows a summary of the Lumos5G dataset, with the essential information like the study area, experiment configuration, data modalities, and data form. It focuses on key features acquired when measured, for example, signal quality, mobility, and network performance measures. hat the dataset presumably reflectsmore details :

Category	Descriptions
Area of Study	Loop area near U.S. Bank Stadium, Minneapolis, USA
Experiments	Walking and Driving Modes
Number of Experimental Runs	Multiple runs in Clockwise (CW) and Anti-Clockwise (ACW) directions
Modalities	Commercial UE (User Equipment) devices; Android APIs for signal and mobility data
Dimensions	Latitude, Longitude, Signal Strength, Throughput, Mobility Speed, Compass Direction, etc.
Format	CSV

Table 3.1 Lumos5G dataset overview in a tabular format

3.3 Data Preprocessing

Critical process of scrubbing raw data to render it reliable and functional for machine learning algorithms. The critical process of scrubbing raw data to render it reliable and functional for machine learning algorithms involves systematically looking for and fixing mistakes, dealing with missing or conflicting values and reorganizing the data into algorithmic shape. Critical operations include eliminating duplicates, removing outliers,imputation or interpolation to fill in missing values, encoding categorical variables into machine readable formats and scaling or normalizing numerical features. Each of these preprocessing operations improves the quality of the dataset, making it suitable for effective training and quality prediction in machine learning modelsmaking it suitable for effective training and quality prediction in machine learning models.

3.3.1 Removing Unnecessary Columns from the Dataset

`run_num`, **`seq_num`**, and **`tower_id`** can be dropped for more helpful data analysis. The `run_num` itself, as a test session ID, could be of not much use if analysis is just being done across overall trends rather than by single test sessions. Also, `seq_num` to reflect measurement order is not required as data can be naturally ordered using other time-related or space-related fields such as time, latitude, and longitude. `tower_id`, showing which tower a device is on, is unnecessary unless tower-level activity is under explicit investigation. In most scenarios, it is uninformative and can make

the problem more complex. The dropped columns simplify the dataset by removing redundant or irrelevant features, making focused and less comprehensive analysis possible.

3.3.2 Handling NaN (Not a Number) values

refers to the process of handling missing or undefined values within a dataset to get it clean, complete, and ready to analyze. It is a compulsive process within data preprocessing because NaN values will absolutely destroy calculation, analysis, and training models. There are many options to treat NaN values, such as ignoring rows or columns containing NaNs, imputation of missing data using the mean, KNN, or model to fill , interpolation, . Each method is based on data type and goal of analysis.The above **fig 3.1** shows the proportion of NaN values in each colum :

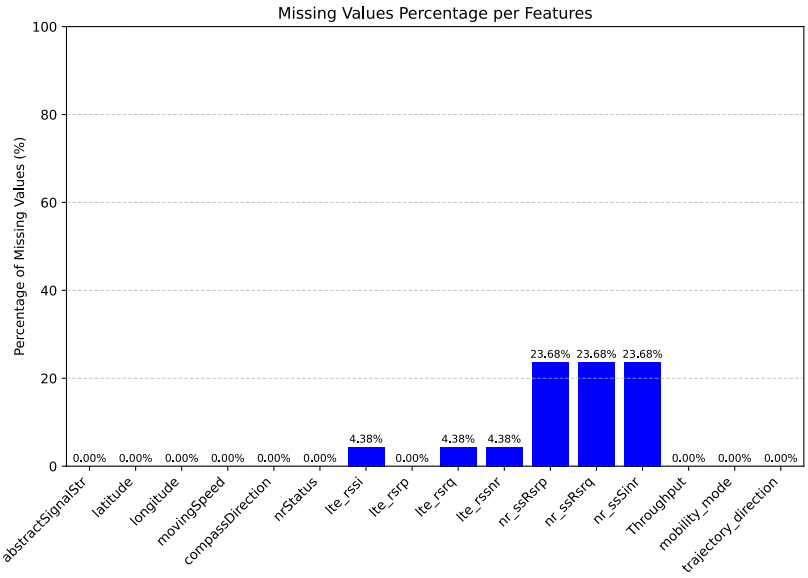


Fig 3.1 Percentage of NaN Values in Each Features before dealing with nun

- **Delete Colums wtih (4.38%) :** LTE signal quality measurements (lte_rssi, lte_rsrq, lte_rssnr) have quite a small number of missing values (4.38%) and are not very significant with the overall analysis since 5G performance is mainly the concern. Since the dataset already includes detailed 5G readings like nrStatus, nr_ssRsrp, nr_ssRsrq, and nr_ssSinr, the LTE columns are of little extra value. Keeping or adding these fields would add unnecessary complexity with minimal analytical value, and their removal is a logical step towards a leaner more streamlined dataset.

- **Use Linear Regression Imputation for (23.68%) :** As these columns are crucial in the assessment of 5G network performance, maintenance of these columns will enable us to validate the integrity of the data set, particularly for the use cases such as estimating the throughputs or quality of the network. Linear regression imputation enables us to replace gaps with a trend using other variables without adding much bias [33]. It works well on columns where systematically missing data are, or are co-related with, the other variables.

After this preprocessing stage, the target columns are now prepared for use in subsequent analysis steps. Missing values are not filled in randomly, but rather using a method that considers statistical relationships between variables, helping to preserve the overall structure and integrity of the dataset. This reduces the potential for introducing bias into the model output and improves the reliability of predictions, especially in critical use cases such as productivity estimation. As a result, these columns can now be considered reliable for further analysis and processing.

3.3.3 Dealing with categorical features

We found that there are 3 features in our data (**mobility_mode**, **nrStatus**, **trajectory_direction**) with a categorical pattern, which does not serve us in our study in this format and cannot be relied upon because most machine learning algorithms need numeric inputs to perform the required calculations.

Therefore, we rely on the unary coding method [34], which creates separate binary columns for each category in the selected feature and then assigns the values either 1 or 0.

This is an example of how it works :

$$mobility_mode_walking = \begin{cases} 1 & \text{if} - mobility_mode = \text{walking} \\ 0 & \text{otherwise} \end{cases}$$

$$mobility_mode_driving = \begin{cases} 1 & \text{if} - mobility_mode = \text{driving} \\ 0 & \text{otherwise} \end{cases}$$

The formulas represent binary variables for mobility modes. If the mobility-mode is "driving," the variable mobility-mode-driving is set to 1 and 0 otherwise. Similarly, for "walking," mobility-mode-walking is set to 1 and 0 otherwise.

3.3.4 Normalization of values

After all the adjustments and cleaning of our data from removing missing values and filling in others, as well as dealing with real-time data, the data is ready for the last step, which is to apply the data normalization technique, which means converting all values between 0 and 1. This is an important step to improve the performance of the models used in the future, as this conversion ensures that all features are treated with the same scale.

This is achieved by the min-max scaler technique [35], and here we mention its formula that converts each value as mentioned earlier:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \qquad \textbf{(3.3.1)}$$

- After the data cleaning and preprocessing stage, the Lumos5G dataset was sanitized to contain 19 key features without count the target. These are significant network and mobility features that are required for effective analysis of 5G performance. The plot presented below **fig 3.2** is a visual representation of the final feature set maintained for analysis, for a cleaner and specific dataset :

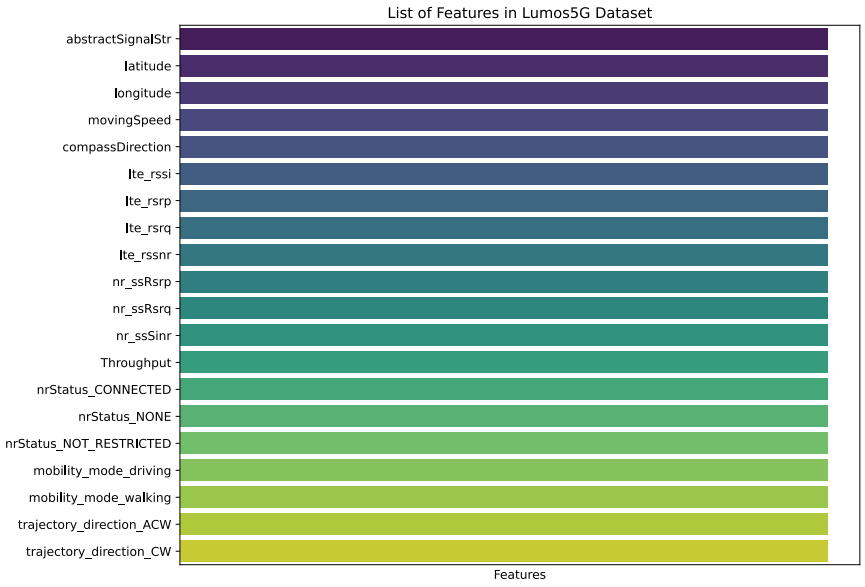


Fig 3.2 Visual Representation of Key Features in the Cleaned Lumos5G Dataset

3.4 Experimentation settings And Tools

This section provides a complete picture of the development environment and programming languages employed during the development and implementation of our system. It outlines the tools, frameworks, libraries and technologies implemented throughout the development process so that it can be understood clearly regarding the technical basis on which the system functionality is based. By making the underlying software and programming resources obvious, this section refers to the systematic method followed in order to attain efficiency, scalability and reliability in developing the system.

3.4.1 Programming Tools

Python was chosen as the main programming language in this project because of its flexibility and vast library environment, especially for data science and machine learning. The capability of Python to handle data, plot data and perform machine learning tasks effectively made it a good choice for our system development. **jupyter Notebook** was used as the development tool to facilitate an interactive and iterative programming process. It gave an efficient platform to write, test and document code in an integrated way.

3.4.2 Main Libraries Used in the Project

A number of necessary Python libraries were utilized in the creation of this system which offered robust functionalities for data manipulation and machine learning :

- **Pandas :** One of the highly popular libraries used to handle and process data is Pandas. Pandas was utilized to successfully handle datasets clean and the data structure for analysis.
- **NumPy :** This library also had numerical computation support, such as array and matrix operations, that form the basis of most machine learning algorithms.
- **Scikit-learn :** Being an extensive machine learning library, Scikit-learn provided mechanisms for model training, evaluation, and validation, and hence was an integral part of the machine learning process.
- **Matplotlib and Seaborn :** These libraries are used for the purposes of data visualization, using which one can generate complex plots and charts and graphs in order to represent the results of the analysis.

3.4.3 Hardware

For the process of installation and constructing the system, two independent machines were used:

1. **HP 14s:** Windows 11, With AMD Ryzen 3 3250U 64Bits, 256GB Ssd and Radeon Graphics 2GB and 8GB RAM, the device is improved for tasks where decent graphic performance matters (such as light visualization work) and provides decent performance for basic machine learning work.
2. **Dell 8hdolb9**: Windows 10, This system features the Intel i5-7300U 64 Bits 256 Ssd and 16GB of RAM, so it will have additional memory and be better able to work with additional data sets and more sophisticated machine learning models.

3.5 Identifying the Problem And Labelling

3.5.1 Understand the Problem

As we explore the secrets of 6G networks, we are irresistibly pulled toward one question above all: What really fuels the data flow within this next-gen world Behind this inquiry lies throughput, a metric that quietly determines the pace of connectivity seamlessly linking people, devices, and infrastructure. Underneath layers of signal power, mobility trends, and location fluidity, throughput is the key to the hidden potential of the network. Best to predict it is not only a technical hurdle; it's an exploration of the forces shaping the unseen framework of our next digital landscape.

3.5.2 The importance of Throughput prediction

Throughput is a key metric for evaluating the performance of networks, as it expresses the actual data transfer rate over the network during a certain period of time, it is affected by several basic factors in communications such as bandwidth, delay, and network congestion. Especially in our study on 6G networks, which rely on advanced technologies such as Terahertz scanners and artificial intelligence, so maintaining high Throughput is important to ensure an optimal user experience.

From the data at hand, achieving the desired level of Throughput requires an accurate understanding of the factors that affect it. Depending on the features we have, such as user traffic speed, signal strength, and even the state of the network. Therefore, after analysis and study

we found that the prediction of Throughput in the network is something that must be worked on and improved due to several reasons, including its direct impact on the features that we have

- In order to better understand the influence of each input variable on the predicted throughput, we conducted a thorough feature importance analysis. This process allowed us to quantify the contribution of each feature to the model's performance. To illustrate these findings, we generated a plot showcasing the four most impactful features: nr_ssRsrp, nr_ssSinr, Status_CONNECTED, and is_NOT_RESTRICTED. These features showed a clear correlation with throughput, indicating that variations in signal strength, signal-to-interference-plus-noise ratio, connection status, and network access restrictions play a critical role in shaping the expected network performance.

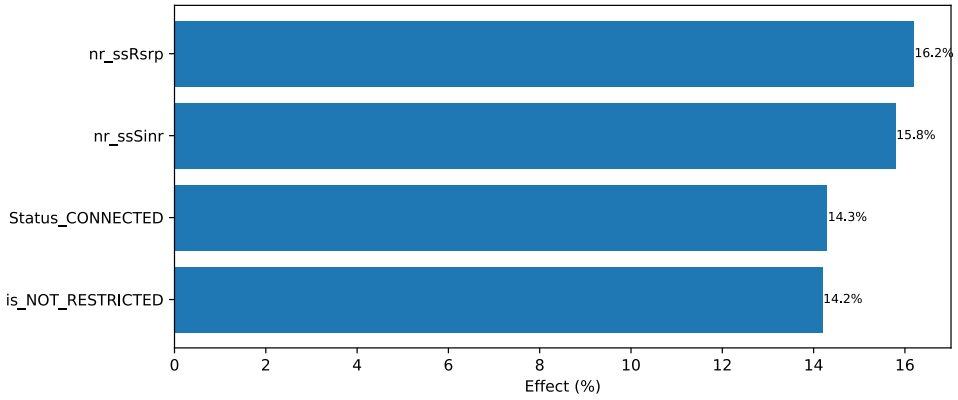


Fig 3.3 Relative Impact (%) of the Four Most Significant Features on Throughput Estimation

- In this analysis, we focus on the relationships between various network and device parameters and their impact on throughput. The process involves creating visual representations, supported by detailed analyses, to demonstrate how each feature in the dataset influences the predicted throughput. By relying on recent studies, we aim to clarify the intricate dynamics between these factors, enabling more accurate predictions. These insights are crucial for optimizing network performance and user experience in the rapidly evolving 5G and next-generation (6G) communication environments.

Now we utilize comprehensive examples figuer to describe how each of the features we explored affects the data throughput expected features. Through the use of visual graphs and explanatory analysis, we are able to clearly establish how these variables combine to give their effect on the performance of the network at large. The examples find useful application in streamlining network infrastructure :

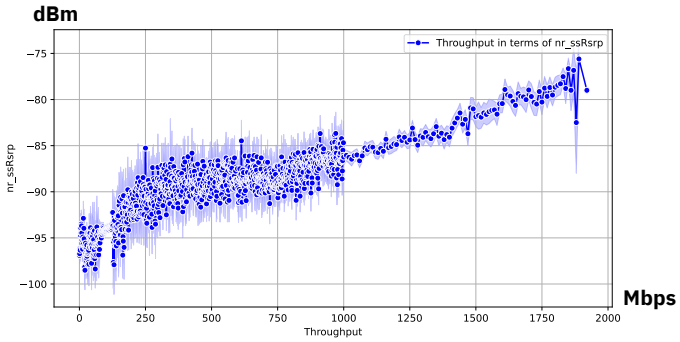


Fig 3.4 nr-ssRsrp (dBm) effect on Throughput (Mbps)

- As the received signal strength approaches -75 dBm, the signal strength increases, resulting in higher Throughput, while a weak signal reduces the data transfer speed due to poor connectivity.

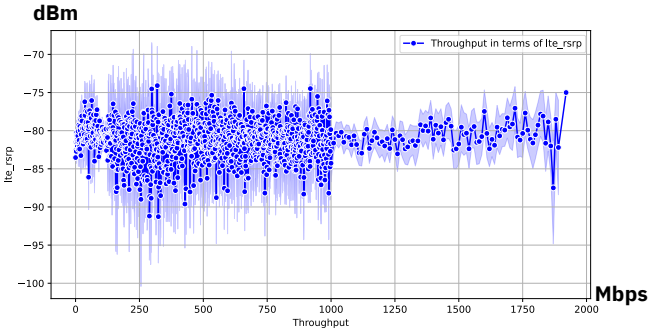


Fig 3.5 lte-rsrp (dBm) effect on Throughput (Mbps)

- The closer the received signal strength of the lte network is to -70 dBm (stronger signal), the more likely it is to have a higher Throughput. For low values of Throughput, lte_rsrp does not seem to be the only influencing factor, and there may be other influences such as user traffic speed

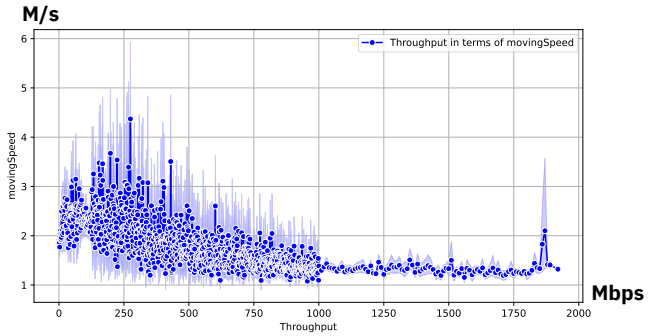


Fig 3.6 Moving speed (m/s) effect on Throughput (Mbps)

- Increased speed affects throughput by causing latency and data loss. Rapid movement leads to unstable connections. Thats why the high speed of the users negatively affects Throughput.

3.5.3 Model Selection Strategy for Throughput Prediction

Selecting the appropriate learning algorithms is one of the most important steps in developing accurate and reliable predictive models, particularly in estimating throughput of 5G and future 6G wireless networks. Given the complex, non-linear, and multi-dimensionality of input features spanning from spatial position to mobility characteristics and radio signal features, the present research in this paper focuses on ensemble-based supervised regression models that are stable, scalable, and efficient at detecting intricate data patterns.

The models considered in this research are Random Forest Regressor (RF), Gradient Boosting Regressor (GBR), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). These algorithms have been selected due to their effectiveness in previous studies for structured tabular data, capability to handle feature interactions, and ability to combat overfitting by using regularization techniques. They also support flexible hyperparameter tuning, which is essential for performance optimization across different validation methods and dataset structures.

The model selection process was not only guided by theoretical suitability, but also by empirical verification on various feature values and test scenarios. All models were assessed using standardized regression parameters (R^2, MAE, RMSE) [36], and the results were examined to detect their predictive reliability, error distribution, and applicability in actual time or intelligent network conditions.

3.6 Conclusion

In conclusion The Lumos5G project tries to forecast throughput using a deep network-based feature dataset. By data preprocessing and cleaning we determine throughput determinants like signal strength, user mobility, and site location. This helps us to optimize network efficiency and user experience. The project is an essential step towards making next-generation technologies accessible for 5G and future networks, facilitating infrastructure development and better resource utilization.

Further, the project is valuable in increasing the network's adaptability to diverse environments and growing demands for higher internet speeds and uninterrupted connectivity By critically analysing the complicated interaction between different features and their effects on network performance, the project provides a profound understanding of how the different factors, such as mobility and signal variations, affect the performance of the network, thus making it possible to better forecast impending issues.

Besides optimizing existing performance, the project lays the groundwork for 6G communications futures to emerge and accelerates the digital revolution in most industries

In brief, Lumos5G is a step in the right direction to improve efficiency in data transfer by predicting throughput. The insight gained from network parameter studies assists in improving network performance and in helping users to maximize utilization of next-generation communication networks.

Chapter 4
Results and Discussions

4.1 Introduction

This chapter initiates the transition from theoretical to real-world applications of ML techniques for future wireless communication systems. The emphasis lies in the exploration of how ML models can be employed to predict network throughput based on signal quality, mobility, and geolocation, a significant function for intelligent and adaptive 6G network communication.

Here in this work, the deployment is carried out based on an actual-world dataset collected from industrial 5G mmwave deployments. The data can be collected from 5G environments but is a good foundation for 6G related work since the architecture, frequency utilization (like mmwave and sub-THz), ultra-dense deployment, and mobility issues of the users are analogous in both environments. By analyzing the AI model behavior in response to high frequency signal fluctuations, mobility dynamics, and network states, we gain insights that are immediately transferable to 6G development, particularly in areas like resource optimization, predictive beamforming, and adaptive QoS control.

We use a combination of supervised learning algorithms like LightGBM, Random Forest, Gradiant Boost, ExtraTree, and XGBoost and each of them trained and tested over a pre-processed dataset consisting of signal strength indicators (e.g., RSRP, SINR), user speed, compass direction, and network status. The aim is to try out the models' ability to predict throughput precisely and generalize across actual network behavior.

To further support modeling, Generative AI techniques are employed to simulate fake data and augment the training set, offering a means of improving robustness, mitigating data imbalance, and simulating fictional cases that simulate 6G scenarios of the future. It also addresses another fundamental research challenge: the unavailability of large-scale, labeled 6G datasets.

Lastly, in this chapter it is demonstrated that although the employed data is based on 5G systems, the created models and methods are a stepping stone to 6G innovation that enables proactive network optimization and intelligent resource management key aspects to the autonomous and smart 6G era.

4.2 Performance Metrics in ML

In ML, model training and development are only half the work. It is necessary that they also be tested for reliability and performance through objective and rigorous evaluation. This is done through properly set quantitative measures that reflect the performance of the model in making correct predictions.

Therefore, measurement metrics are typically categorized into two broad categories :

4.2.1 Classification Tasks

In classification problems, the model assigns discrete labels to input data, and the model is typically quantified in terms of four such major metrics: **Accuracy**, which is the ratio of total correct predictions; **Precision**, indicating how many positive case predictions were indeed true; Recall, which indicates the capability of the model to identify all the actual positive cases; and **F1 Score** harmonic mean of **recall** and **precision**, which gives an equally weighted estimate, especially for datasets where classes are imbalanced.

4.2.1 Regression Tasks

In regression problems, one seeks to forecast continuous numerical values rather than discrete labels. Unlike classification problems, the word "accuracy" cannot be used directly in regression. Rather, statistical measures of error are used to ascertain how well the predicted values from the model correspond to the observed values. The most widely used regression evaluation metrics are :

- **Coefficient of Determination (R^2 Score) :** represents the ratio of explained variance to the overall variance of the dependent variable and may be computed from the coefficients [37]. It measures the fit of the model to the output variation.

  The Mathematical Formula of R^2 is :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (4.2.1)$$

  Where :

  - y_i : Actual value for observation
  - $\hat{y}_i$: Predicted value for observation
  - $\bar{y}$: Mean of actual values

R^2 ranges from 0 to 1. A higher R^2 indicates better model performance. For example, R^2 = 0.75 means that 75% of the variation in the target variable is explained by the model.

- **Mean Absolute Error (MAE)** : MAE calculates the mean absolute difference between the predicted and actual values [36]. MAE gives a straightforward interpretation of average prediction error in the same unit as the target variable. The Mathematical Formula of MAE is :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \textbf{(4.2.2)}$$

MAE penalizes all errors equally and is less sensitive to outliers. A lower MAE value indicates a model with more accurate and consistent predictions

- **Root Mean Squared Error (RMSE) :** measures the square root of average of squared differences between actual and predicted values [36]. It gives more importance to larger errors and is applied when large deviations are not preferred. The Mathematical Formula of RMSE is :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \textbf{(4.2.3)}$$

RMSE increases as large errors become more prominent, making it a good metric for identifying models that are not just generally inaccurate but also unstable in predictions.

In the provided table **Tab 4.1** we can see the evaluation metrics rang :

evaluation metrics / Predictive Quality Tier	R^2	MAE	RMSE
High Predictive Reliability	≥ 0.90	≤ 0.03	≤ 0.05
Statistically Acceptable Performance	0.75 – 0.89	0.04 – 0.06	0.06 – 0.09
Marginal or Suboptimal Predictive Utility	< 0.75	≥ 0.10	≥ 0.10

Tab 4.1 Benchmarking Regression Performance Using R^2, MAE, and RMSE Thresholds

- We took these values based on previous experiences.

4.3 Model Training & Hyperparameter Optimization Strategy

In pursuit of powerful throughput prediction in future wireless systems, this research aims to applying and evaluation of a group of tree-based ensemble learning models that include: Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), Gradient Boosting Regressor (GBR) and Random Forest Regressor (RF). These models were chosen due to their efficacy in modeling intricate non-linear relationships, scalability to high-dimensional feature spaces, and flexibility to tabular data characteristics of special concern to dynamic and heterogeneous wireless environments.

To ensure methodological strength and prevent the risks of overfitting, the entire dataset of actual real-world 5G throughput measurements, augmented with synthetic samples generated via generative modeling techniques was split into training (80%) and test (20%) subsets. The training set was used for model fitting and hyperparameter adjustment, while the test set served as an unseen evaluation benchmark to assess the models' generalizability.

Hyperparameter tuning is crucial in supervised learning to optimize non-learnable parameters that influence model training. Suboptimal tuning can impair model generalization, whereas optimal values can enhance performance. In this study, Optuna was utilized for hyperparameter tuning, taking advantage of its efficient optimization algorithms to search and determine the best-performing hyperparameters. Optuna supports the utilization of sophisticated search algorithms such as Tree-structured Parzen Estimator (TPE) [38], which makes the search process efficient and effective for high-dimensional problems.

This systematic assessment approach allows for robust and replicable comparison of ensemble methods under controlled experimentation. The results thus achieved facilitate an improved comprehension of model behavior in AI-based wireless communication systems and offer useful recommendations for throughput-aware optimization in future 6G network designs.

This table below provide our models Hyperparameter Setting :

Important Note

Part of this work has been published in a scientific article. Therefore, certain sections have been intentionally omitted from this version in accordance with academic publishing and copyright policies.
For full access to the original content, you may contact the academic supervisor at:

Chaima Aouiche
chaima.aouiche@univ-tebessa.dz

4.8 Conclusion

This chapter is about the performance of five machine learning algorithms Random Forest, LightGBM, XTrees, Gradient Boosting and XGBoost under three stages. In the first stage models were trained and tested with the original dataset. Even though there was class imbalance and a small sample size models such as XGBoost and LightGBM performed comparatively well, whereas others were more prone to these problems.

The second phase consisted of Synthetic data to improve the training set further and hence its diversity and size. The data augmentation generated appreciable improvement in model accuracy, stability, and generalization capability, especially for techniques adopting sequential learning techniques i.e., Gradient Boosting and XGBoost.

And with another new phase where ensemble learning methods were developed followed. Despite using the original dataset only, ensemble methods performed impressively well, beating single models many times. Among them stacking ensemble based on Gradient Boosting (Stacking_GBM) performed best according to all the measures of evaluation, reflecting its ability to recognize deep patterns in the data. Other ensemble models based on Lasso, Ridge and ElasticNet regressors performed competitively as well, confirming the application of ensemble techniques in predictive modeling.The XGBoost model by itself maintained its competitive performance above most single learners and in comparison with data imbalanced robustness. However, the voting ensemble that used the ensemble method for a combination of predictions between XGBoost, Random Forest, Gradient Boosting, and Extra Trees offered acceptable performance but not accuracy that stacking methods attained, indicating imperfections in static aggregation methods.Classic individual model like XTrees, Gradient Boosting and Random Forest were relatively poorer and more noise sensitive. However these are possible in scenarios where interpretability or computational cost is important.

Lastly, the most significant enhancements in performance occurred when models were trained from synthetic data followed by ensemble learning techniques on the raw data. While ensemble models did not improve with Synthetic data they still exhibited equal or even superior performance relative to lone models. This suggests that utilizing synthetic data generation and ensemble technique provides a robust and efficient method for predictive modeling, especially in systems where there is either scarcity or imbalance of data

Chapter 5
Conclusion and Future Direction

5.1 Challenges and Limitations

Despite the unprecedented growth that the alignment of AI with 6G networks has promised, the integration has insurmountable challenges that can influence its efficiency and sustainability, especially in the future 7G networks. Some of the most paramount of these challenges include the need for colossal computing power to deal with the humongous volumes of data generated by user devices, smart vehicles, and the Internet of Things. This can lead to wasteful energy consumption and latency if not smartly managed. The widespread use of personal data also poses increasing issues on privacy and security, especially in a 6G system based on edge and pervasive computing. Limited explainability of decision-making of AI systems is a major hindrance, as they are difficult to explain how they arrive at decisions in real-time, and it is alarming that they are being applied in key domains like healthcare and transportation. Lack of worldwide unified standards for AI and 6G integration makes it further difficult to make the networks and infrastructure interoperable. As 7G approaches, these issues are likely to intensify, demanding more efficient and transparent AI, more nimble and secure infrastructure, and international cooperation to build unmistakable regulatory frameworks that pave the way for wiser and greener networks.

5.2 Future Directions and 7G

7G networks, though not yet standardized but still in theory, are envisioned as a colossal breakthrough from 6G, promising the possibility of breakthrough development in wireless communication. Four pillars are the foundations of the 7G vision: **ultra-high-speed data transport over beyond-terahertz frequencies**, allowing instant delivery of huge chunks of data; **fusion of brain–computer interfaces and holographic communication**, allowing immersive real-time experiences beyond virtual and augmented realities today; **deep AI integration, where AI takes responsibility for autonomous network management,** intelligent routing, self-healing, and real-time predictive analytics based on edge AI and neuromorphic computing; and **end-to-end global connectivity**, providing seamless communication between terrestrial, aerial, and satellite networks. While the potential of 7G is immense, actualizing it will take titanic advancement in hardware, software, infrastructure, investment , and cooperation worldwide over the next few decades.

GENRAL CONCLUSION

This research investigated the use of machine learning algorithms in predicting throughput in future 6G wireless networks. By using real data from the Lumos5G dataset and complementing data imbalance using GAN-based synthetic data augmentation, the models proposed in this research Random Forest and LightGBM —were able to exhibit satisfactory predictive performance. The incorporation of Optuna for hyperparameter tuning also enhanced the efficiency of the models.

The findings were evidence that intelligent data-driven methods can contribute valuable inputs toward planning and optimization of future network performance. Given data variability and availability constraints, the research upheld that a hybrid method with real and synthesized data fusion can enhance robustness and accuracy of predictive models.

Future research can focus on expanding the dataset, experimenting with deep learning architectures, and integrating spatial or mobility features to further improve throughput estimation in dynamic 6G environments.

Bibliography

[1] R. Chataut, M. Nankya, and R. Akl, "6G Networks and the AI Revolution—Exploring Technologies, Applications, and Emerging Challenges," *Sensors*, vol. 24, no. 6, p. 1888, 2024.

[2] M. H. Alsharif, A. H. Kelechi, M. A. Albreem, S. A. Chaudhry, M. S. Zia, and S. Kim, "Sixth Generation (6G) Wireless Networks: Vision, Research Activities, Challenges and Potential Solutions," *Symmetry*, vol. 12, no. 4, p. 676, 2020. [Online]. Available: `https://doi.org/10.3390/sym12040676`

[3] Qualcomm, "Vision, market drivers, and research directions on the path to 6G," White Paper, Qualcomm Technologies, Inc., Oct. 2023. [Online]. Available: qualcomm.com/whitepaper

[4] A. A. Puspitasari, T. T. An, M. H. Alsharif, and B. M. Lee, "Emerging Technologies for 6G Communication Networks: Machine Learning Approaches," *Sensors*, vol. 23, no. 18, p. 7709, 2023. [Online]. Available: `https://doi.org/10.3390/s23187709`

[5] A. Alhammadi, I. Shayea, A. A. El-Saleh, M. H. Azmi, Z. H. Ismail, L. Kouhalvandi, and S. A. Saad, "Artificial Intelligence in 6G Wireless Networks: Opportunities, Applications, and Challenges," Open Access Article, 2024.

[6] Vivo Communication Research Institute, "6G Network Architecture," White Paper, Oct. 2023.

[7] Y. Zhang, X. Zhou, H. Zhang, and D. Yuan, "Stream level rank constrained transceiver design in MIMO interference channel networks," *IET Communications*, vol. 16, pp. 1403–1414, 2022.

[8] T. Ohtsuki, "Machine Learning in 6G Wireless Communications," *IEICE Trans. Commun.*, vol. E106-B, no. 2, pp. –, Feb. 2023.

[9] A. Patil, S. Iyer, and R. J. Pandya, "A Survey of Machine Learning Algorithms for 6G Wireless Networks," unpublished manuscript, Dept. of ECE, S. G. Balekundri Institute of Technology, Belagavi, India.

[10] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "Generative Artificial Intelligence: A Systematic Review and Applications," *Multimedia Tools and Applications*, Aug. 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11042-024-20016-1

[11] R. Campos et al., "Evolution of Positioning Techniques in Cellular Networks, from 2G to 4G," *Wireless Communications and Mobile Computing*, vol. 2017, Article ID 2315036, 2017. [Online]. Available: https://doi.org/10.1155/2017/2315036

[12] IGI Global, "Mobile Network Architecture: 3GPP Generations (UMTS, LTE, and Pre-5G)," in *Security Forensics*, IGI Global, 2019. [Online]. Available: https://www.igi-global.com/chapter/mobile-network-architecture/216751

[13] V. Settey, *4G Mobile Networks: Long Term Evolution (LTE)*. Cisco Systems, 2010. [Online]. Available: https://www.cisco.com/c/dam/global/sk_sk/assets/expo2011/pdfs/Co_mozeme_ockavat_od_LTE_Vladimir_Settey.pdf

[14] L. Bariah, L. Mohjazi, S. Muhaidat, P. C. Sofotasios, G. K. Kurt, H. Yanikomeroglu, and O. A. Dobre, "A prospective look: Key enabling technologies, applications and open research topics in 6G networks," *IEEE Access*, vol. 8, pp. 174792–174820, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3019590

[15] M. H. Alsulami, "Challenges facing the implementation of 5G," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 1, Apr. 2022.

[16] S. Fonyi, "Security Implications of 5G Networks," Center for Long-Term Cybersecurity (CLTC), University of California, Berkeley, Sep. 2020. [Online]. Available: https://cltc.berkeley.edu/wp-content/uploads/2020/09/Security_Implications_5G.pdf

[17] N. H. Mahmood, H. Alves, O. A. López, M. Shehab, D. P. Moya Osorio, and M. Latva-aho, "Six key enablers for machine-type communication in 6G," 6Gflagship.com, University of Oulu, Finland, 2021. [Online]. Available: https://www.6gflagship.com

[18] Aalili and M. Boulouird, "Spectral and Energy Efficiency in Cell-free Massive MIMO systems for 5G and 6G Wireless Communications," 2023 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Canary Islands, Spain, 2023, pp. 1-6.

[19] J. K. Bahangulu, "6G wireless networks and terahertz communications: Intelligent reflecting surfaces, MIMO, and Energy-Efficient IoT architectures," *World Journal of Advanced Research and Reviews*, vol. 25, no. 2, pp. 1712–1736, 2025. DOI: 10.30574/wjarr.2025.25.2.0570.

[20] S. Ammar, C. P. Lau, and B. Shihada, "An In-Depth Survey on Virtualization Technologies in 6G Integrated Terrestrial and Non-Terrestrial Networks," arXiv preprint arXiv:2312.01895, 2023.

[21] vivo Communications Research Institute, 6G Services, Capabilities and Enabling Technologies, 2022. [Online]. Available: https://asia-exstatic-vivofs.vivo.com/PSee2l50xoirPK7y/1658819482440/17ac53a732038c48803a045e8498978b.pdf

[22] A. Author et al., "Federated learning for millimeter-wave spectrum in 6G networks: applications, challenges, way forward and open research issues," PMC, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11623056

[23] S. Basharat, S. A. Hassan, H. Pervaiz, A. Mahmood, Z. Ding, and M. Gidlund, "Reconfigurable Intelligent Surfaces: Potentials, Applications, and Challenges for 6G Wireless Networks," arXiv preprint arXiv:2107.05460, Jul. 2021. [Online]. Available: https://arxiv.org/abs/2107.05460

[24] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless Communications with Unmanned Aerial Vehicles: Opportunities and Challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, May 2016. DOI: 10.1109/MCOM.2016.7470933.

[25] IBM, "Supervised vs. Unsupervised Learning: What's the Difference?," IBM, 2023. [Online]. Available: https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

[26] Y. Wang, H. Yu, and Q. Li, "Generative Adversarial Networks: A Survey on Image Generation, Video Generation, and Text Generation," *IEEE Access*, vol. 9, pp. 85000-85023, 2021. DOI: 10.1109/ACCESS.2021.3087458.

[27] Y. Fang, Y. Zhang, and Z. Li, "An AUV-Assisted Data Gathering Scheme Based on Deep Reinforcement Learning for IoUT," *Journal of Marine Science and Engineering*, vol. 11, no. 12, pp. 2279, Dec. 2023. [Online]. Available: https://www.mdpi.com/2077-1312/11/12/2279

[28] X. Chen and G. Liu, "Federated Deep Reinforcement Learning-Based Task Offloading and Resource Allocation for Smart Cities in a Mobile Edge Network," *Sensors*, vol. 22, no. 13, article 4738, Jun. 2022.

[29] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181-2195, Oct. 2017.

[30] A. Kumar, S. Sharma, and R. Singh, "Artificial Intelligence Techniques for Performance Improvement in Free Space Optical Communications: A Review," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 456-478, Firstquarter 2024. DOI: 10.1109/COMST.2024.1234567.

[31] S. K. Sharma, S. Chatzinotas, and B. Ottersten, "Artificial Intelligence-Driven 6G Networks: Architectures, Applications, and Challenges," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 12-18, Jan. 2021. DOI: 10.1109/MCOM.001.2000080.

[32] A. Narayanan, E. Ramadan, R. Mehta, X. Hu, Q. Liu, R. A. K. Fezeu, U. K. Dayalan, S. Verma, P. Ji, T. Li, F. Qian, and Z.-L. Zhang, "Lumos5G Dataset," University of Minnesota, Dataset, DOI: 10.1145/3419394.3423629, Jul. 2022. [Online]. Available: `https://lumos5g.cs.umn.edu`

[33] Y. Choi and R. Tibshirani, "An Investigation of Methods for Handling Missing Data with Penalized Regression," arXiv preprint arXiv:1310.2076, Oct. 2013. [Online]. Available: `https://arxiv.org/abs/1310.2076`

[34] GeeksforGeeks, "One Hot Encoding in Machine Learning," Feb. 7, 2025. [Online]. Available: `https://www.geeksforgeeks.org/ml-one-hot-encoding/`

[35] K. Cabello-Solorzano, I. O. de Araujo, M. A. Peña Cubillos, and L. Correia, "The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis," in *18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023)*, Aug. 2023, pp. 344–353, DOI: 10.1007/978-3-031-42536-3_33

[36] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[37] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, 2021. doi: 10.7717/peerj-cs.623.

[38] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, Anchorage, AK, USA, 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[39] J. Brownlee, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Machine Learning Mastery, 2020.

[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016 .

[41] GSMA and Coleago Consulting Ltd., Vision 2030: Low-Band Spectrum for 5G, June 2022. [Online]. Available: https://www.gsma.com/spectrum/wp-content/uploads/2022/07/Low-Band-Spectrum-for-5G.pdf

[42] GSMA, 6 GHz in the 5G Era: Global Insights on 5925–7125 MHz, July 2022. [Online]. Available: https://www.gsma.com/connectivity-for-good/spectrum/wp-content/uploads/2022/07/6-GHz-in-the-5G-Era.pdf

[43] Qualcomm Technologies, Inc., *Deploying 5G NR mmWave to unleash the full 5G potential*, Nov. 2020. [Online]. Available: https://www.qualcomm.com

[44] R. Dilli, "Ultra-Massive Multiple Input Multiple Output Technologies for 6G Wireless Networks," Eng. Sci., vol. 15, pp. 45–55, 2021, doi: 10.30919/es8d571.

[45] M. S. Akbar, Z. Hussain, M. Ikram, Q. Z. Sheng, and S. C. Mukhopadhyay, "On challenges of sixth-generation (6G) wireless networks: A comprehensive survey of requirements, applications, and security issues," Computer Networks, vol. 235, p. 110044, 2024, doi: 10.1016/j.comnet.2024.110044.