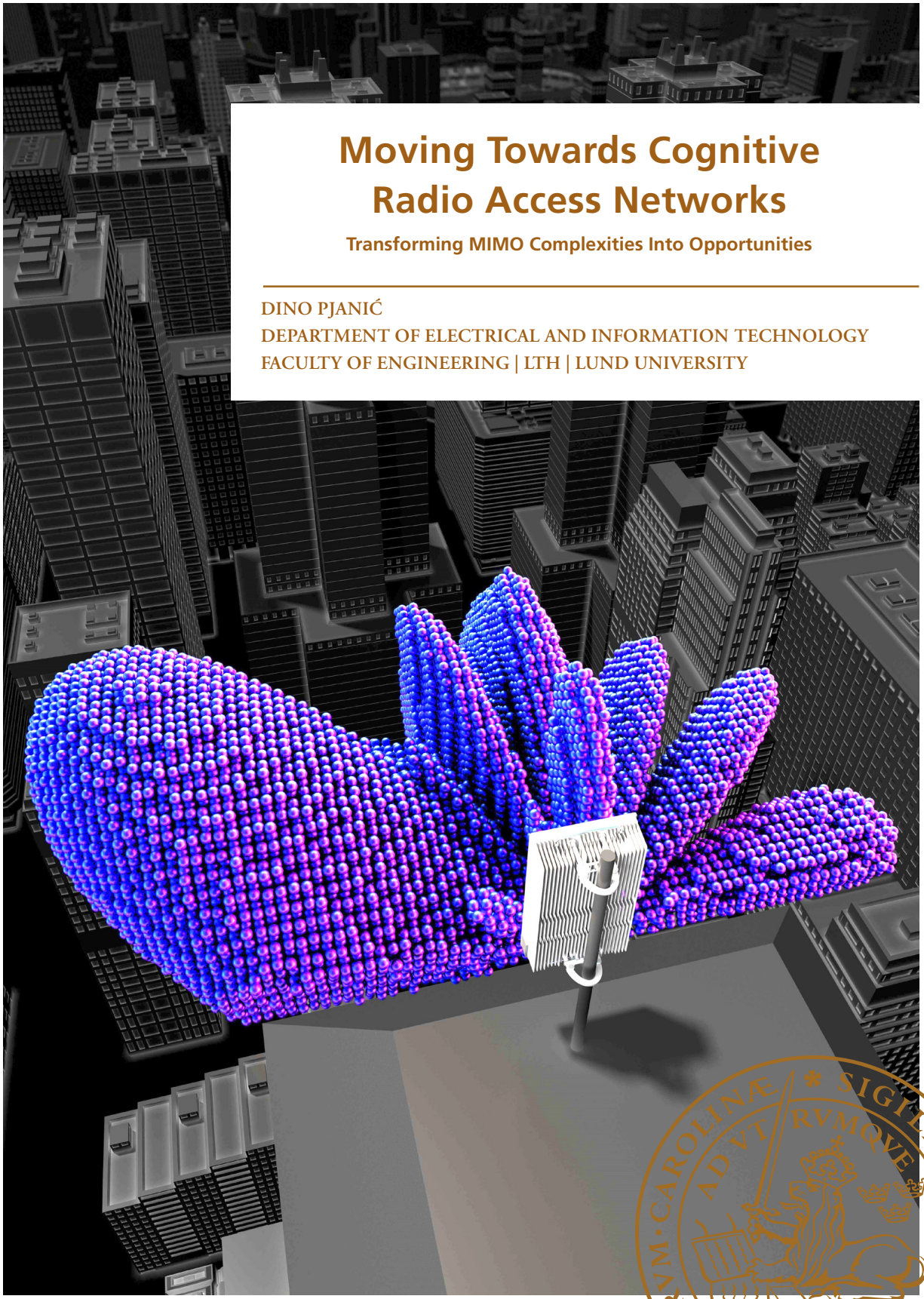


Moving Towards Cognitive Radio Access Networks

Transforming MIMO Complexities Into Opportunities

DINO PJANIĆ
DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING | LTH | LUND UNIVERSITY

Moving Towards Cognitive Radio Access Networks

Transforming MIMO Complexities Into Opportunities

Doctoral Thesis

Dino Pjanić

LUND
UNIVERSITY

Department of Electrical and
Information Technology
Lund, May 2025

Academic thesis for the degree of Doctor of Philosophy, which, by due permission of the Faculty of Engineering at Lund University, will be publicly defended on Thursday, 5 June, 2025, at 9:15 a.m. in lecture hall E:1406, Department of Electrical and Information Technology, Ole Römers Väg 3, 223 63 Lund, Sweden. The thesis will be defended in English.

The Faculty Opponent is Professor Laurent Clavier, Mines-Telecom Institute (IMT Nord Europe), University of Lille, France.

Organisation:
LUND UNIVERSITY
Department of Electrical and Information Technology
Ole Römers Väg 3
223 63 Lund
Sweden

Author:
Dino Pjanić

Document Type:
DOCTORAL THESIS

Date of Issue:
8 May 2025

Sponsoring Organisation(s):
Ericsson AB, Sweden
Swedish Foundation for Strategic Research (SSF)

Title: Moving Towards Cognitive Radio Access Networks: Transforming MIMO Complexities Into Opportunities

Abstract:

The introduction of Multiple-Input Multiple-Output (MIMO) systems has dramatically transformed wireless communication systems, in particular in the Fifth Generation (5G) New Radio (NR) systems, fundamentally changing how signals are transmitted and received. MIMO technology deploys numerous antennas to transmit and receive multiple data streams simultaneously. As user devices move, the interaction of electromagnetic radio waves with surrounding objects and devices generates distinct patterns, referred to as spatial fingerprints. The presence of obstructions and scatterers in wireless environments, varying in location, size, and shape, contributes to a high-dimensional feature space. By analyzing the behavior of the radio channel in real-time through these spatial fingerprints and their temporal evolution, MIMO systems unlock significant opportunities for deeper insights into channel dynamics. These insights lay the groundwork for previously unforeseen functionalities in the Radio Access Network (RAN) domain of cellular networks, moving beyond the constraints of traditional approaches based on mathematical models and solutions. This thesis primarily aims to utilize channel measurements generated in commercial MIMO systems to explore the underlying statistical structures from real-time data, eliminating or mitigating the need for precise mathematical modeling. Recognizing that conventional mathematical models and solutions are unlikely to deliver the performance enhancements required for future wireless networks, the research explores innovative approaches and tools to push the boundaries of this field. Over the past decade, a new era of Machine Learning (ML) and Artificial Intelligence (AI) techniques, particularly Deep Learning (DL), has emerged as a powerful alternative for designing and optimizing wireless networks, as demonstrated in this thesis. The subsequent chapters of this thesis begin with an introductory section that outlines the theoretical background that serves as the foundation for the research topic. This is followed by a collection of paper publications detailing the conducted studies. The six papers included in this thesis encompass three key research areas: user device clustering, user positioning, and traffic pattern-related predictions. The first and third papers focus on user classification, or grouping, based on channel fingerprints derived from measurement data. The first paper explores the feasibility of using channel measurements as a data source for classifying users based on their spatial proximity, density, and velocity while the third paper demonstrates grouping based on user position and heading direction using commercial 5G measurement data. The second research area focuses on cellular positioning, with the second paper being among the first to demonstrate the feasibility of user positioning using commercial 5G New Radio (NR) beam measurement data. Building upon the findings of the second paper, the fifth paper refines positioning accuracy through an attention-based AI model and advanced statistical post-processing techniques. The results showcase a sub-meter level of positioning accuracy. As part of the third research area, traffic pattern-related predictions, the fourth paper proposed a cell handover prediction strategy tailored for dense urban environments. This work emphasizes a user-context-aware handover process, aiming to enhance the efficiency and reliability of handovers in complex network scenarios. Finally, the sixth paper provides novel insights into 5G beam management strategies for both long- and short-term channel predictions. This innovative approach introduces a highly accurate, attention-based prediction model capable of deriving the complete downlink transmission chain in a commercial-grade 5G system. The model demonstrates precise beam predictions extending far beyond coherence time, specifically addressing the challenges posed by Non Line-of-Sight (NLoS) environments characterized by complex, high-dimensional channel dynamics.

Keywords:
Artificial Intelligence, Machine Learning, Massive MIMO, Radio Access Networks

Classification System and/or Index Terms
Electronic Engineering, Communications Engineering

Language:
English

Supplementary Bibliographical Information:
–

ISBN (printed):
978-91-8104-485-0

Key title and ISSN:
Series of Licentiate and Doctoral Theses; 1654-790X, No. 184

ISBN (digital):
978-91-8104-486-7

Recipient's Notes:

Number of Pages:
195

Price:
No commercial issue

Security Classification:
Unclassified

Signature:  *Date:* 8 May 2025

Moving Towards Cognitive Radio Access Networks: Transforming MIMO Complexities Into Opportunities

Doctoral Thesis

Dino Pjanić

LUND
UNIVERSITY

Department of Electrical and
Information Technology
Lund, May 2025

Dino Pjanić
Department of Electrical and Information Technology
Lund University
Ole Römers Väg 3, 223 63 Lund, Sweden

Series of Licentiate and Doctoral Theses
ISSN 1654-790X, No. 184
ISBN 978-91-8104-485-0 (printed)
ISBN 978-91-8104-486-7 (digital)

This thesis is typeset using LaTeX 2ε with the body text in Palatino and Goudy Initials, headings in Helvetica, text in figures in Arial.

Frontispiece: Electromagnetic waves transmitted by a base station equipped with a MIMO antenna system are shaped into beams. Copyright: Ericsson AB

Printed by Tryckeriet i E-huset, Lund University, Lund, Sweden.

"Att våga är att förlora fotfästet för en sekund. Att inte våga är att förlora sig själv."

Sören Kirkegaard

Abstract

The introduction of Multiple-Input Multiple-Output (MIMO) systems has dramatically transformed wireless communication systems, in particular in the Fifth Generation (5G) New Radio (NR) systems, fundamentally changing how signals are transmitted and received. MIMO technology deploys numerous antennas to transmit and receive multiple data streams simultaneously. The presence of obstructions and scatterers in wireless environments, varying in location, size, and shape, contributes to a high-dimensional feature space. As user devices move, the interaction of electromagnetic radio waves with surrounding objects and devices generates distinct patterns, called spatial fingerprints. By analyzing the behavior of the radio channel in real time through these spatial fingerprints and their temporal evolution, MIMO systems unlock significant opportunities for deeper insight into channel dynamics. These insights lay the groundwork for previously unforeseen functionalities in the Radio Access Network (RAN) domain of cellular networks, moving beyond the constraints of traditional approaches based on mathematical models and solutions.

This thesis aims primarily to utilize channel measurements generated in commercial MIMO systems to explore the underlying statistical structures from real-time data, eliminating or mitigating the need for precise mathematical modeling. Recognizing that conventional solutions are unlikely to deliver the performance enhancements required for future wireless networks, the research presented in this thesis explores innovative approaches and tools to push the boundaries of this field. Over the past decade, a new era of Machine Learning (ML) and Artificial Intelligence (AI) techniques, particularly Deep

Learning (DL), has emerged as a powerful alternative for designing and optimizing wireless networks, as demonstrated in this thesis. The subsequent chapters of this thesis begin with an introductory section that outlines the theoretical background that serves as the foundation for the research topic. This is followed by a collection of papers that detail the conducted studies. The six papers included in this thesis encompass three key research areas: user device clustering, user positioning, and traffic pattern-related predictions.

The first and fourth papers focus on user classification, or grouping, based on channel fingerprints derived from measurement data. The first paper explores the feasibility of using channel measurements as a data source to classify users based on their spatial proximity, density, and velocity. In contrast, the fourth paper demonstrates grouping based on user position and direction using commercial 5G measurement data.

The second research area focuses on cellular positioning, with the second paper among the first to demonstrate the feasibility of user positioning using commercial 5G NR beam measurement data. Building upon the findings of the second paper, the fifth paper refines positioning accuracy through an attention-based AI model and advanced statistical post-processing techniques. The results showcase a sub-meter level of positioning accuracy.

As part of the last research area, traffic pattern-related predictions, the third paper proposed a customized cell handover prediction strategy for dense urban environments. This work emphasizes a user-context-aware handover process, with the aim of improving the efficiency and reliability of handovers in complex network scenarios. Finally, the sixth paper provides novel insights into 5G beam management strategies for long- and short-term channel predictions. This innovative approach introduces a highly accurate, attention-based prediction model capable of deriving the complete downlink transmission chain in a commercial-grade 5G system. The model demonstrates precise beam predictions extending far beyond coherence time, specifically addressing the challenges posed by Non Line-of-Sight (NLOS) environments characterized by complex, high-dimensional channel dynamics.

Populärvetenskaplig sammanfattning

Socialt samspel och kommunikation är inneboende i människans natur. Medan tidig mänsklig kommunikation huvudsakligen var ansikte mot ansikte, har det de senaste två århundradena uppstått ett behov av fjärranslutning över avstånd bortom räckvidden för den mänskliga rösten. Denna efterfrågan ökade avsevärt under den industriella eran, vilket drev utvecklingen av mer praktiska och allmäntillgängliga telekommunikationssystem. Dessa framsteg ledde till uppfinningen av trådbundna teknologier, såsom telegrafen och senare telefonen. Dock hade trådbundna lösningar begränsningar eftersom de krävde att användarna befann sig på specifika platser vid bestämda tider för att skicka eller ta emot meddelanden eller samtal. Denna brist på flexibilitet drev på sökandet efter globala lösningar och banade väg för den sammanlänkade värld vi upplever idag. Utvecklingen av trådlös cellulär kommunikation sträcker sig över århundraden och började med grundläggande teorier om elektromagnetism på 1800-talet. Banbrytande arbete, såsom Maxwells ekvationer och Hertzs experiment, lade grunden för Marconis trådlösa kommunikation över långa avstånd via morsekod. Dessa innovationer blev avgörande för modern vetenskap och ingenjörskonst och visade på potentialen och fördelarna med trådlös kommunikation. Betydande framsteg under 1900-talet gjorde många av dessa idéer praktiskt genomförbara, inklusive introduktionen av radioutsändningar och konceptualiseringen av cellulära nätverk på 1940-talet, vilken fokuserade på trådlös röstkommunikation.

Under 1980-talet lanserades första generationen av analoga nätverk, 1G, följt av den digitala revolutionen på 1990-talet med 2G-nätverk, som introducerade förbättrad röstkvalitet, textmeddelanden och internetåtkomst. På 2000-talet möjliggjorde 3G mobilt bredband, medan 2010-talet inledde 4G-eran, som erbjöd höghastighetsuppkoppling för dataintensiva applikationer. Vid 2020-talet hade 5G-teknik introducerats, kännetecknad av ultralåg för-

dröjning, massiv uppkoppling samt stöd för autonoma system och industriell automation. Historien om trådlös kommunikation visar samspelet mellan vetenskaplig innovation, teknologisk utveckling och samhällsbehov, vilket har gjort uppkoppling till en grundpelare i det moderna samhället.

Introduktionen av avancerade antennsystem under den senare delen av 4G-eran markerade en avgörande utveckling genom användning av MIMO-teknik. Dessa system öppnade nya möjligheter för att förstå kanalegenskaper. En passande analogi för dessa system är ett stort astronomiskt observatorium som använder en uppsättning teleskop spridda över hela världen för att observera samma avlägsna objekt. Varje teleskop samlar in en unik del av ljuset eller radiovågorna, påverkad av dess position och vinkel. Genom att kombinera data från alla teleskop skapas en mer detaljerad och multidimensionell bild, som överträffar vad ett enskilt teleskop kan åstadkomma.

Under de senaste åren har moderna mobilnät ställts inför nya krav utöver den traditionella kommunikationen mellan användare och nätverket, inklusive integrerad trådlös avkänning. Trådlös avkänning utnyttjar befintliga kommunikationssignaler för att uppfatta och tolka omgivningen. Ur ett nätverksperspektiv innebär detta att mobilnätsinfrastrukturen används för att möjliggöra avkänning utan behov av dedikerade sensorer. Istället för att enbart upprätthålla datakommunikation kan nätverket analysera signaler för att upptäcka objekt, rörelser, användare och omgivande miljöförändringar. Denna avhandling undersöker delvis olika aspekter av trådlös avkänning, ett forskningsområde som fortfarande är i ett tidigt skede vid tiden för skrivandet.

När denna avhandling skrivs genomgår telekommunikationsindustrin en övergång mot nästa generation av nätverk, 6G, med fokus på terabithastigheter, holografisk kommunikation och AI-drivna nätverk. Till skillnad från traditionella kommunikationsparadigmer, som bygger på ett reaktivt förhållningssätt där mottagare väntar på signaler, kommer kognitiva cellulära nätverk att använda prediktiva funktioner för att förutse trafikbehov och kommande händelser. Denna avhandling utnyttjar AI, som till skillnad från traditionella programmerade system är adaptiv, probabilistisk och kapabel att lära sig från data för att fatta autonoma beslut. Frågan om hur AI kan möjliggöra kognitiva mobila nätverk genom användning av MIMO-baserade "teleskop" för att förutsäga användares rörelsemönster eller radiokanalens egenskaper, både på korta och långa tidshorisonter, behandlas i det följande. AI-teknologi har en transformativ potential att revolutionera driften av mobila nätverk genom att utnyttja historiska data för att möjliggöra intelligent, autonom funktionalitet med minimalt mänskligt ingripande.

Popular Scientific Summary

OCIAL interaction and communication are intrinsic to human nature. While early human communication was primarily face-to-face, the past two centuries have brought about a need for remote connectivity over distances beyond the range of the human voice. This demand grew significantly during the industrial era, driving the development of practical and accessible telecommunication systems. These advancements led to the invention of wired technologies, such as the telegraph and later the telephone. However, wired solutions had limitations, requiring users to be in specific locations at designated times to send or receive messages or calls. This inflexibility spurred the quest for global solutions, paving the way for the interconnected world we experience today. The evolution of wireless cellular communication spans centuries, beginning with foundational theories of electromagnetism in the 19th century. Groundbreaking work, such as Maxwell's equations and Hertz's experiments, laid the foundation for Marconi's long-distance wireless communication via Morse code. These innovations became pivotal in modern science and engineering, showcasing the potential and advantages of wireless communication. Significant milestones in the 20th century made many of these ideas practical, including the advent of radio broadcasting and the conceptualization of cellular networks in the 1940s, which focused on wireless voice communication.

The 1980s saw the launch of the first analog networks, 1G, followed by the digital revolution in the 1990s with 2G networks, which introduced improved voice quality, text messaging and internet access. In the 2000s, 3G enabled mobile broadband, while the 2010s ushered in 4G, offering high-speed connectivity to support data-intensive applications. By the 2020s, 5G technology emerged, characterized by ultra-low latency, massive device connectivity, support for autonomous systems, and industrial automation. The history of wireless cellular communication demonstrates the interplay

of scientific innovation, technological progress, and societal needs, making connectivity a cornerstone of modern society.

The introduction of advanced antenna systems in the late 4G era marked a pivotal advancement, employing MIMO technology. These systems opened new avenues for understanding channel characteristics. An apt analogy for these systems is a large astronomical observatory employing an array of telescopes scattered across the globe to observe the same distant object. Each telescope collects a unique portion of the light or radio waves, influenced by its position and angle. By combining the data from all telescopes, a more detailed and multi-dimensional image is created, surpassing what a single telescope could achieve.

In recent years, modern cellular networks have faced new requirements beyond traditional communication between users and the network, including integrated wireless sensing. Wireless sensing utilizes existing communication signals to perceive and interpret the surrounding environment. From a network perspective, this involves leveraging cellular infrastructure to enable sensing without the need for additional dedicated sensors. Rather than solely transmitting data, the network can analyze signals to detect object presence and movement, user positioning, and environmental changes. This thesis partly examines various aspects of wireless sensing, a field still in its early stages at the time of writing.

At the time of this writing, the telecommunications industry is transitioning toward the next generation of networks, 6G, focusing on terabit-level speeds, holographic communication, and AI-driven networks. Unlike traditional communication paradigms, which rely on a reactive approach where receivers wait for signals, cognitive cellular networks will utilize predictive capabilities to anticipate traffic demands and upcoming events. This thesis leverages AI, which, unlike traditional programmed devices, is adaptive, probabilistic, and capable of learning from data to make autonomous decisions. The question of how AI can enable cognitive mobile networks by utilizing MIMO-based "telescopes" to predict user movement patterns or radio channel characteristics over both short and long time horizons is addressed in the following. AI technology holds transformative potential to revolutionize cellular network operations by leveraging historical data to enable intelligent, autonomous functionality with minimal human intervention.

Acknowledgments

THERE are so many people who have shaped me from a young age and in a way made this journey possible. Trying to mention everyone meaningfully within the limits of two pages is quite a challenge; however, in the following I name those who meant the most to me during this journey and acted as formal and informal mentors and patrons.

I would like to thank Ericsson AB for being such a supportive employer, including many talented colleagues who pushed me forward over the finish line. I am very humbled and grateful to *Andres Reial, Niclas Holmqvist, Linda Persson, Fredrik Dahlgren, Johan Eker, Anders Henriksson, Daniel Landström, Mats Melander, and Björn Ekelund* who, in different ways, conducted me to and through the industrial PhD program at the company. Special thanks to the Swedish Foundation for Strategic Research (SSF) for believing in my research project and funding me throughout my PhD pursuit.

I express my sincere gratitude to my main supervisor, Prof. *Fredrik Tufvesson*, for believing in me and my vague initial research idea, and for helping me shape and refine it until it could truly fly. Thank you for providing me with support throughout this academic marathon — all the way to the finish line!

My sincere and deepest thanks to my co-authors and mentors, *Alexandros Sopasakis* from the Department of mathematics for our lengthy discussions and for sharing your extensive knowledge, and to *Harsh Tataria* for his valuable advice and for helping me rediscover the fascinating enigmas of wireless technology. To *Guoda Tian* and *Xuesong Cai* for their invaluable collaboration and contributions. My sincere thanks also go to the master's students I had the pleasure of supervising, *Andre Ráth* and *Korkut Emre Arslantürk*, your curiosity and dedication brought fresh energy to my research.

I also express my gratitude to my co-supervisor Prof. *Bo Bernhardsson* and also Prof. *Ove Edfors* and Prof. *Maria Kihl* for their support.

I am deeply grateful to my dear colleagues at LTH, with whom I had the pleasure of working, collaborating, and celebrating. Thank you for your support, insightful discussions, and for being there when I needed guidance and a helping hand to reach the finish line - you all ROCK, *Sara, Xuhong, Christian, Harsh, Ali, Ilayda, Anders, Umar, Fredik Rusek, Jesús, Juan Vidal Alegría, Meifang, Aleksei, Aleksandar, Naharika, Yingjie, Juan Sanchez, Hedieh, Ashkan, Xuesong, Michiel, Vincent, Emil, and William*.

To industrial PhD colleagues, *Russ* and *Junshi* for all good and late discussions at the very back of the EIT corridor.

To my dear parents, I thank you for your unconditional love and dedicated support throughout my life. Finally, there is a *real* "Dr." in our family, but I am not doing this again!

Dragi roditelji, hvala vam na vašem neumornom zalaganju, strpljenju i bezrezervnoj podršci. Vaša vjera u mene bila je temelj mog uspjeha. Volim vas.

Finally, I thank my dear wife, my best friend and soulmate *Anjali*, for her endless encouragement and support.

Lund, May 2025

Contents

Abstract v

Populärvetenskaplig sammanfattning vii

Popular Science Summary ix

Acknowledgments xi

Contents xiii

Preface xvii

Structure of the Thesis xvii

Included Papers xviii

Acronyms and Symbols xxv

Acronyms and Abbreviations xxv

Introduction 1

1 Background and Overview of the Research Field 3

1.1 Evolution of Cellular Networks 4

1.2 Cellular System Architecture 5

1.2.1 Core network Domain 5

1.2.2 Radio Access Network Domain 7

1.3 AI in cellular networks: Trends and Future Outlook 8

1.4 Research Questions 9
1.5 Research Boundaries and Practical Restraints 11

2 Fundamentals of Wireless Communications 13
2.1 Basics of Electromagnetic Wave Theory 13
2.2 Channel Modeling and Propagation 14
2.2.1 Physical Channel Modeling 16

3 Introduction to Massive MIMO systems 19
3.1 Antenna Arrays as a Key Enabler of MIMO Technology 20
3.2 Signal Model and Processing 22
3.3 Massive MIMO Techniques 24
3.3.1 Beamforming 24
3.3.2 Spatial Multiplexing 25
3.3.3 Nullforming 26
3.4 Channel State Information in MIMO Systems 26

4 Machine Learning and Artificial Intelligence 29
4.1 Machine Learning Basics 30
4.1.1 Types of Machine Learning 31
4.1.2 Supervised Learning 32
4.1.3 Unsupervised Learning 32
4.1.4 Reinforcement Learning 33
4.2 Machine Learning Algorithms Explored 33
4.2.1 Clustering 33
4.2.2 Neural Networks 34
4.2.3 Transformers 35
4.3 Wireless Channel As Input To ML Models 36
4.3.1 Channel Transfer Function 36
4.3.2 Channel Impulse Response 36
4.3.3 Fingerprinting 37
4.3.4 A Note on SRS data 39

5 Conclusions and Outlook 41
5.1 Research Contributions 41
5.2 Future Perspectives of MIMO and AI In Cellular Networks . . . 42

5.2.1 Data Collection and Data-Processing Aspects 42
5.2.2 Integration of Native AI/ML Technologies 43
5.2.3 Data Integrity and AI/ML Trustworthiness 43
5.2.4 Beyond 5G and 6G Vision 44

Bibliography **45**

PAPERS **55**

I Learning-Based UE Classification in Millimeter-Wave Cellular Systems With Mobility **57**

II ML-Enabled Outdoor User Positioning in 5G NR Systems via Uplink SRS Channel Estimates **73**

III Early-Scheduled Handover Preparation in 5G NR Millimeter-Wave Systems **91**

IV Dynamic User grouping based on Location and Heading in 5G NR System **121**

V Attention-aided Outdoor Localization In Commercial 5G NR Systems **135**

VI Illuminating the Path: Attention-Assisted Beamforming and Predictive Insights in 5G NR Systems **173**

Preface

This thesis represents the realisation of research conducted between February 2019, when I joined the Industrial PhD program at Ericsson Sweden, and May 2025. During this period, I was part of the *Wireless Communications Engineering* group at Lund University, under the supervision of Professor *Fredrik Tufvesson* and Principal Researcher *Andres Reial* at Ericsson.

STRUCTURE OF THE THESIS

This thesis comprises an introduction section that offers a high-level overview of the research topic, including preliminary knowledge and an in-depth discussion of three key pillars: cellular network architecture, wireless channel modeling, and machine learning and artificial intelligence. The introduction is designed to be self-contained, providing sufficient material for readers interested in the research topic but with emphasis on topics covered by the included papers. The thesis includes three conference papers and three journal papers authored in collaboration with others, which were published in or submitted to scientific journals and conference proceedings. These papers are reprinted with permission from the publishers and form the main body of the thesis. The concluding chapter summarizes the thesis and presents a vision for future research in this field.

- **INTRODUCTION**
  The primary focus of this thesis was to explore whether and how 5G MIMO systems can be optimized using ML and AI, targeting various aspects of network optimization across both the lower bands below 6 GHz and the higher band around 30 GHz. Massive MIMO systems, with their numerous antennas, provide unique opportunities to gain deeper insights into channel behavior. By studying real-world channel behavior, most

often through data collected from commercial systems, the viability of spatial fingerprints and their temporal evolution were examined in detail. The research specifically investigated how accurate user positioning can be achieved by leveraging radio features inherent in the environment. Additionally, it explored fingerprint-based machine learning approaches that operate without explicitly modeling these radio features. The work demonstrated that by utilizing the unique spatial fingerprints users create at the base station, it is possible to achieve accurate predictions of channel behavior, mobility patterns, and time-advance protocol messages. The structure of this thesis is as follows:

  - Chapter 1 provides a high-level overview of cellular networks and their architecture, tracing the evolution from early 1G systems to today's advanced 5G networks.
  - Chapter 2 introduces the theoretical foundations of wireless channel modeling, which serve as the basis for the conducted research.
  - Chapter 3 presents the fundamentals of MIMO systems.
  - Chapter 4 delves into machine learning and artificial intelligence, highlighting their applications in the wireless industry.
  - Chapter 5 concludes the introductory section and outlines a vision for future research directions.

- **PAPERS**
  The six papers that constitute the main body of this thesis are reproduced in the dedicated chapter and are listed below, accompanied by a brief description of my contributions to each.

INCLUDED PAPERS

The following papers form the main body of this thesis and the respective published or draft versions are appended in the back.

Paper I: D. Pjanić, A. Sopasakis, H. Tataria, F. Tufvesson, and A. Reial, "Learning-Based UE Classification in Millimeter-Wave Cellular Systems With Mobility", *IEEE International Workshop on Machine Learning for Signal Processing*, Oct. 2021, Gold Coast, Australia, doi: 10.1109/MLSP52302.2021.9596275.

► ***Research Contributions:*** *The paper explores the clustering and classification of UEs based solely on their network measurement reports,*

without relying on physical positioning or additional supporting information. The findings demonstrate that it is possible to infer the mobility mode of UEs through such an approach. Higher-layer channel measurement reports, including time-evolving Channel State Information - Reference Signal (CSI-RS) signals from dynamic millimeter-wave scenarios, can be utilized as input to both traditional supervised and unsupervised machine learning methods for UE classification based on velocity and mobility patterns. This classification can, in turn, aid in predicting and optimizing radio resource requirements. The work provides valuable insights into developing new beam prediction mechanisms for mobility-aware MIMO scenarios. At the time, I didn't realize that we were conducting a study on wireless sensing. Furthermore, when combined with positioning or trajectory estimation (the focus of Paper II and V), these results could prove instrumental in preparing handovers (the focus of Paper IV) and anticipating resource demands.

► ***Personal Contributions:*** *This was my first hands-on experience with machine learning, integrated with dimensionality reduction techniques like Principal Component Analysis (PCA), a concept I had learned during the first two years of my PhD studies. I originated the idea and also designed and conducted all simulations, performed data collection, data post-processing and analysis, authored the paper, and incorporated feedback from co-authors.*

Paper II: A. Ráth, D. Pjanić, B. Bernhardsson, and F. Tufvesson, "ML-Enabled Outdoor User Positioning in 5G NR Systems via Uplink SRS Channel Estimates", *IEEE International Conference on Communication*, Rome, May 2023, doi: 10.1109/ICC45041.2023.10279249.

► ***Research Contributions:*** *The paper originated from a master's thesis that I supervised. This paper marks an early exploration into commercial data generated by a 5G base station processing uplink (UL) Sounding Reference Signal (SRS) channel estimates, which served as the primary training dataset. The BS handles a time series of SRS measurements that represent the angular delay spectrum of the radio channel in the beam domain. We employed supervised machine learning methods, using the UE's GNSS-defined location in space as the label. Despite several limitations related to the accuracy of the ground truth positioning and the accessibility of the full bandwidth of SRS measurement data, we successfully demonstrated that SRS channel estimates are viable for UE positioning in commercial systems. This study is also one of the first to show that the fingerprinted features of the surrounding environment, captured in UL CSI based on commercial SRS measurements, contain sufficient information for positioning, without relying on traditional spatial parameters like Angle of Arrival (AOA) or Time of Arrival (TOA). In*

addition, this study is an example of wireless sensing for positioning.

► ***Personal Contributions:*** *In addition to my supervisory responsibilities, I developed the original research idea, defined the thesis scope, actively engaged in discussions on ML modeling, and carried out all the measurements and analyses. Obtaining UL SRS channel measurements from a commercial 5G BS presents significant challenges, especially when dealing with large, complex data structures like SRS measurement samples. These measurements, generated at millisecond intervals, are typically confined to the BS's baseband unit for internal use, with external access often restricted due to hardware and software limitations. I modified the BS software and created a framework to facilitate the transfer of data out of the baseband unit. Additionally, I performed all post-processing on the raw SRS data before preparing it for ML processing. Since Andre Ráth was the sole author of his master's thesis, he was granted the role of first author for the paper.*

Paper III: D. Pjanić, A. Sopasakis, A. Reial, F. Tufvesson, "Early-Scheduled Handover Preparation in 5G NR Millimeter-Wave Systems", *IEEE Open Journal of The Communications Society*, vol. 5, pp. 6959 - 6971, Oct. 2024, doi: 10.1109/OJCOMS.2024.3488594.

► ***Research Contributions:*** *This paper represents my first journal manuscript utilizing simulated data from a commercial-grade simulator at Ericsson, as in Paper I. The study primarily aimed to investigate the feasibility of traffic predictions in 5G NR systems within dense cell deployments featuring high-speed users. The handover preparation phase is widely regarded as the most critical part of the handover process. The insights from this research enable the development of a new handover preparation scheme, introducing a novel, user-aware, and proactive approach to handover decision-making in MIMO scenarios that account for user mobility. I designed the test scenarios, conducted all simulations and data analyses, and authored the manuscript, integrating feedback from my co-authors and reviewers. This work culminated in the filing of two patent applications.*

► ***Personal Contributions:*** *I envisioned the research idea, designed the test scenarios, ran all simulations, handled data post-processing and analysis, and wrote the manuscript. I explored various machine learning models and techniques to achieve the final results. Additionally, I integrated feedback from my co-authors and reviewers.*

Paper IV: D. Pjanić, K. E. Arslantürk, X. Cai, F. Tufvesson, "Dynamic User grouping based on Location and Heading in 5G NR System", *IEEE*

Vehicular Technology Conference (VTC), Oct. 2024, Washington DC, USA, doi: 10.1109/VTC2024-Fall63153.2024.10757679.

► ***Research Contributions:*** *The positioning results from Paper II appeared unsatisfactory, intuitively due to the limited bandwidth resolution of the SRS measurements and the low accuracy of the GNSS data used as the positioning ground truth. It seemed that positioning estimation accuracy could be significantly improved. To validate this assumption, I initiated a master's thesis work, where we utilized a highly accurate GNSS device alongside a higher-resolution bandwidth for collecting SRS channel measurements. With the new, highly accurate UE position estimates generated by the combined CNN/FNN ML model, we added a new dimension to our dataset: the heading direction of the users. Using these two dimensions, we applied a clustering model to dynamically group UEs while in connected mode. Furthermore, this study also serves as an example of wireless sensing for positioning and tracking. The work culminated in filing a patent application.*

► ***Personal Contributions:*** *As with Paper II, I conceptualized the original research idea, defined the entire scope of the thesis, and performed all measurements and analyses. I modified the BS software to support the generation of high-resolution SRS channel measurements within a commercial setup. I actively suggested the ML models to be deployed and analyzed the results. In addition to my supervisory responsibilities, I created the figures and wrote the paper, incorporating feedback from my co-authors and reviewers.*

Paper V: G. Tian, D. Pjanić, X. Cai, B. Bernhardsson, F. Tufvesson, "Attention-aided Outdoor Localization In Commercial 5G NR Systems", *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 2, Nov. 2024, doi: 10.1109/TMLCN.2024.3490496.

► ***Research Contributions:*** *Building on the promising positioning results from Paper IV, we simultaneously explored whether positioning accuracy could be further improved by utilizing a Transformer architecture. Being highly effective at capturing long-term dependencies and correlations, Transformers are particularly well-suited for time-series applications, such as SRS channel measurements, enabling accurate estimations. Furthermore, at the time we initiated this work, only a few studies had examined Transformer applications in the wireless research field, making our exploration innovative, especially given the commercial setup. This work achieved highly accurate UE positioning, with precision reaching sub-1 meter levels and is a further example of wireless sensing for positioning purposes.*

► ***Personal Contributions:*** *Guoda Tian and I contributed equally to this paper. The decision to employ a new AI-based technology, the attention-driven Transformer for this task was my idea, inspired by ChatGPT's remarkable evolution at the time, driven by its attention mechanism. I had a strong belief that the advancements in generative AI were promising tools. As with Paper III, I defined the test campaigns and routes, conducted all measurements, and performed post-processing of the raw I/Q data collected at the 5G BS before it was conveyed to the generative AI model. This also required modifying the BS software to support the generation of high-resolution SRS channel measurements.*

Paper VI: D. Pjanić, G. Tian, A. Reial, X. Cai, B. Bernhardsson, F. Tufvesson, "Illuminating the Path: Attention-Assisted Beamforming and Predictive Insights in 5G NR Systems", *IEEE Transactions on Vehicular Technology*, submitted May 2025.

► ***Research Contributions:*** *Before embarking on my PhD journey, as a baseband software developer implementing beamforming functionality, I often encountered challenges posed by the current models and algorithms used to manage beamforming procedures. These methods were both time- and energy-intensive, jeopardizing the efficiency of beamforming, particularly in scenarios with many UEs in the system, where computational complexity scaled linearly with the number of antennas in the MIMO system. In this final paper, I had the opportunity to explore whether AI could assist a BS in deriving the entire downlink transmission chain within a commercial-grade 5G system. For this purpose, the Transformer architecture, established in Paper V, proved to be a suitable candidate for predicting the strongest beams in a 5G NR system. The predicted downlink beams were specifically designed to address the challenges of NLOS environments, which are characterized by high-dimensional channel dynamics and signal variations caused by scatterers. The presented beam prediction results showcased remarkable robustness, even for long-term prediction horizons extending well beyond the channel coherence time, by leveraging high-dimensional fingerprinted features. This study in many ways overlaps with the concept of wireless sensing for network optimization related to improving coverage, based on real-time environmental data the network can automatically adjust its coverage strategy. The work culminated in filing a patent application.*

► ***Personal Contributions:*** *I utilized the data from Paper V, incorporating*

additional post-processing tailored for this specific study and modified the Transformer model to fit this task. I generated the figures, and wrote the paper.

RELATED WORK

A publication to which I contributed during the course of this thesis work, but which is not included in the thesis itself, is listed below:

CONFERENCE CONTRIBUTIONS

Paper viii: I. Yaman, G. Tian, D. Pjanić, F. Tufvesson, O. Edfors, Z. Zhang, L. Liu, "Adaptive Attention-Based Model for 5G Radio-based Outdoor Localization", *https://doi.org/10.48550/arXiv.2503.23810*, May 2025.

Acronyms and Symbols

Here, important acronyms, abbreviations, and symbols are listed, which are recurring throughout the thesis. Some abbreviations, which only occur in a narrow context, are intentionally omitted; some abbreviations are used in more than one way, but the context is always explicitly clarified in the corresponding text. Some (compound) units are provided with prefixes to reflect the most commonly encountered notations in the literature.

ACRONYMS AND ABBREVIATIONS

3G	Third Generation
3GPP	3rd Generation Partnership Project
4G	Forth Generation
5G	Fifth Generation
6G	Sixth Generation
AAS	Advanced Antenna System
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AOA	Angle of Arrival
AOD	Angle of Departure

CIR	Channel Impulse Response
CN	Core Network
CNN	Convolutional Neural Network
CSI	Channel State Information
CSI-RS	Channel State Information Reference Signal
CTF	Channel Transfer Function
DL	Deep Learning
DNN	Deep Neural Network
DOA	Direction of Arrival
DOD	Direction of Departure
FDD	Frequency-Division Duplex
FNN	Feedforward Neural Network
FT	Fourier Transform
gNB	gNodeB
GNSS	Global Navigation Satellite Systems
GOB	Grid of Beams
GPS	Global Positioning System
GSCM	Geometry-Based Stochastic Channel Model
LOS	Line-of-Sight
MIMO	Multiple-Input Multiple-Output
ML	Machine Learning
MPC	Multipath Component

MU-MIMO	Multi User Multiple-Input Multiple-Output
NLOS	Non Line-of-Sight
NLP	Natural Language Processing
NN	Neural Network
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
RAN	Radio Access Network
RL	Reinforcement Learning
RLC	Radio Link Controller
RNN	Recurrent Neural Networks
RRC	Radio Resource Control
RSS	Received Signal Strength
SRS	Sounding Reference Signal
SU-MIMO	Single User Multiple-Input Multiple-Output
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
TDD	Time-Division Duplex
TDOA	Time-Difference of Arrival
TOA	Time of Arrival
UE	User Equipment
VR	Virtual Reality

XR Extended Reality

INTRODUCTION

1

Background and Overview of the Research Field

"When wireless is perfectly applied the whole earth will be converted into a huge brain, which in fact it is, all things being particles of a real and rhythmic whole. We shall be able to communicate with one another instantly, irrespective of distance. Not only this, but through television and telephony we shall see and hear one another as perfectly as though we were face to face, despite intervening distances of thousands of miles; and the instruments through which we shall be able to do this will be amazingly simple compared with our present telephone. A man will be able to carry one in his vest pocket."

Nikola Tesla in 1926

The purpose of this chapter is to describe the evolution of cellular networks toward 5G and beyond. Years of research went into perfecting the enhanced cellular broadband experience and 5G was the latest technology, with 6G being standardized, at the time when this thesis was being written. A 5G system enables user connectivity through ultra-reliable and low-latency communications and the research presented in this thesis addresses 5G from a perspective of RAN and many-antenna technology such as MIMO systems. As the fundamental tool for addressing the research questions presented later in this thesis, ML and AI are examined and discussed.

Note: The terms user equipment *UE*, *device* and *user* are used interchangeably in the following sections to refer to a mobile phone or wireless device. In addition, the relatively recent shift in terminology from ML to AI has gained momentum in recent years, primarily due to the remarkable success of generative AI models such as chatGPT (2020) [1], DALL·E (2021–2023) [2], and

AlphaFold (2021) [3]. These models show advanced capabilities, including human-like text generation, AI-driven image synthesis from text descriptions, and solving the long-standing protein-folding problem, which had remained an open challenge in biology for more than 50 years. This thesis explores and demonstrates the capabilities of generative AI models in the field of wireless communications, and therefore, the terms ML and AI are used interchangeably, as ML is a subfield within the broader domain of AI.

1.1 EVOLUTION OF CELLULAR NETWORKS

Cellular networks have been continuously evolving since their introduction around the 1980s. In general, approximately every 10 years, a technology shift towards a new generation has been introduced. Each generation of cellular networks has advanced communication technology, focusing on speed and capacity. The first two generations of cellular networks opted for an analogue and later a digital co-called circuit-switched telephony. This technology shares the same technical core as landline telephony, besides the wireless communication between the user and the network. 1G, introduced in the 1980s [4], was the first generation of mobile networks, offering analog voice communication. However, it had limited capacity. In the 1990s, 2G [5–8] brought digital voice and SMS, improved voice quality, and first access to internet via packet-switched data transmission [9]. With the introduction of the 3rd and 4th generation of cellular networks, the focus shifted towards mobile broadband connectivity. The 2000s saw the rise of 3G [10–12], which enabled mobile internet and multimedia services such as video calls. We became increasingly interconnected with others through the internet, even outside our homes. With 4G in the 2010s [13–15], high-speed broadband became a reality. The 5G, specified in [16,17] and described in detail in [18,19], expanded the scope of use cases beyond the mobile broadband objectives initially defined in 4G, primarily by leveraging NR technologies such as MIMO systems, which facilitate enhanced data rates.

Based on the assessment of the continuously growing traffic in cellular networks, the targeted 5G user scenarios were significantly broader than those of earlier generations of cellular networks. Before 5G, the main focus of cellular communications was human-centric, from telephony to mobile broadband services. To meet the increasing demands of increased data traffic triggered by new services such as 3D video, holographic-type communications [20–22] or Augmented and Virtual Reality (AR/VR) using the umbrella term Extended Reality (XR), an enhanced mobile broadband was envisioned beyond 5G. Figure 1.1 illustrates the progression of cellular network generations and the anticipated future advances, as envisioned when this thesis was composed.

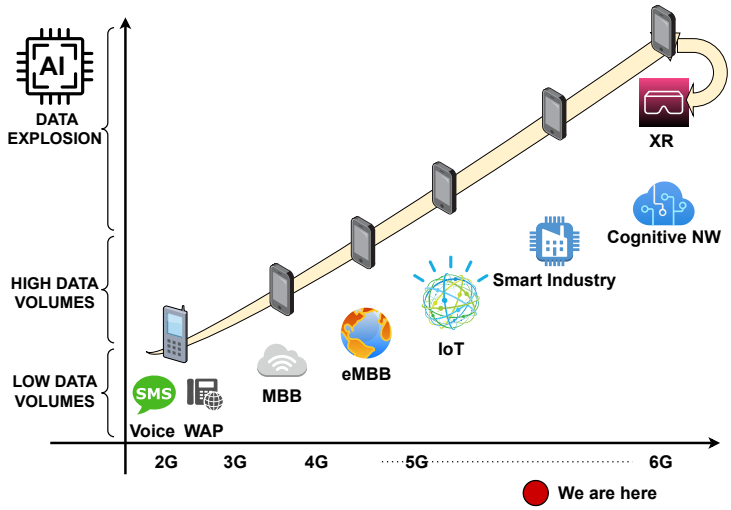


Figure 1.1: The progression of cellular networks and future prospects envisioned by Ericsson.

1.2 CELLULAR SYSTEM ARCHITECTURE

A cellular 5G network consists of a core and a *Radio Access Network* (RAN). The RAN domain is responsible for all radio-related functionality of the cellular network including radio resource handling, transmission protocols, channel coding, scheduling, and different multi-antenna schemes.

1.2.1 CORE NETWORK DOMAIN

The *Core Network* (CN) is responsible for functions complementing the functionality included in the radio access domain such as subscriber setup, authentication, end-to-end connections etc. This architectural separation between the two entities is driven by the fact that one 5G core network may serve multiple 5G radio access networks combined with other radio access technologies such as 2G, 3G and 4G [23]. The architecture of cellular networks comprises numerous sub-functions that are thoroughly detailed and standardized by the 3rd Generation Partnership Project (3GPP), primarily in [24], [25], and [26]. In general, data flows in a cellular network are divided into control-plane and user-plane. Control-plane is designed to ensure reliable communication and manages signaling and network control functions, enabling seamless communication, mobility, authentication, authorization and service continuity. The user-plane data is responsible for handling the actual data traffic between the users and the network like handling packet routing and forwarding user data

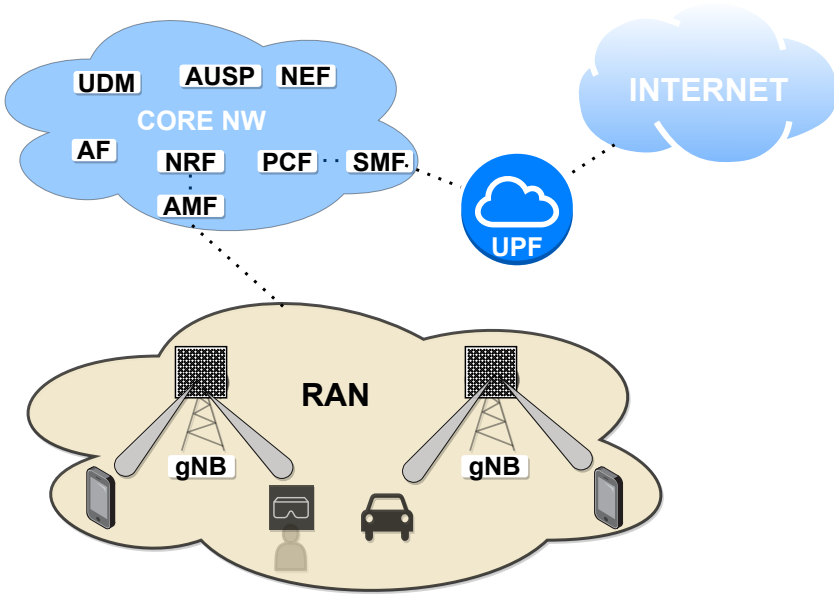


Figure 1.2: The high-level architecture of a cellular network consists of the core network and radio access network. The radio node, gNB, serves various wireless devices through MIMO antenna systems over the air interface. This thesis focuses on optimizing radio communication between the gNB and users, illustrated at the bottom of the figure.

packets through different network nodes (e.g., internet traffic, voice over IP, video streaming).

Due to the complexity of cellular network architecture, only the most significant functions are highlighted in this section as illustrated by Fig. 1.2. For a more comprehensive understanding, users are encouraged to refer to the above-mentioned 3GPP specifications for further details. User-plane functions in a 5G network primarily involve the User Plane Function (UPF), which acts as a gateway to external data networks such as the internet. The control-plane functions include the Session Management Function (SMF), which is responsible for managing sessions and handling IP address allocation for devices. Additionally, the access and Mobility Management Function (AMF) manages control signaling between the core network and the user. This includes key functions such as authentication, security, and idle-state mobility management. A crucial distinction in 5G architecture is between the non-Access Stratum (NAS) and Access Stratum (AS). NAS handles direct communication protocols between UEs and the core network, ensuring

session and mobility management, while AS operates between UEs and the RAN, managing lower-layer functionalities such as radio resource control and transmission scheduling. This layered approach ensures an efficient separation of control and user traffic.

1.2.2 RADIO ACCESS NETWORK DOMAIN

In cellular networks, the RAN domain serves as a bridge between the UE, such as smartphones, and the CN. A main component of a radio access network in 5G is a radio node called the *gNodeB* (gNB). The gNB is responsible for radio-related functions in one or multiple cells such as radio resource management, admission control, connections establishment etc. It is a common implementation that a single gNB covers a three-sector site where a base station handles transmissions in three cells thus, a base station is a possible implementation of, but not necessarily the same as gNB. A gNB incorporates a purpose-built baseband hardware designed to handle the intensive signal processing and computational tasks required in cellular networks, ensuring efficient communication between user equipment UE and the network. This hardware is optimized for high performance, low latency, and energy efficiency, enabling seamless operation across different protocol layers in the network stack. As

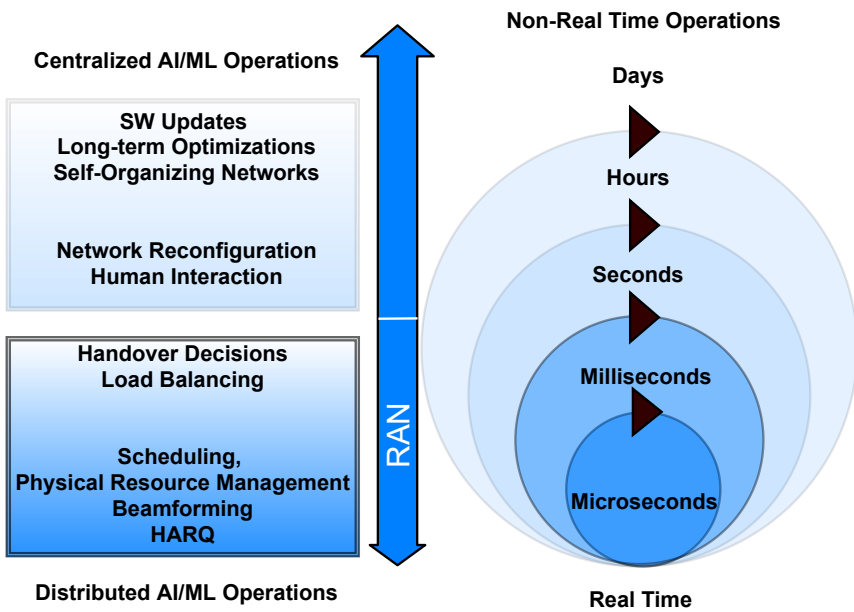


Figure 1.3: Multiple time scale operations of cellular networks.

the wireless connection, or air interface, specifies how data is formatted, transmitted, and received over Orthogonal Frequency Division Multiplexing (OFDM), MIMO, and control channels, the RAN domain operates on multiple time scales to manage communication effectively. These time scales are vital for ensuring efficient resource allocation, synchronization, and meeting the performance demands of diverse applications, all while dynamically adapting to user needs and environmental conditions. Synchronizing operations on these scales is critical for the advancement of modern 5G and future 6G networks. The various time scales in cellular networks are illustrated in Fig. 1.3, highlighting that the RAN is responsible for the most critical aspect: real-time signal processing. The focus of this thesis was on the RAN domain, as it forms the most fundamental part of cellular networks. Its wireless nature introduces the largest technical challenges, making it a pivotal area for research and innovation.

1.3 AI IN CELLULAR NETWORKS: TRENDS AND FUTURE OUTLOOK

Cellular networks have traditionally operated as closed, self-sufficient systems, designed to provide users with wireless connectivity without the need for interaction with other technologies. However, with the advent of late 2G developments, cellular networks began to integrate packet-switched data, enabling the delivery of services over IP-based traffic through the internet. This marked a significant shift, paving the way for more interconnected and versatile communication systems. However, user needs have generally been met through existing software and hardware technological solutions. These solutions were built on robust mathematical and statistical models, which have evolved alongside cellular networks to address the growing demands for higher data transmission capacity. AI is transforming industries across the globe, and telecom is no exception, on the contrary, it has the potential to drive this transformation with other industries being a generator of one of the largest datasets in the industry. The rise of 5G technology has faced new capacity demands, driven by the vast number of devices connected through cellular networks worldwide. These devices generate an immense volume of data including images, videos, and text transmitted over the *air interface* of the RAN domain. Notably, this data is inherently structured in a format optimized for computational processing. This is also the starting point for the rise of the ML and AI solutions being explored within the field of telecommunication.

Future cellular networks will need to interact not only with other telecommunication systems, such as satellite-based systems [27, 28], but also with systems that are not necessarily telecommunication-based, yet connected to

the internet through a multitude of mobile devices. This will necessitate a broader understanding of users' contexts about their "ecosystem", extending beyond the basic communication between a user and the mobile network over the air. Understanding all these contexts will be an overwhelming task that traditional methods for controlling data traffic will no longer be able to handle. These will be especially difficult for a human to comprehend.

There is no doubt, that the greatest technical challenges will lie within the radio access domain, the communication that occurs over the air interface, i.e., between users and the network. Due to its wireless nature, which already involves numerous challenges in the physical layer, radio communication will need to meet new demands for latency, but, above all, for the data capacity required to support all the connected, data-hungry devices that will increase exponentially. The vision at the time of writing this dissertation revolves around a world where the physical and virtual worlds converge and merge into one, through various digital avatar-like solutions [29–31]. Everything in the real, physical world as we know it today will have a counterpart in the digital world [32]. *Metaverse*, and its vision studied in detail by [33–37], is a collective term for a virtual space that merges physical reality with digital environments, allowing people to interact, work, play, and socialize through 3D avatars and immersive XR technologies like VR and AR. Deep Learning is expected to be one of the key technological enablers of 5G Advanced and 6G by offering a new paradigm for the design and optimization of networks with a high level of intelligence [38]. DL has already proven valuable in tough wireless communication problems, especially when it's hard to model the system or when the model's complexity makes practical solutions difficult [39]. The radio access domain of the network already operates within very short time intervals, such as milliseconds and microseconds (Fig. 1.3), and to manage these enormous amounts of data in short periods, ML/AI technologies will be key players in handling the task [40]. To enable the convergence of the physical and digital worlds, new and complementary support technologies will be essential to aid decision-making, provide support, and ease the load on the RAN domain. Hybrid approaches like cloud-based solutions are expected to integrate AI/ML technologies to enhance the RAN. These approaches can harness the computational capabilities of the cloud to process large-scale data efficiently and offload tasks from native AI deployments in the RAN domain of cellular networks.

1.4 RESEARCH QUESTIONS

The numerous antennas in a modern MIMO system, massive MIMO, initially proposed by Thomas Marzetta [41], provide unprecedented insights

into wireless channel behavior offering a unique prospect within wireless communication according to [42, 43]. By examining real channel behavior, including spatial fingerprints and their evolution over time, we gain access to details not observable in the pre-MIMO era. As described by Maxwell's equations, electromagnetic waves interact with their environment as they propagate through media like air. These interactions depend on factors such as wavelength, electrical conductivity, and the geometry of surrounding objects. The majority of signal power is scattered along multiple directions or paths, creating complex patterns as waves interact with various objects in the environment. Instead of analyzing individual propagation paths, an alternative approach involves examining sequences of received wireless signals associated with specific physical locations. By capturing a time series of measurements linked to these locations, fingerprinted patterns emerge, which ML/AI models can interpret. This approach avoids the need for complex and time-intensive calculations of physical channel parameters or estimating individual signal paths. Numerous studies have demonstrated that the theoretical predictions regarding spectral and energy efficiency were, to a large extent, achievable in practice [44–46].

ML has a long history but has gained significant attention over the last decade due to the increased availability of data and computational power. When this thesis was initiated, the application of ML to the lower, physical layers of radio networks remained relatively unexplored, with limited studies on the topic especially those leveraging data from commercial 5G systems. However, a few overview papers had examined related areas. For instance, [47] analyzed the prediction of CSI in LTE systems using LTE channel condition maps, demonstrating promising results in forecasting channel conditions at specific physical locations based on historical data. [48] employed Support Vector Machines [49] and Gaussian Processes [50] to predict received signal strength over a long-term horizon. Meanwhile, [51] investigated the challenge of predicting wireless channel features that are not directly observable at a Base Station (BS), using machine learning techniques driven by large-scale channel data. Additionally, [52] analyzed traffic patterns in real 4G networks, implementing a Gaussian process-based predictor to model traffic variations. The study demonstrated that wireless traffic prediction could effectively reduce uncertainties in network demand and supply.

This thesis investigates whether and how ML/AI can optimize the RAN domain of cellular networks by leveraging fingerprinted radio features derived from 5G MIMO systems, which are known for their complexity even with reasonably sized antenna arrays. The research focuses on various aspects of network optimizations via massive MIMO antennas,

targeting lower frequency bands below 6 GHz and higher frequency bands around 30 GHz. These commercially deployed bands exhibit distinct and often contrasting propagation characteristics when interacting with their surrounding environments. The primary focus of this thesis pivots around addressing the following key questions and the proposed methods to tackle them:

- ***Q1****: Is it possible to achieve long and short-term channel predictions utilizing historical channel measurements?*
  Method: Analysis of long-term channel behavior to improve channel gain prediction based on historical channel estimates from the uplink pilot signals.
- ***Q2****: How to help mobility and traffic pattern estimation?*
  Method: Analysis of uplink and/or downlink channel estimates and the large-scale parameters.
- ***Q3****: Are handover predictions feasible?*
  Method: Analysis of whether uplink and/or downlink channel measurements can enable new prediction methods.
- ***Q4****: How to achieve physical and virtual UE positioning?*
  Method: Analysis of uplink and/or downlink channel estimates for user positioning.

1.5 RESEARCH BOUNDARIES AND PRACTICAL RESTRAINTS

Telecommunication systems of commercial-grade are the product of extensive academic and industrial research, and in most cases represent a compromise between theoretical performance and practical limitations. This includes not only technical constraints but also economic viability and product feasibility. They represent a compromise between theoretical optimality and practical constraints like: hardware limitations, deployment costs, spectrum availability, and energy efficiency. Economic factors such as cost of infrastructure, user equipment, and market dynamics play a major role in which technologies have to persist. Consequently, certain technological solutions prevail over others for example, TDD deployments have become more common than FDD in many 5G NR deployments, especially above 2.5 GHz.

While millimeter-Wave (mmWave) frequency bands, like 28 GHz, offers huge bandwidth and high data rates, in practice, sub-6 GHz (e.g., 3.5 GHz) deployments today dominate because mmWave coverage is limited and short-ranged. Due to this, the cost of dense mmWave deployment is higher, which makes operators prioritize mid-band 5G first to balance coverage, cost, and performance. On a related note, four of the included papers, II, IV,

V and VI, utilize measurement data collected from a TDD MIMO system. In these studies, different NN-driven ML/AI models applied fingerprinting techniques to investigate positioning, location-based user clustering, and short and long-term channel prediction tasks.

For Papers I and III, a simulated environment, operating with a 28 GHz setup, was created to generate measurement data from a MIMO system. Paper I investigates user clustering by analyzing measured channel patterns across various traffic deployments in urban environments.

Paper III proposes a novel approach to optimize 5G handover procedures using the historical MIMO measurement data described above. A neural network-based model predicts incoming users by analyzing MIMO measurement patterns along a predefined urban-inspired trajectory.

2

Fundamentals of Wireless Communications

This chapter provides a high-level overview of the technical challenges faced by wireless communication systems and addresses these challenges at a fundamental level.

2.1 BASICS OF ELECTROMAGNETIC WAVE THEORY

Wireless communication is fundamentally based on the presence and utilization of electromagnetic waves. By generating waves that propagate through space, energy and information can be transmitted from one location to another. The information is encoded by modulating the amplitude, phase, and frequency of these waves. The electromagnetic spectrum, encompassing both naturally occurring and artificially generated waves, spans a vast range of frequencies and wavelengths. In 5G wireless communication systems, frequencies typically range from some gigahertz to tens of gigahertz. A very important characteristic of electromagnetic waves is that they can be added together, or *superimposed*. The interaction of waves propagating in different directions can lead to the formation of standing wave patterns, causing a significant reduction in average field strength in certain regions, a phenomenon known as *fading*, which is discussed in this section and illustrated in Fig. 2.1. The basic concepts are extensively discussed in a book written by professor Andreas Molisch [53], which also serves as a key technical reference for many researchers in the field of wireless communications. Traditionally, fading has posed a challenge to reliable wireless communication. A moving transmitter or receiver will experience rapid time variations of the signal strength due to the fading, often changing completely within a fraction of a second. This

kind of fading is often referred to as fast fading, or small-scale fading, and has always challenged reliable and efficient communication. However, the increasing adoption of directional antennas, discussed in the next chapter, which can differentiate between waves based on their propagation direction, is transforming this challenge into an advantage. Properties of electromagnetic waves can be constructively superimposed to enhance signal transmission and reception.

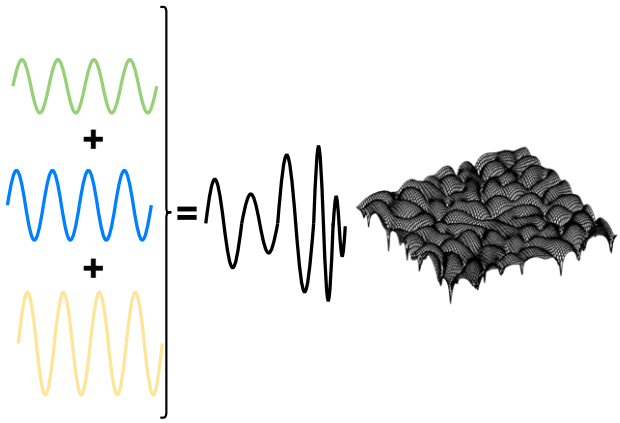


Figure 2.1: Waves can be added together, or superimposed creating standing wave patterns.

2.2 CHANNEL MODELING AND PROPAGATION

When electromagnetic waves travel from a transmitter to a receiver, they do not always follow a single direct path. Instead, they propagate along multiple distinct routes, referred to as Multipath Components (MPCs). This phenomenon, known as multipath propagation, arises because signals can reflect off surfaces, diffract around obstacles, and scatter in different directions before reaching the receiver. In some cases, there may be a direct LOS connection, but often the received signal consists of a combination of multiple indirect paths. Each propagation path has unique characteristics, including signal amplitude, propagation delay (travel time), Direction of Departure (DoD) from the transmitter, and Direction of Arrival (DoA) at the receiver. One crucial aspect of multipath propagation is that signals traveling along different paths undergo varying phase shifts, which depend on the distances they have covered. These phase shifts lead to constructive or destructive interference,

meaning that the total received signal strength fluctuates dynamically as the transmitter, receiver, or surrounding objects move. Since a conventional receiver cannot differentiate between individual MPCs, it simply adds up, creating a signal interaction called *interference*. Depending on how the phases of the arriving signals align, interference can either amplify (constructive) or weaken (destructive) the received signal. As a result, the total signal strength varies over time. These fluctuations occur across time, space, and frequency and also affect signal polarization—the orientation of the electric field. The larger the range of directions in which propagation paths occur, *Angular spread*, the more rapid the variations are. Angular spread describes the range of directions from which signals arrive. In environments where signals arrive from many different angles, fading patterns exhibit short-distance fluctuations, with peaks and nulls often occurring within distances of half a wavelength. Conversely, at a BS, which is typically positioned at an elevated height for wider coverage, the angular spread is more limited. As a result, the fading pattern is smoother, and the distance between peaks and nulls becomes larger, as illustrated in Fig. 2.2. The multipath and angular

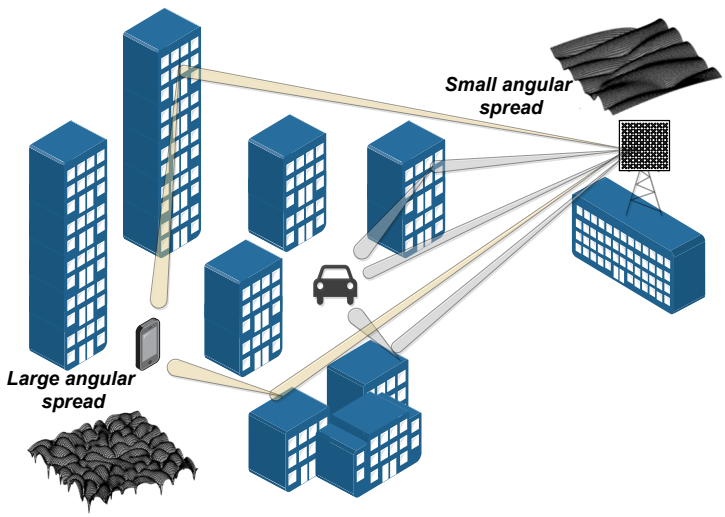


Figure 2.2: Multipath propagation. The receiver sees multiple copies of the original transmitted signal in short succession with differences in arrival time and power.

spread strongly impacts how antenna directivity can be used and whether simple beam shapes are adequate or more irregular radiation patterns are needed to optimize the communication. A multi-antenna system includes methods and algorithms for acquiring information about complex and rapidly

changing multipath conditions and using the degrees of freedom of large antennas with dual-polarized elements. This way, they phase-shift different MPCs such that they add constructively and enhance communication quality. The fading variations are not only confined to time and space. Multi-antenna techniques started as methods to mitigate the fading by adding diversity, but have subsequently evolved to more directly take advantage of the multipath channel by using the different propagation paths and polarizations as separate and parallel communication channels through MIMO schemes.

2.2.1 PHYSICAL CHANNEL MODELING

The physical channel refers to the medium through which data and control signals are transmitted over the air interface. Modeling the physical channel is essential for understanding signal propagation, interference, and system performance in real-world deployments. Unlike previous generations, 5G operates across a wide frequency range (sub-6 GHz and mmWave bands) and supports advanced features like massive MIMO, beamforming, and ultra-dense networks. In summary, while the RAN air interface specifies how data is formatted, transmitted, and received over OFDM, MIMO, and control channels, the physical-layer channel models describe the statistical or deterministic behavior of the propagation environment. Together, they form the foundation for evaluating and optimizing radio-access performance from link-level algorithms all the way up to system-level network deployments. 3GPP has defined various standardized models for 5G NR to ensure accurate simulations and testing. Here, a few are mentioned to provide the reader with a basic understanding of the theoretical field.

- Geometry-Based Stochastic Channel Models (GSCMs): GSCMs combine geometry and stochastic elements to model wireless channels and represent scatterers and MPC explicitly, rather than relying purely on statistical models [54, 55]. These models provide realistic spatial characteristics, such as angular spread, Doppler shift, and delay dispersion.

- Cluster-Based Geometry Model: Is a subclass of GSCMs that represents MPCs as being grouped into clusters and used for both sub-6 GHz and mmWave frequencies. The channel is modeled as clusters of MPCs, each containing multiple rays and includes both LOS and NLOS components. The most significant feature of this widely used model structure is that it supports spatial consistency, which means that the channel changes smoothly over time as the user moves. In accordance with the guidelines outlined in [56], Papers I and III employed a

simulated system based on the specifications provided therein.

- Ray-Tracing Models: These models is used in high-precision simulations for urban environments and it computes exact reflections, diffractions, and scattering based on environment geometry. It is particularly important for mmWave frequencies where blockage effects are significant, as discussed in detail in [57, 58]. Today, both open-source [59] and commercial ray-tracing simulation tools are available [60], with the former frequently referenced in academic research for realistic wireless propagation studies. Major telecommunications equipment providers, such as Ericsson and Nokia, also possess their in-house simulation software.

3

Introduction to Massive MIMO systems

Multi-antenna transmission is a fundamental feature of 5G NR and this chapter provides an overview of multi-antenna transmission in general. The use of multiple antennas at the receiver and/or transmitter has been a key technology in cellular networks for over two decades, despite antenna arrays first being introduced during World War II. Understanding MIMO systems is a challenging task, as it is an interdisciplinary field that encompasses communication and information theory, signal processing, propagation channel modeling, and antenna system design. In the following sections, the fundamental principles of MIMO-based communications are introduced, focusing on their core concepts without delving into the complexities of each contributing discipline. In academic research, massive MIMO is typically defined based on theoretical principles, emphasizing asymptotic system properties where the number of base station antennas significantly exceeds the number of served users. Academic studies often assume idealized conditions, such as favorable propagation environments and channel hardening, to explore fundamental limits and potential gains. In contrast, the wireless industry adopts a more practical definition, focusing on deployable massive MIMO solutions that account for real-world constraints, including hardware limitations, power consumption, and deployment complexity. This section primarily addresses massive MIMO from an industrial perspective, emphasizing practical implementations, challenges and solutions.

3.1 ANTENNA ARRAYS AS A KEY ENABLER OF MIMO TECHNOLOGY

As mentioned in chapter 2, wireless communication is fundamentally based on the transmission and reception of radio waves using antennas which serve as the tool of transmitting electromagnetic waves from one location and capturing a portion of these waves at the receiving end. Once the wave has been generated it will continue to propagate and radiate into the surrounding space. The effectiveness with which antennas can radiate and capture waves is fundamental to the wireless communication quality. All antennas are directive to some extent, meaning that they are more effective in radiating waves in certain directions and with certain polarizations. Antennas are usually reciprocal, which means that an antenna has the same radiation pattern when receiving as when transmitting. Pairs of antennas with orthogonal polarizations are commonly used in wireless communications to transmit or receive waves with arbitrary polarizations. Antennas that are small in relation to the wavelength usually have low directivity and therefore spread their radiated energy in many directions. Antennas that are large with respect to the wavelength have more freedom in shaping the radiation pattern, such as creating high directivity by focusing the transmitted energy in a narrow range of directions. High directivity is very useful for improving the communication quality but also presents a challenge since the transmission and reception directions need to be carefully aligned. As radio waves travel between a transmitter and a receiver, they experience path loss, which increases with both distance and frequency which is particularly pronounced at higher frequencies making the use of directive antennas even more crucial. In cellular networks, where mobile devices are scattered in different directions relative to the base station and have random orientations, fixed directional antennas are impractical. Instead, antenna arrays composed of numerous small antenna elements enable highly directive transmission by adjusting the phase, and potentially the amplitude of signals applied to each element.

With the introduction of massive MIMO these capabilities can be dynamically adapted to the spatial distribution of multiple mobile users, traffic patterns, and the diverse propagation paths between transmitters and receivers. Massive MIMO is a concept in which multi-antenna techniques leverage a large number of antennas to enable dynamically adaptable input and/or output signals. Additionally, a massive MIMO solution refers to the practical implementation of this concept, encompassing both hardware components (such as massive MIMO radio units) and software elements (such as massive MIMO signal processing features and beamforming algorithms). Massive MIMO generally uses planar antenna arrays of dual-polarized element pairs divided into subarrays as depicted in Fig. 3.1 (a). Orthogonal polarizations (like ±45 degrees, or 0/90 degrees), systems can significantly improve spectral

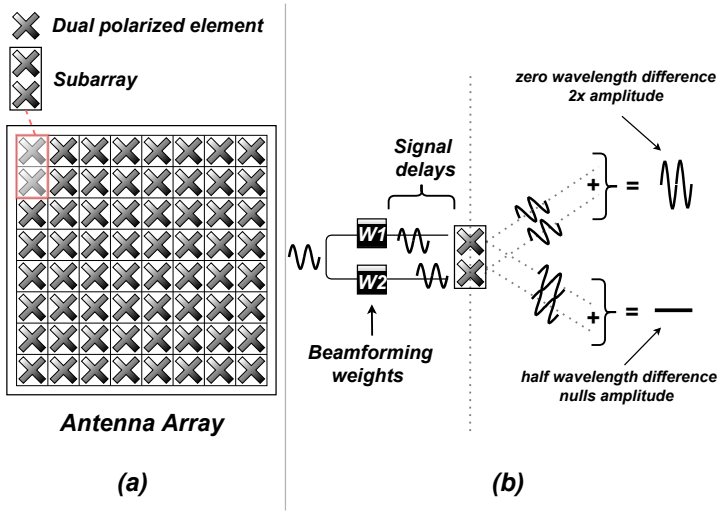


Figure 3.1: Antenna array (a) equipped with cross-polarized antenna elements. Beamforming weights control the relative phase shifts between antenna elements, which enables that transmitted waves combine constructively at the intended receiver while minimizing interference in other directions (b).

efficiency [61]. This technique effectively allows for the transmission of multiple layers of information over the same channel. In NLOS conditions, the isolation between the two polarizations can be especially advantageous. Unlike traditional single-antenna systems, these antenna arrays are a fundamental enabler of 5G and future 6G wireless networks and allow precise radio energy steering through beamforming, which enhances data rates and extends transmission range by directing power toward the intended receivers. By transmitting copies of the same signal and appropriately adjusting the different amplitudes and delays from all the elements it is possible to control the radiation pattern. By dynamically adjusting the beamforming weights, the transmission beam can be directed toward a specific user or spatial region, enhancing signal strength and reception quality. In contrast, destructive interference can be utilized to suppress unwanted signals, thus improving overall network efficiency and minimizing interference between users. This process is illustrated in Fig. 3.1 (b), assuming a free-space channel and the simplest case with two antennas. The electromagnetic field generated by an antenna array results from the superposition or summation of contributions from individual antenna elements. This means that a beam can be formed by transmitting the same signal from multiple antennas. Since signal propagation takes time, the receiver receives multiple delayed versions of the transmitted

signals. These signals combine in the air and embody a summed signal at the receiver side. If all signals arrive at the receiver with identical phase, they add up constructively, resulting in maximum signal strength and gain. Conversely, if the signals are entirely out of phase, they cancel each other out, producing a zero signal and no gain. For intermediate directions, the phase differences cause the signal strength to vary between zero and the maximum gain. In 5G massive MIMO systems, the delays needed are small, and a delay of a signal is equivalent to a phase shift of the signal. A complex number, or weight may therefore represent the amplitude and the delay adjustment of a signal. The set of weights for all the antennas is often collected in a beamforming weight vector, where each element of this vector represents the delay and amplitude of that specific element. Additionally, it is important to note that the array, like any antenna, can also be used for reception. This enables the amplification of desired signals arriving from specific directions while simultaneously suppressing or nulling interfering signals from other directions.

Another significant advantage of multi-antenna deployments is their capability to enhance signal diversity. Simply put, multiple antennas can be deployed at one end of the communication link typically at the BS. In the case of uplink transmission from a single-antenna UE, the BS's multiple antennas can receive multiple versions of the same signal through independent propagation paths, improving the reliability of signal reception in rich scattering environments. This technique is known as receive diversity. Conversely, transmit diversity enhances the reliability of signal reception by leveraging multiple antennas at the transmitter side, typically the BS, to send redundant copies of the signal over different paths. This helps mitigate deep fades and ensures that at least one copy of the transmitted signal reaches the receiver with sufficient quality.

It is worth noting that concepts formerly called Advanced/Active Antenna Systems (AAS) closely resemble or align with massive MIMO, which is a combination of a Massive MIMO radio and a set of Massive MIMO features [62]. While the terms are sometimes used interchangeably, AAS is often seen as an industry-driven term for deployable multi-antenna solutions. In contrast, massive MIMO is more rooted in academic and theoretical discussions.

3.2 SIGNAL MODEL AND PROCESSING

In a MIMO configuration, a base station equipped with N multiple antennas communicates with K users, each having one or more antennas [63]. In general, any linear multi-antenna transmission scheme with N_L layers, represented by the vector $\tilde{x}$, being mapped to N_T transmit antennas, represented

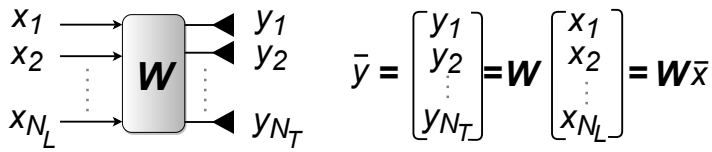


Figure 3.2: General model of MIMO transmission mapping of N_L layers to N_T antennas according to [23].

by the vector $\tilde{y}$, can be modelled as a multiplication with a matrix **W** of size N_T x N_L according to Fig. 3.2. This general model applies to most cases of MIMO transmission. However, the specific implementation of the matrix **W** within the physical transmitter chain can vary, influencing system performance and design considerations [23]. Beamforming, which is the function that maps information signals to multiple antennas, can be implemented in different ways. A key design choice is whether the signal processing should be implemented using analog or digital components, which directly impacts the architecture of the transmitter chain, as illustrated in Fig. 3.3 (a) and (b). The transmitter chain refers to the sequence of components and processing stages responsible for generating, modulating, amplifying, and transmitting the radio signals from the baseband to the antenna elements. In antenna arrays, the transmitter chain must carefully coordinate the signals across multiple antenna elements. The placement of the beamforming matrix **W** in the chain is crucial. For digital beamforming, the matrix is applied at the baseband before Digital-to-Analog Conversion (DAC), while for analog beamforming, the matrix is applied at the RF stage before amplification. The primary disadvantage of

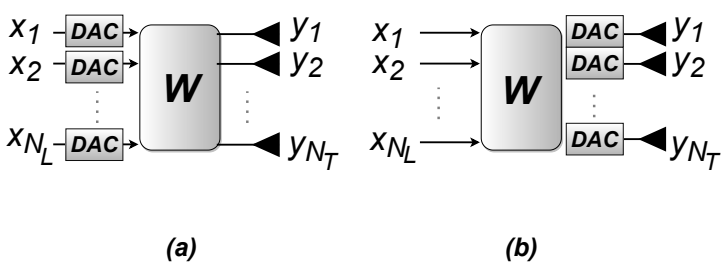


Figure 3.3: Analog (a) vs Digital (b) MIMO processing.

digital beamforming is its high implementation complexity, as it requires a dedicated DAC for each antenna element. This may increases the hardware cost, making it more expensive compared to analog beamforming, where a single RF chain can control multiple antenna elements.

3.3 MASSIVE MIMO TECHNIQUES

Three multi-antenna technology components contribute to the increased performance of massive MIMO, beamforming, nullforming and spatial multiplexing. These technologies apply to both downlink and uplink and are briefly introduced below.

3.3.1 BEAMFORMING

A massive MIMO radio is a hardware unit that includes the antenna array along with a large number of radio transmitter and receiver chains. Massive MIMO features can be implemented in the massive MIMO radio itself, in the baseband unit, or distributed between both, as illustrated in Fig. 3.4. The beam pattern can either be static or dynamic, depending on the deployment scenario. UE-specific beamforming dynamically adapts the beam pattern in both time and frequency, tailoring transmission for each user to optimize signal quality and reduce interference. Static beamforming, in contrast, maintains a fixed beam pattern over time.

- Static beamforming: In cellular networks, a static beam pattern in the elevation domain is common. For instance, a single-column antenna with a fixed beam pattern—characterized by a narrow main lobe in the vertical plane and broad sector coverage in the horizontal plane is often deployed to serve all UEs in a cell. To achieve effective beamforming gains with a fixed beam system, the targeted UEs must be positioned close to the peak of the beam; otherwise, the beamforming gain declines rapidly. In rooftop deployments and large cells, this condition is often well met in the vertical domain for a significant portion of UEs. From the BS's perspective, many UEs are located near the horizon, leading to a relatively small vertical angular spread. However, the spatial distribution of UEs in the horizontal domain is typically much wider and more uniform across the sector. As a result, static beamforming in the horizontal direction is less effective either requiring a wide beam to encompass all UEs, which reduces beamforming gain, or a narrow beam that fails to reach most users.

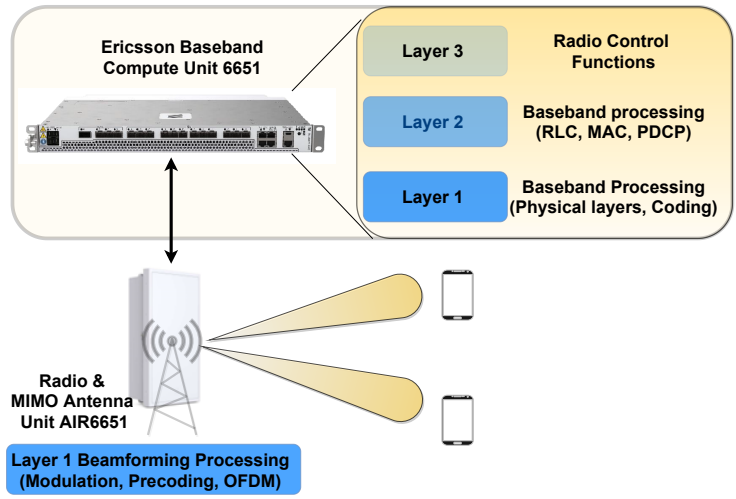


Figure 3.4: Ericsson purpose-built baseband hardware and functionality across protocol layers.

- UE-specific beamforming: A significant advantage of massive MIMO systems is their capability to generate UE-specific beams directed toward individual users. This requires dynamic adaptation of the beam pattern to suit each UE in real-time, allowing for a narrow, and focused beam that can track user movement. Unlike static beamforming, BS ensures that each UE benefits from a dedicated beam with a strong and high-gain peak resulting in higher SINR levels.

3GPP has standardized advanced beam management techniques to dynamically control and direct beams toward users dispersed in different directions [64, 65]. The difficulties of 5G beam management, only outlined in these specifications, are left for those eager to explore, including research studies such as [66–70]. However, due to the limited scope of this work, only the fundamental theory behind beamforming is considered.

3.3.2 SPATIAL MULTIPLEXING

Antenna arrays can be used to transmit multiple streams, or layers, simultaneously with different beam weights on the same time-frequency resource. This is referred to as spatial multiplexing, and the streams of data symbols multiplexed are referred to as layers. There are two basic use cases:

- Single-user MIMO (SU-MIMO), where multiple layers are transmitted to a single user. This requires a multipath propagation channel as well as a receiver with multiple receiver antennas.
- Multi-user MIMO (MU-MIMO) where multiple layers are transmitted to different users in different directions.

3.3.3 NULLFORMING

Nullforming is a beamforming technique aimed at reducing or eliminating signal transmission in certain directions. By carefully shaping the radiation pattern to introduce nulls or low-gain regions in areas where interference-sensitive transceivers are positioned, nullforming helps mitigate unwanted signal interference. This approach is predominantly applied in downlink transmissions to enhance overall network performance.

3.4 CHANNEL STATE INFORMATION IN MIMO SYSTEMS

The network must first acquire channel knowledge to effectively perform beamforming, nullforming, or spatial multiplexing. CSI refers to the knowledge of the properties of the communication channel. It includes parameters like signal strength, amplitude, phase, noise levels, interference, and other metrics that affect the transmission quality. The receiver produces complex-valued estimates per subcarrier for OFDM systems after performing down conversion. CSI can be obtained through various methods, but each comes with a cost, typically in increased signaling overhead. Since the radio channel is a finite resource, different trade-offs are required depending on the network's primary objective, such as enhancing coverage, increasing capacity, or maximizing throughput. A challenge in massive MIMO systems is the availability of CSI. The 3GPP standard defines various sounding and feedback methods [71], but different UEs may support different capabilities and CSI acquisition modes. As a result, the network must support multiple CSI acquisition methods. The process of acquiring CSI differs between Frequency-Division Duplex (FDD) and Time-Division Duplex (TDD) systems. Unlike TDD, where uplink and downlink share the same frequency and reciprocity can be leveraged to estimate CSI, FDD operates on separate frequencies for uplink and downlink. As a result, reciprocity cannot be directly applied, and alternative CSI acquisition methods must be used. These methods are primarily classified into codebook-based approaches, each offering distinct advantages and limitations.

- Codebook-Based Beamforming: The user measures the channel and selects the best beam from a predefined set of beamforming vectors, known as a codebook. This method is robust and effective even in environments with limited channel reciprocity. However, it requires feedback from the user, which introduces signaling overhead. To mitigate this, an implicit feedback mechanism is employed, where a predefined set of candidate beamformer, referred to as a precoder codebook, is specified in the 3GPP standard [71]. The user then recommends the beamformer that best matches the measured downlink channel based on CSI-RS, allowing the base station to apply the most suitable beamforming configuration. Since codebook-based feedback provides the base station with only the dominant downlink channel direction, it is particularly well-suited for SU-MIMO transmission. While it can still be used for MU-MIMO in certain scenarios such as with well-separated users. As a result, FDD systems, which rely on codebook-based beamforming, typically achieve higher efficiency in SU-MIMO compared to MU-MIMO deployments. Also, codebooks are often based on a Grid of Beams (GoB) precoders.

- Reciprocity-Based Beamforming: In TDD systems, channel reciprocity enables the BS to estimate the downlink channel from uplink transmissions, eliminating the need for explicit feedback. Instead of relying on UE-reported measurements, the BS acquires downlink CSI by analyzing the uplink channel, with the UE transmitting a Sounding Reference Signal (SRS). This approach is particularly efficient in dynamic environments but requires precise calibration of transceivers to ensure accurate reciprocity. While the propagation channel itself is reciprocal, the transceiver hardware introduces non-reciprocal effects that must be compensated for. The key advantage of reciprocity-based beamforming is that the full channel information, including small-scale fading characteristics, is available at the transmitter. This provides the base station with highly detailed channel information in both the spatial and frequency domains, allowing for advanced and adaptive beamforming techniques. However, a notable limitation is that the entire downlink bandwidth must be sounded in the uplink for each UE antenna. Given that most UEs are equipped with only one or two transmit chains but have up to four receiver antennas, proper SRS antenna switching functionality is required to ensure that all UE antennas can be sounded. Additionally, each UE must be allocated a dedicated SRS resource, but these resources are limited from an air interface perspective. In scenarios with a high number of connected

users, not all UEs may receive an SRS allocation. Therefore, SRS resources should be prioritized for users who would benefit the most from reciprocity-based beamforming to maximize system efficiency.

CSI can be classified based on its level of detail. It may include short-term (small-scale) channel characteristics or be limited to long-term (large-scale) properties. TDD reciprocity, which captures small-scale fading dynamics, exemplifies short-term CSI, whereas FDD reciprocity primarily provides long-term CSI. While short-term CSI offers more detailed insights into the channel, its reliability depends on a sufficiently strong signal, making it less effective in poor coverage areas. Nevertheless, when coverage conditions permit, short-term CSI enables better system performance.

The 5G mid-band (below 7 GHz, TDD) is a key enabler for 5G deployments, as it meets system performance requirements while allowing for practical antenna array sizes. While both codebook-based and reciprocity-based CSI acquisition methods have been investigated in this thesis, a significant portion of the research has been conducted in a commercial 5G TDD system operating in the mid-band frequency n77 and n78, as specified in [72], primarily focusing on SU-MIMO scenarios. To fully grasp the interaction between NR physical channels and massive MIMO, relevant 3GPP specifications outline the physical layer, measurement methodologies, and beam management in [73–75].

4

Machine Learning and Artificial Intelligence

"We are drowning in information and starving for knowledge."

Rutherford D. Roger

This chapter provides a brief overview of the most important principles of ML which, as a subset of AI, focuses on learning patterns from data and making predictions. AI is a broader field that aims to create systems that can perform tasks requiring human-like intelligence, such as understanding language, recognizing images, etc.

With the emergence of computers and the information age, the ability to solve statistical problems has grown exponentially in both size and complexity. Vast amounts of data are generated in numerous domains, with telecommunication systems forging one of the biggest data amounts, creating a demand for systems capable of learning from data. Fundamentally, ML and AI technologies are rooted in statistical principles to uncover patterns and relationships within data, with the primary goal of extracting meaningful insights by recognizing patterns, correlations, or underlying structures [76]. It is essentially about understanding "what the data says" with increased emphasis on the use of computers to statistically estimate complex tasks. The author perceives ML and AI as a successful fusion of statistical methods, computational power, and knowledge representation. While many aspects of statistical learning have been established for decades, as thoroughly discussed in [77], the rapid advancements in hardware and computational capacity over the past decades were previously unimaginable, enabling the breakthroughs seen today. Since computation plays a critical role in processing and analyzing

large datasets, much of the modern development in ML and AI has been driven by researchers from disciplines outside traditional statistics, such as computer science and engineering. Writing an introductory section on ML and AI is challenging due to the fast growth and evolution of the field over the years. It is impossible to cover the entire field in substantial depth within the scope of a PhD thesis like this. However, to provide an overview, this section emphasizes the most commonly used methods and algorithms that were explored throughout the research presented in this thesis.

4.1 MACHINE LEARNING BASICS

The main ingredients of ML are tasks, models and features, and how these link to each other is illustrated in Fig. 4.1.

- *Feature*: A collection that has been quantitatively measured from some object/event/domain. Features determine much of the success of an ML application and directly affect the quality of the ML output. Features can be thought of as a kind of measurement that can be performed on any instance. As measurements are often numerical, commonly features are represented as real numbers.

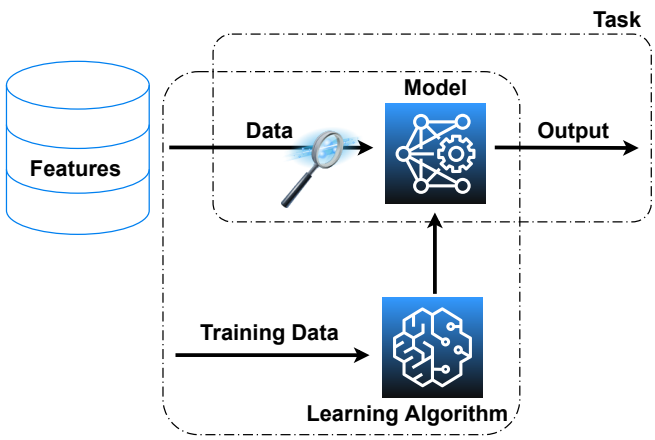


Figure 4.1: General overview of a machine learning model addressing a given task according to [78].

- *Task*: A task is an abstract representation of a problem to be solved.
- *Learning Algorithm*: A machine learning algorithm that is able to learn from data [80].
- *Model*: A model is a central concept and represents what has been learned from data to solve a given task.

  An important distinction is made between tasks and learning problems: *Tasks are addressed by models, whereas learning problems are solved by learning algorithm that produces models*, according to [78].

4.1.1 TYPES OF MACHINE LEARNING

ML uses data and algorithms to mimic human learning, allowing machines to improve over time and enhance their accuracy in making predictions, classifications, or extracting data-driven insights. Figure 4.2 contrasts traditional programming with the ML/AI approach. The key distinction between tradi-

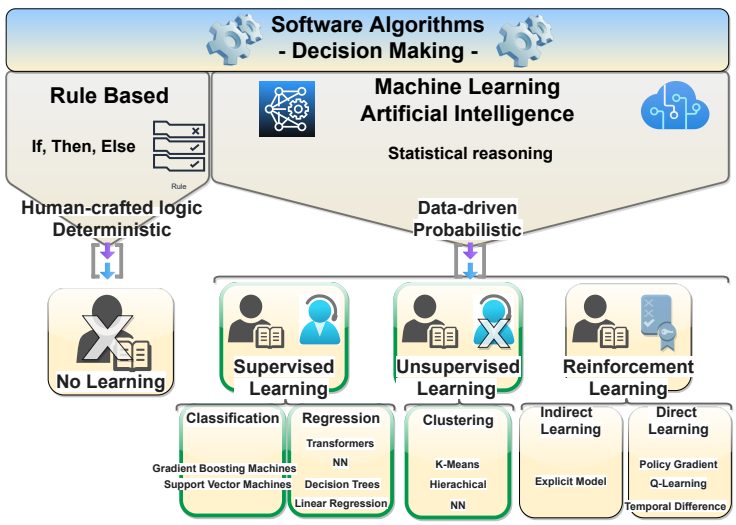


Figure 4.2: General overview of types of learning for machine learning models according to [79]. Machine learning involves showing a large volume of data to a machine so that it can learn and make predictions, find patterns, or classify data. The types of learning explored and utilized in this thesis are highlighted in green.

tional programming and ML lies in their methodologies for problem-solving. Traditional programming requires a programmer to explicitly define rules and instructions that dictate how a computer processes input data to generate the desired output. This approach demands a deep understanding of the problem and a well-defined method to encode the solution in a programming language. Since it relies on predefined logic, the output quality is determined mainly by the programmer's ability to anticipate all possible scenarios. In contrast, ML eliminates the need for explicitly defined rules. Instead, a model is trained on large datasets, allowing it to recognize patterns and relationships that enable decision-making without being explicitly programmed for every possible case [79]. This data-driven approach is especially valuable for solving complex problems where defining explicit rules is impractical or infeasible. The effectiveness of an ML model depends heavily on the quality and quantity of training data, as these factors significantly influence its accuracy and performance. An interesting aspect of these two approaches is the predictability of the generated outcome. In traditional programming, the result is highly predictable if the inputs and the logic are known. For decades, well-established mathematical models of wireless channels have served as the foundation for software development in telecommunications systems, ensuring they meet the necessary requirements for reliable and in a way predictable communication. In the case of ML, the outcomes of predictions or decisions can sometimes be less interpretable, particularly with complex models such as Deep Neural Networks (DNN). However, when trained on well-structured and well-understood data, these models can effectively learn and adapt. This adaptability is essential for the future of cellular networks.

4.1.2 SUPERVISED LEARNING

Supervised learning [81] is a fundamental branch of ML where a model learns from labeled data. In this approach, the training dataset consists of input-output pairs, where each input (feature set) is associated with a corresponding correct output (label or ground truth) that the model aims to iteratively learn, i.e how to map inputs to the proper outputs by minimizing a predefined error function. From a statistical perspective, this means that we have the outcomes for all experiments within the training set. This method typically involves dividing the complete dataset into distinct subsets: a training set for model learning, a validation set to assess the model's generalization ability, and a test set, which is used after training to evaluate the model's accuracy.

4.1.3 UNSUPERVISED LEARNING

This type of learning relies solely on sample data from the problem domain without prior knowledge or predefined labels to guide the learning process

[82]. A model learns patterns, structures, or relationships in data without labeled outputs. Unlike supervised learning, where the model is trained with input-output pairs, unsupervised learning works with unstructured or unlabeled data and finds hidden patterns or clusters within it. Essentially, unsupervised learning has no prior knowledge of the class or value of any sample, it shall infer it automatically. Solving the task involves grouping items based on similarities or shared characteristics, as there is no prior information about predefined classes available.

4.1.4 REINFORCEMENT LEARNING

Reinforcement Learning (RL) [83, 84] has recently demonstrated remarkable capabilities in the field of wireless communications. Unlike supervised learning, RL does not rely on labeled datasets but instead operates through interaction with an environment. It is built on key concepts such as an agent, an environment, and an optimal policy. The learning process is goal-driven, where the agent continuously interacts with the environment, learning through trial and error by maximizing rewards and minimizing penalties. The ultimate objective is to develop an optimal policy that maximizes cumulative rewards over time.

4.2 MACHINE LEARNING ALGORITHMS EXPLORED

In general, ML/AI-based algorithms are computational methods that enable systems to learn from data, recognize patterns, and make decisions with minimal human intervention. They may serve different applications, from predictive analytics to computer vision, speech recognition, and autonomous systems. However, selecting the appropriate ML algorithm for a given task depends on various factors such as the problem type, data availability, and the nature of the decision-making process. Below is a brief summary of the ones explored in this thesis.

4.2.1 CLUSTERING

Humans naturally categorize objects, people, and behaviors into groups based on shared characteristics. Similarly, in ML, technique known as *clustering* is a fundamental unsupervised learning method that organizes unlabeled data into groups based on common, defining features [85]. This approach helps identify inherent structures within the data by grouping similar components jointly without prior knowledge of their classifications unlike supervised learning, where models learn from the labelled data, clustering algorithms operate without predefined labels.

4.2.2 NEURAL NETWORKS

The concept of *Neural Networks* (NNs) was first introduced in [86, 87] as a computational approach aimed at mathematically modeling information processing in biological systems. Neural networks are engineered systems inspired by the workings of the human brain, where interconnected units, or neurons, process information in layers. The term *Deep Learning* (DL) [80] has since evolved beyond its initial neuro-scientific roots to encompass a wide range of advanced learning architectures. Within this framework, Artificial Neural Networks (ANNs) represent a specific type of NN, consisting of artificial neurons organized into multiple layers as illustrated in Fig. 4.3. At each layer, the output of a node depends on inputs from preceding layers, the corresponding weights and biases. The optimization process involves fine-tuning all hyperparameters, specifically the weights and bias terms within each layer, to minimize the losses. The broader NN category includes diverse architectures such as Recurrent Neural Networks (RNNs) [88] and Temporal Convolutional Networks (TCNs) [89], which take inspiration from biological neural functions, but are designed for specialized computational tasks.

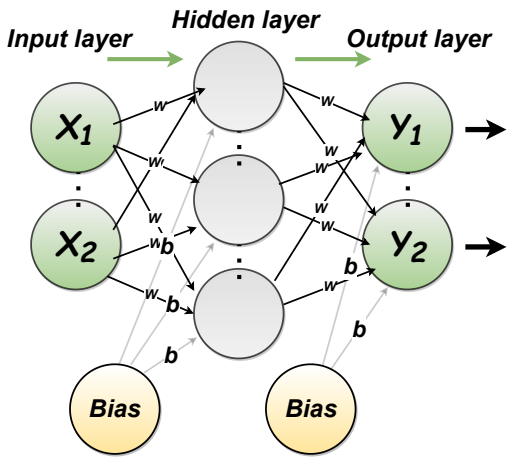


Figure 4.3: Illustration of an ANN. The input, hidden, and output variables are represented by nodes and the weight parameters are represented by links between the nodes (neurons). Bias **b** is equivalent to the intercept in linear models, it is an additional parameter used to adjust the output along with the weighted sum of the inputs to the neuron. Arrows denote the direction of information flow through the network during forward propagation.

4.2.3 TRANSFORMERS

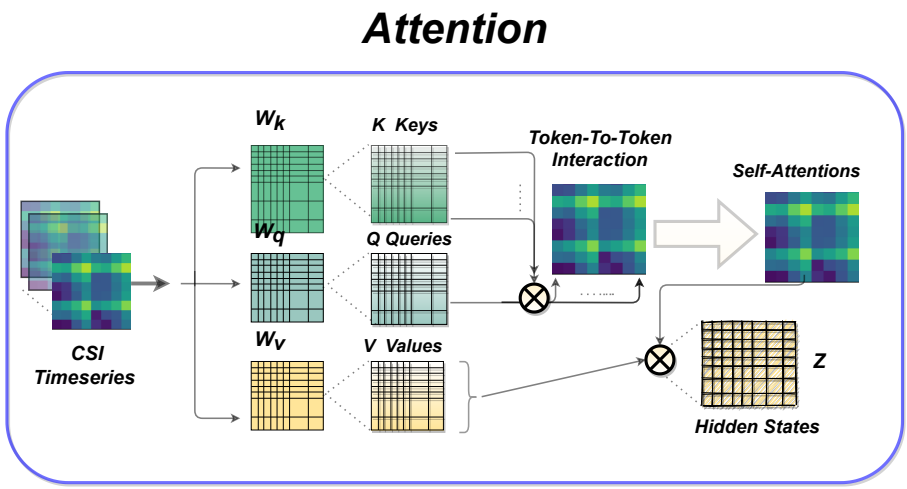


Figure 4.4: Attention mechanism: Instead of processing all input information equally, attention assigns different weights to different elements, enhancing the model's ability to capture dependencies, even across long sequences.

Transformer models, as suggested by [90], have emerged as powerful tools for tackling a wide range of tasks. Originating in Natural Language Processing (NLP), transformers utilize a unique mechanism known as self-attention, see Fig. 4.4, enabling them to capture long-range dependencies more effectively than traditional RNNs. The attention mechanism is a core technique, particularly in deep learning models like *Transformers*, that allows a model to focus on the most relevant parts of input data when making predictions. This makes transformers especially suitable for analyzing long sequences in time-series data, such as CSI measurements influenced by user mobility patterns and surrounding scattering characteristics in wireless environments. Transformers may be classified as AI because they demonstrate capabilities like reasoning, generating human-like text, and answering complex queries in a new way that we did not see before. However, while it is an ML-driven AI system, it is still narrow AI (focused on specific tasks) rather than artificial general intelligence (AGI) which would match human cognitive abilities across all domains.

4.3 WIRELESS CHANNEL AS INPUT TO ML MODELS

A cornerstone for producing high-quality ML and RL models is access to good data. Currently, there are many mechanisms for the generation, collection and analysis of data from the RAN. An example is debug and trace data, which includes detailed information about the internal RAN system state. The latter data is often generated on demand in a real system, which was also the case for the majority of the research publications presented in this thesis. Being the major tool for research presented in this thesis, the input data into the ML/AI model has a pivotal role. It was of the highest significance to understand the measurement data of the wireless channel and its characteristics. The letter, of course, applies in general to any problem to be solved by an ML/AI model as quality, quantity, and relevance of input data are crucial for the performance. High-quality data can lead to better predictions and more effective decision-making processes. Consequently, proper data pre-processing and cleaning techniques are essential to maximize the efficacy of ML applications in the wireless domain. Input data in wireless systems encompasses a diverse range of information, from channel metrics to user behavior, and understanding this data is key to leveraging ML effectively.

4.3.1 CHANNEL TRANSFER FUNCTION

The *Channel Transfer Function* (CTF) is fundamental in wireless communications, describing the frequency-domain characteristics of a channel and how signal components are altered during transmission. Accurate channel knowledge is essential for designing robust wireless communication systems. However, precise CTF estimation is challenging due to the dynamic and complex nature wireless channel, which exhibits variations across frequency and spatial domains. To address this, advanced channel estimation algorithms are required to efficiently process large volumes of data and adapt to highly dynamic environments.

4.3.2 CHANNEL IMPULSE RESPONSE

The *Channel Impulse Response* (CIR) is a function that describes how a wireless channel responds to an impulse signal, which is a very short and high-energy signal, see Fig. 4.5. CIR and CTF are Fourier pairs and are both complex-valued. The CIR captures the amplitude, phase, and delay of the multipath components that reach the receiver after reflecting, refracting, or scattering from the environment. The CIR can be measured by sending a known impulse signal from the transmitter and recording the received signal at the receiver. It is also used to evaluate the quality and capacity of the wireless channel, as well as to design and implement wireless techniques, such as modulation,

coding, and beamforming.
Frequency domain analysis involves transforming the CIR into the frequency

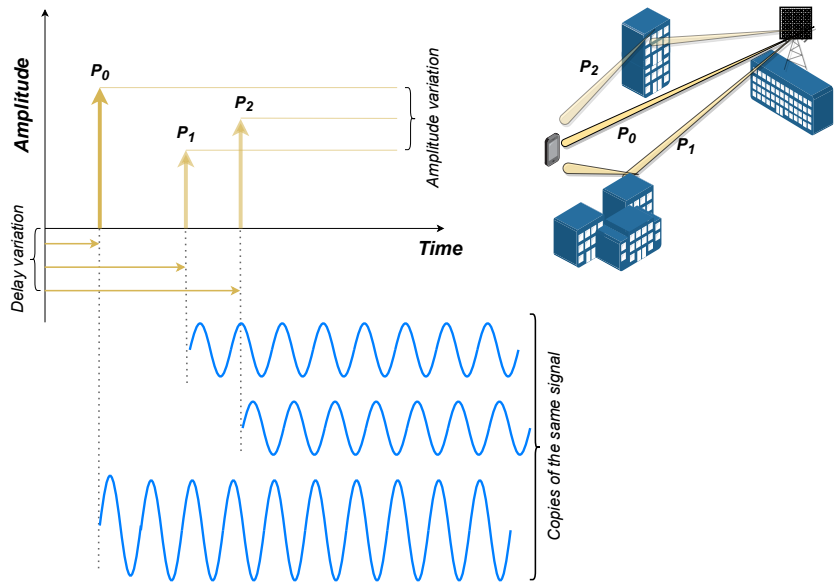


Figure 4.5: The CIR plot represents the different paths the signal takes. The timing or position of each spike on the x-axis shows the delay for that path. The height of the spike on the y-axis indicates the amplitude or strength of that path.

domain by using the *Fourier transform* (FT), which gives the CTF. Statistical analysis involves the use of probabilistic models to describe the distribution and correlation of channel parameters. This can be used to estimate the channel capacity. Using the CIR, the transmitter or receiver can adjust the phase and amplitude of signals from multiple antennas to form a directional beam that maximizes the channel gain and minimizes the channel interference.

4.3.3 FINGERPRINTING

Fingerprinting, within the wireless domain, draws an analogy to scene understanding in other fields. Location fingerprinting refers to techniques that match the specific characteristics of a signal that are dependent on the user's location [91], as shown in Fig. 4.6. In general, two approaches are used for sensing, localization, user tracking, there are two broad strategies in

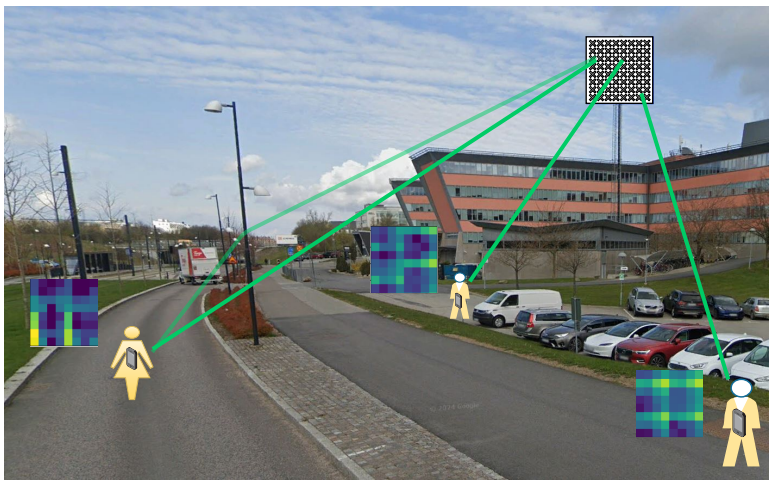


Figure 4.6: The users position significantly affects the location-dependent fingerprints.

wireless systems:

- Parametric approach: Is based on known mathematical models of the radio channel. This approach assumes specific parameters, e.g., AoA, time delay, or Doppler shift. The fingerprinted features can be extracted from the CTF and common examples include *Received Signal Strength* (RSS), the CIR beam matrices etc. The parametric approaches rely on predefined models of the environment and can provide a more structured way to estimate user location based on a model [92–96]. The estimation quality of this approach depends on how accurately these models capture the real-world conditions in the scenario. The localization accuracy may suffer if the model does not align well with the actual conditions hence requires accurate modeling.

- ML approach: This method uses data-driven techniques to learn and recognize environmental signal patterns [97, 98]. Instead of modeling the environment explicitly, a database of "fingerprints" is created by measuring beam-space signals at known locations or conditions. ML-based fingerprinting techniques adapt to more complex, dynamic environments by learning patterns in the data. Using fingerprinting, ML-based approaches can, for instance, predict the location of the

user based on these learned signal characteristics. However, they do come with the challenge of requiring large, diverse datasets and may struggle to generalize to new, unseen conditions. The downside is that performance may degrade as the environment changes, especially if the conditions during deployment differ significantly from those in the training phase.

Note: Throughout the following text, the terms *space* and *domain* are used interchangeably when referring to antennas and beams.

4.3.4 A NOTE ON SRS DATA

As noted in the final section of the Introduction chapter, Papers II, IV, V, and VI use measurement data obtained in a commercial 5G TDD NR system. A shared characteristic across these papers is that the collected SRS measurement data consists of a time series of uplink measurements that have been transformed from the antenna domain into the beam domain. This approach utilizes the data being processed at a later stage of the signal processing chain, once the channel parametric processing is complete within the baseband unit illustrated in Fig. 3.4. When a radio signal is transmitted or received using a MIMO array, the initial representation of the signal is in antenna space, where each antenna element processes the data independently. However, in systems with large antenna arrays, it is often beneficial to transform the signal into the beam space or angular domain. The beam space transformation involves applying a mathematical operation, typically a discrete Fourier or similar spatial transformation. The idea is to combine the signals from different antennas to form beams pointing in specific directions which in LOS scenarios directly translates into the geographical location of users.

In the above-mentioned studies, the fingerprinting approach was applied by collecting beam-space signals at known locations or conditions. Various NN-driven ML/AI models were trained on these fingerprints to recognize patterns and explore UE positioning, location-based user clustering, and both short- and long-term channel prediction tasks. The SRS is primarily used for uplink CSI estimation to support tasks such as beam management, uplink scheduling, and uplink optimization, the angular information captured in the SRS dataset after transformation into the beam domain can also be used for UE localization. A noteworthy aspect is that SRS data is typically transmitted by the UE while in connected mode that is, after an active signaling connection with the BS has been established via the Radio Resource Control procedure [99]. This effectively makes SRS-based positioning a form of active sensing. Moreover, the requirement for an active connection is requirement for the BS to be able to reliably perform user positioning.

When it comes to other signals suitable for positioning applications, there is also DL Positioning Reference Signal (PRS) introduced in [100] to help the NW determine user positions by measuring time differences (like OTDOA) based on signals sent by multiple cells. The network can determine the UE position using different estimation methods, such as precise measurements of ToA, TDoA, and AoA key parameters for location estimation. PRS is transmitted in a structured pattern across multiple frequency and time resources, allowing the UE to measure arrival times from different transmission points. These measurements are then processed to estimate the UE's location with sub-meter accuracy. The flexible allocation of PRS in the time-frequency grid helps minimize interference and improve robustness in multipath environments. In 5G, PRS is particularly important for Ultra-Reliable and Low-Latency applications (URLLC) such as autonomous driving, emergency response, and industrial automation, where accurate location tracking is essential. PRS is specifically designed to deliver the highest possible levels of accuracy, coverage, and interference avoidance and suppression. To design an efficient PRS, special care was taken to give the signal a large delay spread range, since it must be received from potentially distant neighboring base stations for position estimation. A key difference from SRS is that the UE can use PRS for positioning in both idle and connected modes, depending on the specific positioning scenario. Since idle mode consumes less battery, it is more favorable from the UE's perspective compared to the active signaling required for uplink SRS transmissions.

This thesis did not explore the positioning capabilities offered by PRS pilots, which continues to be a promising area for future research, as it relies on measurements from multiple receiving base stations involved in the positioning process to enhance positioning accuracy. As of the writing of this thesis, the majority of the studies have used simulated environments as discussed in [101, 102], and no prior studies have utilized PRS data generated in a commercial 5G NR system. At the time of writing this thesis, most studies still rely on simulated environments, [101, 102], and no prior work has used PRS data obtained from a commercial 5G NR system. However, the findings in this thesis suggest that as RAN continues to evolve, integrating PRS with machine learning and AI-enhanced signal processing will further improve positioning accuracy down to the centimeter level.

Beyond positioning, Paper VI demonstrated that both long-term and short-term DL channel predictions can achieve high accuracy, highlighting the versatility of 5G SRS measurements.

5

Conclusions and Outlook

Managing cellular network complexity has long been an active area of research, and with the advent of 6G deployments featuring terabit-level data rates, XR, holographic communication, IoT, and more this challenge will become even more critical. To meet these demanding requirements, the integration of automation and AI-based solutions is increasingly seen as essential, offering transformative potential to revolutionize cellular network operations. Moreover, the remarkable success of MIMO systems has inspired extensive research into leveraging multi-antenna configurations to better understand radio channel dynamics and their interactions with the environment during signal propagation, a topic I explored in depth during my PhD journey. This chapter concludes the thesis by summarizing its research contributions, speculating on potential future applications, and discussing key challenges and open issues for future research directions.

5.1 RESEARCH CONTRIBUTIONS

While cellular networks have been around for more than half a century, AI technology is still in its early stages. Over the past 2-3 years, AI has shown tremendous progress, with applications like generative chatGPT, which can sometimes be difficult to fully grasp. Cellular networks, previously isolated from external technologies, are now evolving to support integration with innovations like AI and ML [103–105], aligning with a broader vision of openness. This cross-pollination has the potential to elevate cellular wireless technology to remarkable technical heights, revolutionizing the way we communicate, heights we can likely only speculate about today.

Despite its complexity, we demonstrate that the MIMO-enabled RAN domain offers significant potential to help cellular networks become more autonomous and cognitive. This thesis scratches the surface of opportunities, showcasing a few potential optimization approaches by utilizing fingerprinted radio features in measurement data.

5.2 FUTURE PERSPECTIVES OF MIMO AND AI IN CELLULAR NETWORKS

The results presented in this thesis demonstrate that the cross-domain interplay between RAN and AI/ML is not only a successful approach but also an undeniable cornerstone for the future of cellular communications, with MIMO systems providing the perfect "lens" to investigate and advance cellular radio technologies. However, current cellular deployments come with several constraints, and the research on integrating AI/ML in cellular communication system design is still in its infancy, and many key issues are still open. In the following, I discuss several potential directions for future study, based on experiences and constraints that I encountered throughout my research, identifying and highlighting the most noticeable ones, summarized as follows.

5.2.1 DATA COLLECTION AND DATA-PROCESSING ASPECTS

Most of the publications included in this thesis are based on real-life data collected from commercial 5G systems. As the 5G architecture shares many similarities with previous generations of cellular networks, accessing data within the RAN domain remains a highly demanding and time-consuming process. As earlier mentioned, commercial 5G BS impose significant limitations when retrieving data, which is typically confined within the baseband unit of the BS for internal processing. External access to these data is often restricted by hardware and software constraints.

These limitations required extensive and time-consuming efforts to establish software pipelines for data collection, requiring significant human interaction. A substantial portion of the time was spent on preparation and post-processing of data before the AI/ML models could be engaged for training. This approach will not be sustainable for future cellular networks, which are envisioned to be fully autonomous and operate without human intervention. In addition, a more comprehensive perspective on AI/ML applications will be essential, where distributed AI/ML solutions, such as those explored in this thesis, will need to interact extensively with other centralized components of the network to achieve fully autonomous and cognitive behavior.

5.2.2 INTEGRATION OF NATIVE AI/ML TECHNOLOGIES

Network efficiency and generalization challenges have been extensively discussed in the context of AI/ML-assisted communication systems, with generative AI architectures being particularly affected. These issues, especially the high computational and memory complexity driven by the self-attention mechanism, may hinder the broader adoption of generative AI. Algorithmic efficiency and resource optimization in AI development will become increasingly important, especially when integrated into systems like cellular networks. Rather than relying solely on brute-force computation, achieving high performance with significantly fewer resources will be crucial, challenging the traditional notion that larger models and datasets are always superior. This paradigm shift will promote the development of more innovative and sustainable approaches. With growing concerns about the carbon footprint of AI, adopting sustainable AI technologies will be essential to reduce energy consumption and minimize the use of computational resources with fewer environmental impacts.

In this thesis, the training and evaluation of AI/ML models, based on commercial data, were performed offline, meaning that none of the solutions presented here have been deployed in real networks as of yet. That being said, for the successful integration of AI architectures in future networks, it is essential to optimize their computational efficiency and generalization capabilities through tailored AI/ML architectures for wireless applications, supported by dedicated hardware resources. This requires a thorough evaluation of AI's real-time performance. Leveraging fast GPUs or purpose-built hardware will be vital in minimizing delays and ensuring efficient operation within commercial cellular networks. Also, hybrid approaches, such as cloud-based solutions, are likely to incorporate AI/ML technologies to enhance and support the RAN. These approaches can leverage the scalability and computational power of the cloud to process large-scale data, enabling advanced AI/ML-driven functionalities like network optimization, real-time decision-making, and predictive analytics. This integration can address the growing complexity of cellular networks.

5.2.3 DATA INTEGRITY AND AI/ML TRUSTWORTHINESS

Trust and reliance on modern telecom systems are widespread. However, the adoption of AI introduces new risks widely discussed as this thesis is been written. How this is going to be regulated is not clear yet and is a subject discussed outside of the wireless research field as well. To realize the full potential of AI, trustworthiness is a prerequisite and shall be built into the system by design, addressing aspects spanning from explainability and

human oversight to security and built-in safety mechanisms. A clear example of this is the positioning studies presented in Papers II, IV, and V. In these studies, the location of the UE was estimated using fingerprinted features derived from channel measurement data. However, this approach indirectly exposed the user's geographical position, which is typically not exchanged between the UE and the network. Given the sensitive nature of positioning information and its close association with personal privacy, such data must always be handled with the utmost caution. The final word on how AI/ML technologies should be regulated has yet to be spoken, and it will eventually become a research field of its own in the future.

5.2.4 BEYOND 5G AND 6G VISION

Smartphone-based video streaming has historically dominated mobile network traffic. The smartphone-driven app economy has already transformed daily life, from banking to shopping to healthcare. As technology advances, XR and AI hold the potential to extend this digital revolution, offering immersive experiences that go beyond screens and integrate digital objects into the physical world, as illustrated Fig. 1.1. The evolution of XR and AI technologies will push network capabilities, particularly within the challenging RAN domain, far beyond current limits. These immersive and interactive applications will demand the real-time transmission of vast volumes of data, ultra-low latency to ensure seamless interactivity, and dynamic adaptability to manage fluctuating demands from multiple users in shared digital environments. 6G will play a pivotal role in enabling next-generation experiences, while 5G technology provides a strong foundation for transitioning into the 6G era to meet the demands of future society.

Another aspect that will add significant complexities to the RAN is the integration of high-band and ultra-high frequency bands, such as those in the terahertz range. These frequency bands, which extend beyond the millimeter-wave frequencies commonly used in 5G, offer tremendous potential for massive data throughput and ultra-low latency communication. However, at the time this thesis is being finalized, terahertz frequencies exhibit significant challenges, primarily related to propagation and coverage. These challenges will require sophisticated antenna arrays bringing increased complexity in RAN architectures. Overcoming these limitations will necessitate the development of advanced signal processing techniques, with the integration of AI/ML technologies being essential for the management of these networks.

Bibliography

[1] OpenAI, "chatGPT," https://chatgpt.com/, accessed: 2024-05-05.

[2] ——, "DallE," https://openai.com/dall-e, accessed: 2025-02-15.

[3] DeepMind, "Alphafold," https://www.deepmind.com/research/highlighted-research/alphafold, accessed: 2025-05-15.

[4] W. C. Y. Lee, *Mobile Cellular Telecommunications: Analog and Digital Systems*. McGraw-Hill Professional; 2nd edition, Feb 1995.

[5] 3GPP-44.xxx Series, "3rd Generation Partnership Project (3GPP), GSM Core Network." 3rd Generation Partnership Project (3GPP), 2001. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/44_series/

[6] 3GPP-45.xxx Series, "3rd Generation Partnership Project (3GPP), GSM Radio Access." 3rd Generation Partnership Project (3GPP), 2001. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/45_series/

[7] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications*. Telecom Pub, Jan 1992.

[8] T. S. Rappaport, *Wireless Communications: Principles and Practice*. Prentice Hall, Dec 2001.

[9] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS*. Wiley–Blackwell, Apr 2002.

[10] 3GPP-25.xxx Series, "3rd Generation Partnership Project (3GPP), UMTS Radio Access Network (UTRAN)." 3rd Generation Partnership Project (3GPP), 2001. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/25_series/

[11] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. John Wiley Sons, Ltd, July 2002.

[12] H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Naghian, and V. Niemi, *UMTS Networks: Architecture, Mobility and Services, 2nd Edition*. Wiley, July 2005.

[13] 3GPP-36.xxx Series, "3rd Generation Partnership Project (3GPP), LTE and LTE-Advanced, Release 8 and 13." 3rd Generation Partnership Project (3GPP), 2024. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/36_series/

[14] H. Holma and A. Toskala, *LTE for UMTS: Evolution to LTE-Advanced*. Wiley, Apr 2011.

[15] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. Wiley, March 2011.

[16] 3GPP-38.xxx Series, "3rd Generation Partnership Project (3GPP), 5G NR, Release 15, 16 and 17." 3rd Generation Partnership Project (3GPP), 2024. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/

[17] 3GPP-23.xxx Series, "3rd Generation Partnership Project (3GPP), 5G Core, Release 15, 16 and 17." 3rd Generation Partnership Project (3GPP), 2024. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/

[18] A. Osseiran, J. F. Monserrat, and P. Marsch, *5G Mobile and Wireless Communications Technology*. Cambridge University Press, June 2016.

[19] P. Marsch, Ömer Bulakci, O. Queseth, and M. Boldi, *5G System Design: Architectural and Functional Considerations and Long Term Research*. Wiley, May 2018.

[20] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334 – 366, Feb 2021. doi: 10.1109/OJCOMS.2021.3057679

[21] A. Clemm, M. T. Vega, H. K. Ravuri, and T. Wauters, "Toward truly immersive holographic-type communication: Challenges and solutions," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 93 – 99, Jan 2020. doi: 10.1109/MCOM.001.1900272

[22] R. Petkova, I. Bozhilov, A. Manolova, K. Tonchev, and V. Poulkov, "On the way to holographic-type communications: Perspectives and enabling technologies," *IEEE Access*, vol. 12, pp. 59 236 – 59 259, Apr 2024. doi: 10.1109/ACCESS.2024.3393124

[23] E. Dahlman, S. Parkvall, and J. Sköld, *5G NR, The Next Generation Wireless Access Technology*. Academic Press, 2018.

[24] 3GPP-23501, "3rd Generation Partnership Project (3GPP), System architecture for the 5G System (5GS)." 3rd Generation Partnership Project (3GPP), 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.501/

[25] 3GPP-23502, "3rd Generation Partnership Project (3GPP), Procedures for the 5G System (5GS)." 3rd Generation Partnership Project (3GPP), 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.502/

[26] 3GPP-23503, "3rd Generation Partnership Project (3GPP), Policy and charging control framework for the 5G System (5GS); Stage 2." 3rd Generation Partnership Project (3GPP), 2022. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.503/

[27] G. Charbit, A. Medles, , P. Jose, D. Lin, and X. Z. andI Kang Fu, "Satellite and cellular networks integration - a system overview," Jul. 2021. doi: 10.1109/EuCNC/6GSummit51104.2021.9482585

[28] S. Chen, Y.-C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, "Vision, requirements, and technology trend of 6G: How to tackle the challenges of system coverage, capacity, user data-rate and movement speed," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 218 – 228, Feb. 2020. doi: 10.1109/MWC.001.1900333

[29] Ericsson, "6G Explained," Tech. Rep., Nov 2024. [Online]. Available: https://www.ericsson.com/en/6g

[30] H. Pennanen, T. Hänninen, O. Tervo, A. Tölli, and M. Latva-Aho, "6G: The intelligent network of everything," *IEEE Access*, vol. 13, pp. 1319 – 1421, Dec. 2024. doi: 10.1109/ACCESS.2024.3521579

[31] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, and C. Zhang, "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Communications Surveys Tutorials*, vol. 25, pp. 905 – 974, Feb 2023. doi: 10.1109/COMST.2023.3249835

[32] Nokia, "Technology Vision," Tech. Rep., Jan 2025. [Online]. Available: https://www.nokia.com/innovation/technology-vision/

[33] L. U. Khan, Z. Han, D. Niyato, M. Guizani, and C. S. Hong, "Metaverse for wireless systems: Vision enablers architecture and future directions," *IEEE Wireless Communications*, vol. 31, no. 4, pp. 245 – 251, Aug. 2024. doi: 10.1109/MWC.013.2300287

[34] A. M. Aslam, R. Chaudhary, A. Bhardwaj, I. Budhiraja, N. Kumar, and S. Zeadally, "Metaverse for 6G and beyond: The next revolution and deployment challenges," *IEEE Internet of Things Magazine*, vol. 6, no. 1, pp. 32 – 39, Mar. 2023. doi: 10.1109/IOTM.001.2200248

[35] M. Zawish, F. A. Dharejo, S. A. Khowaja, S. Raza, S. Davy, and K. Dev, "Ai and 6G into the metaverse: Fundamentals challenges and future research trends," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 730 – 778, Jan. 2024. doi: 10.1109/OJCOMS.2024.3349465

[36] F. Tang, X. Chen, M. Zhao, and N. Kato, "The roadmap of communication and networking in 6G for the metaverse," *IEEE Wireless Communications*, vol. 30, no. 4, pp. 72 – 81, Aug. 2022. doi: 10.1109/MWC.019.2100721

[37] M. H. Alsamh, A. Hawbani, S. Kumar, and S. H. Alsamhi, "Multisensory metaverse-6G: A new paradigm of commerce and education," *IEEE Access*, vol. 12, pp. 75 657–75 677, Apr. 2024. doi: 10.1109/ACCESS.2024.3392838

[38] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gunduz, and V. Poor, "Transformer-empowered 6G intelligent networks: From massive MIMO processing to semantic communication," *IEEE Wireless Communications*, vol. 30, no. 6, pp. 127 – 135, Dec. 2023. doi: 10.1109/MWC.008.2200157

[39] Y. C. Eldar, A. Goldsmith, D. Gunduz, and H. V. Poor, *Machine Learning and Wireless Communications*. Cambridge University Press, Aug. 2022.

[40] E. Dahlman, S. Parkvall, and J. Sköld, *5G/5G-Advanced, third edition*. Academic Press, 2023.

[41] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," Nov. 2010. doi: 10.1109/TWC.2010.092810.091092

[42] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," Feb. 2014. doi: 10.1109/MCOM.2014.6736761

[43] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," Dec. 2012. doi: 10.1109/MSP.2011.2178495

[44] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3899–3911, July 2015. doi: 10.1109/TWC.2015.2414413

[45] X. Li, K. Batstone, K. Åstrom, M. Oskarsson, C. Gustafson, and F. Tufvesson, "Robust phase-based positioning using massive MIMO with limited bandwidth," Oct. 2017. doi: 10.1109/PIMRC.2017.8292362

[46] J. Vieira, E. Leitinger, M. Sarajlić, X. Li, and F. Tufvesson, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," Oct. 2017. doi: 10.1109/PIMRC.2017.8292280

[47] X. Wang, E. Grinshpun, D. Faucher, and S. Sharma, "On medium and long term channel conditions prediction for mobile devices," May 2017. doi: 10.1109/WCNC.2017.7925951

[48] F. Chiariotti, D. D. Testa, M. Polese, A. Zanella, G. M. D. Nunzio, and M. Zorzi, "Learning methods for long-term channel gain prediction in wireless networks," Mar. 2017. doi: 10.1109/ICCNC.2017.7876120

[49] C. Cortes and V. Vapnik, "Support-vector networks," Sep. 1995. doi: doi:10.1007/BF00994018

[50] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[51] S. Navabi, C. Wang, O. Y. Bursalioglu, and H. Papadopoulos, "Predicting wireless channel features using neural networks," May 2018. doi: 10.1109/ICC.2018.8422221

[52] Y. Xu, W. Xu, F. Yin, J. Lin, and S. Cui, "High-accuracy wireless traffic prediction: A gp-based machine learning approach," Jan. 2018. doi: 10.1109/GLOCOM.2017.8254808

[53] A. Molisch, *Wireless Communications*. John Wiley & Sons, 2012.

[54] P. Kyösti, J. Meinilä, L. Hentila, and X. Zhao, "Ist-4-027756 winner ii d1.1.2 v1.2, winner ii channel models." Information Society Technologies, Dec. 2007.

[55] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, and F. Quitin, "The cost 2100 mimo channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92 – 99, Dec. 2012. doi: 10.1109/MWC.2012.6393523

[56] 3GPP-38901, "3rd Generation Partnership Project (3GPP), Study on channel model for frequencies from 0.5 to 100 GHz." 3rd Generation Partnership Project (3GPP), 2017. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/38.901/

[57] T. S. Rappaport, R. H. Jr., and R. Daniels, *Millimeter Wave Wireless Communications*. Prentice Hall, Sept. 2014.

[58] Z. Yun and M. F. Iskander, "Ray tracing for radio propagation modeling: Principles and applications," *IEEE Access*, vol. 3, p. 1089 – 1100, Jul 2015. doi: 10.1109/ACCESS.2015.2453991

[59] Fraunhofer HHI: Fraunhofer Heinrich-Hertz-Institut. Quadriga. [Online]. Available: https://quadriga-channel-model.de/

[60] Remcom. Wireless insite. [Online]. Available: https://www.remcom.com/wireless-insite-em-propagation-software

[61] Ericsson, *Massive MIMO Handbook, extended version*. Ericsson AB, 2024.

[62] ——, "Massive MIMO for 5G networks," Tech. Rep., Aug. 2017. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/advanced-antenna-systems-for-5g-networks

[63] A. Paulraj, R. Nabar, and D. Gore, *Introducton To Space-Time Wireless Communications*. Cambridge University Press, 2008.

[64] 3GPP-36133, "3rd Generation Partnership Project (3GPP), 5G NR – Requirements for support of radio resource management." 3rd Generation Partnership Project (3GPP), 2020. [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/36_series/36.133/

[65] 3GPP-38802, "3rd Generation Partnership Project (3GPP), Study on New Radio Access Technology - Physical Layer Aspects." 3rd Generation Partnership Project (3GPP), 2017. [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/38_series/38.802/

[66] R. M. Dreifuerst and R. W. Heath, "Massive MIMO in 5G: How Beamforming, Codebooks, and Feedback Enable Larger Arrays," *IEEE Communications Magazine*, vol. 61, no. 12, pp. 18 – 23, Dec. 2023. doi: 10.1109/MCOM.001.2300064

[67] E. Bećirović, E. Björnson, and E. G. Larsson, "Massive MIMO in 5G: How Beamforming, Codebooks, and Feedback Enable Larger Arrays," *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 10 065 – 10 080, Nov. 2022. doi: 10.1109/TWC.2022.3182749

[68] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74 – 80, Feb. 2014. doi: 10.1109/MCOM.2014.6736746

[69] R. W. Heath and A. Lozano, *Foundations of MIMO Communication*. Cambridge University Press, Dec. 2018.

[70] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 173 – 196, Sep. 2018. doi: 10.1109/COMST.2018.2869411

[71] 3GPP-38214, "3rd Generation Partnership Project (3GPP), NR; Physical layer procedures for data." 3rd Generation Partnership Project (3GPP), 2020. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.214/

[72] 3GPP-38104, "3rd Generation Partnership Project (3GPP), NR; Base Station (BS) radio transmission and reception." 3rd Generation Partnership Project (3GPP), 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.104/

[73] 3GPP-38211, "3rd Generation Partnership Project (3GPP), NR; Physical channels and modulation." 3rd Generation Partnership Project (3GPP), 2020. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.211/

[74] 3GPP-38213, "3rd Generation Partnership Project (3GPP), NR; Physical layer procedures for control." 3rd Generation Partnership Project (3GPP), 2020. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.213/

[75] 3GPP-38215, "3rd Generation Partnership Project (3GPP), NR; Physical layer measurements." 3rd Generation Partnership Project (3GPP), 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.215/

[76] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[77] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Interference, and Prediction*. Springer, 2017.

[78] P. Flach, *Machine Learning*. Cambridge University Press, 2017.

[79] R. Benin, *Machine Learning for Developers*. Packt, Oct. 2017.

[80] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2017.

[81] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, Dec 2018.

[82] H. Barlow, *Neural Computation*. MIT Press, Sep 1989.

[83] R. Sutton and A. Barto, "Reinforcement learning: An introduction," Sep. 1998. doi: 10.1109/TNN.1998.712192

[84] S. P. Singh and D. Bertsekas, *Reinforcement learning for dynamic channel allocation in cellular telephone systems*. The MIT Press, 1997.

[85] G. J. Oyewole and G. A. Thopil, *Data clustering: application and trends*. Springer Nature, Nov. 2022.

[86] A. Turing, "On computable numbers," Nov. 1936.

[87] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," 1943.

[88] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," Mar. 2020. doi: https://doi.org/10.1016/j.physd.2019.132306

[89] Y. Chen, Y. Kang, Y. Chen, and Z. Wang, "Probabilistic forecasting with temporal convolutional neural network," Jul. 2020. doi: https://doi.org/10.48550/arXiv.1906.04397

[90] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Jun. 2017. doi: https://doi.org/10.48550/arXiv.1706.03762

[91] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067 – 1080, Oct. 2007. doi: 10.1109/TSMCC.2007.905750

[92] R. D. Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability latency and robustness of 5G," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 102 – 112, Oct. 2014. doi: 10.1109/MSP.2014.2332611

[93] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "Position and orientation estimation through millimeter wave MIMO in 5G systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1822–1835, Mar. 2018. doi: 10.1109/TWC.2017.2785788

[94] K. Witrisal, P. Meissner, E. Leitinger, Y. Shen, C. Gustafson, and F. Tufvesson, "High-accuracy localization for assisted living: 5G systems will turn multipath channels from foe to friend," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 59 – 70, Mar. 2016. doi: 10.1109/MSP.2015.2504328

[95] A. Guerra, F. Guidi, and D. Dardari, "On the impact of beamforming strategy on mm-wave localization performance limits," *IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2017. doi: 10.1109/ICCW.2017.7962758

[96] ——, "Single-anchor localization and orientation performance limits using massive arrays: MIMO vs. beamforming," *IEEE Transactions on Wireless Communications)*, vol. 17, no. 8, pp. 5241 – 5255, Aug. 2018. doi: 10.1109/TWC.2018.2840136

[97] M. M. Butt, A. Rao, and D. Yoon, "RF Fingerprinting and Deep Learning Assisted UE Positioning in 5G," *IEEE 91st Vehicular Technology Conference*, May 2020. doi: 10.1109/VTC2020-Spring48590.2020.9128640

[98] G. Kia, L. Ruotsalainen, and J. Talvitie, "A CNN Approach for 5G mm Wave Positioning Using Beamformed CSI Measurements," *2022 International Conference on Localization and GNSS (ICL-GNSS)*, June 2022. doi: 10.1109/ICL-GNSS54081.2022.9797028

[99] 3GPP-38331, "3rd Generation Partnership Project (3GPP), NR; Radio Resource Control (RRC); Protocol specification." 3rd Generation Partnership Project (3GPP), 2022. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.331/

[100] 3GPP-38855, "3rd Generation Partnership Project (3GPP), Study on NR positioning support." 3rd Generation Partnership Project (3GPP), 2019. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.855/

[101] S. Dwivedi, R. Shreevastav, F. Munier, J. Nygren, I. Siomina, Y. Lyazidi, D. Shrestha, G. Lindmark, P. Ernstrom, E. Stare *et al.*, "Positioning in 5G networks," Feb. 2021. doi: https://doi.org/10.48550/arXiv.2102.03361

[102] S. Huang, H.-M. Chen, B. Wang, J. Chai, X. Wu, and F. Li, "Positioning Performance Evaluation for 5G Positioning Reference Signal," Aug. 2022. doi: 10.1109/ICFEICT57213.2022.00093

[103] M. Elsayed and M. Erol-Kantarci, "AI-Enabled Future Wireless Networks: Challenges, Opportunities, and Open Issues," *IEEE Vehicular Technology Magazine*, vol. 14, pp. 70 – 77, Sep. 2019. doi: 10.1109/MVT.2019.2919236

[104] Q. Liu, T. Zhang, M. Hemmatpour, H. Qiu, D. Zhang, and C. S. Chen, "Operationalizing AI/ML in Future Networks: A Bird's Eye View from the System Perspective," *IEEE Communications Magazine*, vol. 63, no. 4, pp. 176 – 182, Apr. 2025. doi: 10.1109/MCOM.001.2400033

[105] C. K. Thomas, C. Chaccour, W. Saad, M. Debbah, and C. S. Hong, "Causal Reasoning: Charting a Revolutionary Course for Next-Generation AI-Native Wireless Networks," *IEEE Vehicular Technology Magazine*, vol. 19, no. 1, pp. 16 – 31, Mar. 2024. doi: 10.1109/MVT.2024.3359357

PAPERS

Paper I

Paper I

Reproduced, with permission from IEEE

D. Pjanić, A. Sopasakis, H. Tataria, F. Tufvesson, and A. Reial, "Learning-Based UE Classification in Millimeter-Wave Cellular Systems With Mobility," *IEEE International Workshop on Machine Learning for Signal Processing*, Oct. 2021, Gold Coast, Australia, doi: 10.1109/MLSP52302.2021.9596275.

Learning-Based UE Classification in Millimeter-Wave Cellular Systems With Mobility

Dino Pjanić[1,2], Alexandros Sopasakis[3], Harsh Tataria[2],
Fredrik Tufvesson[2], and Andres Reial[1]
[1] Ericsson AB, Sweden
[2] Dept. of Electrical and Information Technology, Lund University, Sweden
[3] Dept. of Mathematics, Lund University, Sweden
e-mail: {dino.pjanic, harsh.tataria, andres.reial}@ericsson.com,
alexandros.sopasakis@math.lth.se, and fredrik.tufvesson@eit.lth.se

Abstract

Millimeter-wave cellular communication requires beamforming procedures that enable alignment of the transmitter and receiver beams as the user equipment (UE) moves. For efficient beam tracking it is advantageous to classify users according to their traffic and mobility patterns. Research to date has demonstrated efficient ways of machine learning based UE classification. Although different machine learning approaches have shown success, most of them are based on physical layer attributes of the received signal. This, however, imposes additional complexity and requires access to those lower layer signals. In this paper, we show that traditional supervised and even unsupervised machine learning methods can successfully be applied on higher layer channel measurement reports in order to perform UE classification, thereby reducing the complexity of the classification process.

Index Terms— 5G, classification, beam management, machine learning, millimeter-wave, mobility.

1. INTRODUCTION

In wireless communication systems, optimization of the radio access network (RAN) has always been an important area [1, 2]. In the context of fifth-generation

(5G) systems, a key issue from a network performance viewpoint is how the RAN can adapt to the dynamic radio environment [3]. Irrespective of the deployment mode of 5G systems, it is understood that massive multiple-input multiple-output (MIMO) is expected to be the workhorse of its RAN front-haul at the cellular base stations (BSs) [3]. In the 24.250 - 52.6 GHz band, analog or digital beamforming in both azimuth and elevation domains is typically used to increase the received signal levels. The beamforming procedure provides access to control and payload signals for network-level decision making. By studying the characteristics of these signals to/from each UE, *spatial fingerprints* unique to each UE can be found [4]. In the related literature, spatial fingerprinting has been used in conjunction with classical machine learning (ML) methods for UE localization via the learned "features" of the environment [5], or by direct matching [6] without learning the environment features. The authors in [7] train a six-layer fully connected network on real-time observations at sub-6 GHz bands to predict beamforming weight coefficients and blockages, while the study of [8] demonstrates the use of a simple, feed-forward neural network for band assignment to different UEs. While [9] surveys an extensive list of related literature, the vast majority of the works in the literature only consider physical layer (PHY) properties of the transmitted/received signal and do not capture the interaction of the PHY with the data link and media access control layers of the system. In reality, these higher system layers greatly manipulate the PHY signals seen to/from the phased array ports which capture the physical amplitude and phase properties, as well as embed protocol level detail for RAN performance optimization.

Considering a cross-layer perspective, this paper investigates whether classical ML methods are capable of classifying UEs into different groups by simultaneously processing layer 2 (L2) uplink channel state information-reference signals (CSI-RSs). A key consideration in our analysis is that of UE mobility, thus making the CSI-RSs time varying for each UE. We consider different mobility patterns of the UEs in a typical urban setting and include movements such as walking, cycling, and traveling with cars or public transportation. The objective is to investigate the use of learning algorithms towards UE classification based solely on the reported narrow beams and corresponding Reference Signal Received Power (RSRP) values in dynamic millimeter-wave (mmWave) scenarios. Previous studies on beam management have mostly focused on best beam predictions [10] and have not dealt with combining RSRP measurements across multiple wide beams and the corresponding narrow

beams [11]. In particular, we employ both supervised and unsupervised methods such as tSNE [12] and K-means clustering [9], combined with principal component analysis (PCA) in order to classify the UE types. Furthermore we train a number of decision tree classifiers in order to further explore whether it is indeed possible to characterize the modality (i.e. pedestrian, car etc) from the measurement reports.

Our analysis here assumes line-of-sight (LOS) propagation at a center frequency of 28 GHz across a 100 MHz bandwidth. There are several reasons for this: First, one of our objectives is to assess whether unsupervised learning is able to detect different UE types given their individual time evolving CSI-RS reference signals. Second, the majority of the existing studies (see earlier references) confine their focus to PHY parameters and do not consider the effects of PHY-higher layer co-design since it is difficult to decouple the PHY effects from higher layers once certain access control and protocol-level decisions are made. Finally, the results can be useful for approximating the achievable performance obtained by very sparse mmWave channels, since the consensus is that there only seems to be a few *dominant* multipath components which are active [1, 2]. We note that since our target is to study UE classification at standardized mmWave frequencies, bands outside the one mentioned above is not within the direct scope of the study and hence conclusions made here can not be extrapolated to other frequencies.

2. Simulated System Description and Methodology

We now describe the commercial grade simulated system which is responsible for generating our data. Unlike previous works, we consider a radio system simulator supporting detailed beam management procedures compliant with standardized 5G systems at mmWave frequencies [13]. As a result, we are able to simulate a commercial grade 28 GHz phased array antenna module (PAAM) in BS type 1-O configuration containing 192 cross-polarized elements distributed across 8 rows and 12 columns. Analog beamforming capability is modelled with horizontal and vertical inter-element spacings fixed to 0.5 λ and 0.7 λ with each cross-polarized element having a direction-specific gain pattern given in [14]. The PAAM is tuned for operation within a bandwidth of 100 MHz (standards compliant). The generated grid-of-beams (GoB) from the PAAM is depicted in Fig. **1** yielding a total of 12 wide beams (WB) and 136 narrow beams (NB).

The simulated three-dimensional area deploys a single site with one hexagon shaped sector having a radius of 200 m. The BS is deployed in the corner of the

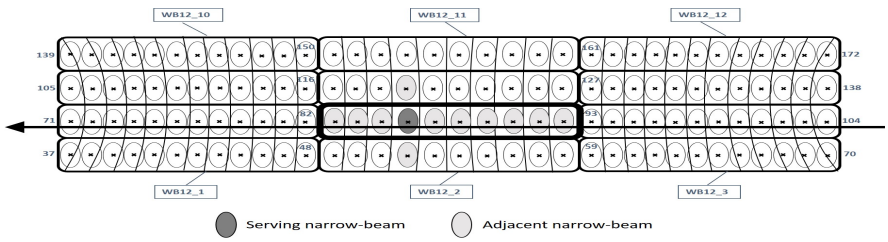


Figure 1: Generated GoB with fixed wide/narrow beam numbering. Conceptual overview of beam refinement procedure: During a typical P2 procedure and with the help of incoming UE measurement reports, NB tracking is performed within the current WB and if RSRP of the best NB is higher than the currently serving NB, a beam switch is initiated. Note: UEs are moving in front of the BS along their lane in same direction as the depicted arrow.

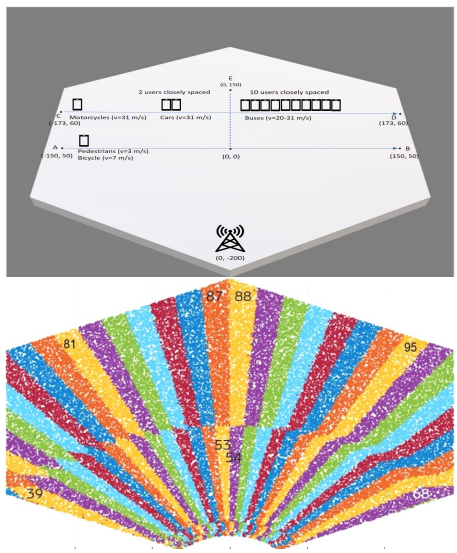


Figure 2: (Top) Site model. The UEs are placed in LOS moving along two different lanes. (Bottom) An example of narrow-beam coverage showing a subset of beams present in the horizontal plane at a given time instant (as depicted in Fig. **1**).

hexagon as depicted in Fig. **2**, with height of 30 m having an overall field-of-view of 120° in the azimuth and 40° in elevation. No mechanical downtilt of the BS is assumed. A protection radius of 5 m is kept from the BS periphery where no UEs can be located within 5 m from the BS. Prior to data collection, two UE mobility patterns have been designed, namely slow and fast moving UEs. The first group mimics pedestrian and bicycle while the latter one mimics vehicular UEs such as, motorcycle, car and bus, 5 main classes in total. In addition, user specific mobility patterns have been modelled, presenting a pavement and a street. In order to reduce built-in bias in our data sets, randomness was applied to the starting points as well as UEs movement trajectory. The UEs are moving from left to right in front of the BS having free LOS propagation. Velocities and spacing between UEs are modelled in a way that corresponds to their mobility pattern, as depicted in Fig. **2**. For instance, a car carrying two persons is represented by two UEs physically separated by one meter and moving together at the same time and velocity. The height of each UE is set to 1.5 m. During simulations UEs emerge at predefined starting points along their route and continue moving to the endpoint covering the entire length of the route. In contrast to those, some UEs would emerge at intermediate points and thereby cover just fraction of the entire route. A predefined number of pedestrian UEs is scheduled to cross the street simulating northward movement away from the BS. During the simulation multiple parameter settings are triggered and evaluated for different loads of UEs, and several seeds are used to ensure statistical confidence. Fig. **2** shows narrow beam coverage with corresponding numbering/indexing across the simulated geographical area (as generated from the system explained in Fig. **1**). We note the above was done in conjunction with uplink file transfer protocol traffic patterns.

3. Problem Definition and Formulation

Maintaining good radio link reliability is a key challenge for mmWave communication systems, especially when mobility is incorporated. Directional links, however, require fine alignment of the transmitter and receiver beams, achieved through a mechanism known as beam management [15]. In line with [16], three downlink layer 1/2 (L1/L2) beam management procedures, commonly known as P1, P2 and P3 are involved. In this study, our primary focus was on the P2 procedure which handles beam tracking at the BS (a.k.a. gNB). Note that tracking refers to gNB refining beams (e.g., sweeping through all the narrow beams over a small range) where UEs detect the best (service) beam and report its index to gNB. As illustrated

in Fig. **1**, the P2 CSI-RS measurements were performed by the moving UEs (as specified in [17]) for all NB within the same WB (synchronization signal block (SSB) beam) and the reported RSRP measurements were collected.

A considerable amount of literature has been proposing different ML approaches for classification of UEs, mostly applied on data sets generated on physical radio channel attributes such as: e.g., precoding schemes, modulation scheme [18] or channel covariance [19]. As UEs tend to move along predefined routes in the physical environment while establishing fingerprints at the BS, e.g. by channel measurement, we argue that using L2 report data-sets can be advantageously utilized in order to simplify the UE classification due its less complex nature compared to physical channel parameters. This approach opens up for new ways of learning from UE mobility patterns and thereby possibly prevent UEs from ending up in unfavorable radio channel conditions.

4. Un/supervised Learning Framework

To perform our analysis we utilize the fact that mmWave communication is sensitive to UE motion. We explore whether machine learning methods can learn motion characteristics and perhaps even users intent from RSRP measurements as reported on different narrow-beams. One difficulty is that our measurement report, as in real life, is typically ill-balanced. Therefore, in the case of K-means, we first perform Principal Component Analysis (PCA) in order to uncover the clustering structure of our data in an unsupervised setting. We use 50 components for the PCA which accounts for 97% of the cumulative explained variation. In order to ascertain the number of clusters K we employ the elbow method [20] and subsequently explore classification of UEs into the main five classes comprising our data: pedestrian, bicycle, car, bus and motorcycle.

In the subsequent analysis, we explicitly avoid comparing multiple unsupervised classification methods and questions about optimality of one approach over another. Our primary objective is to understand whether it is feasible to utilize knowledge of the time evolving CSI-RS signals for classifying different UEs. Such information can then be used for predicting and optimizing their radio resource requirements. As such, from a ML viewpoint, we opt for the simplest combination of the K-means clustering with PCA, rather than a more complex supervised or unsupervised learning approach based on deep learning methods. We point out that while these

are extremely interesting directions, we cover them in detail in a follow up extension of this paper.

A. *Feature extraction and the data*

Each UE measurement report contains 12 RSRP CSI-RS measurements together with their corresponding narrow-beams and arrives on a approximate 40 ms basis. The total number of such reports, which essentially indicates the size of our dataset, is highly dependent on the duration of simulation, as well as the UE velocities as described in Section **2**. Identity information such as the UE identity number, UE location and time are not exposed to the ML algorithm but rather used as labeled references when interpreting results of classification. When dealing with incomplete and unbalanced data sets a number of approaches are possible. To counter the data imbalance we employ PCA as a pre-processing step. As explained previously, we apply PCA with 50 components which accounts for 97% of explained variation in the data. Furthermore, due to data inhomogeneity we apply feature scaling where the range of each feature is normalized so that it contributes proportionately. This approach is essential since the clustering methods calculate distances between data points. Subsequently the two features, RSRP and corresponding narrow-beams, are stacked on top of each other representing one single UE fingerprint at the base station at a certain report time. In tests, not presented here, we observed that having only RSRP or only the narrow beam indices as input to our ML model results in degradation of the classification rate.

5. Results and Discussion

We now present results from both unsupervised and supervised ML models towards our goal: UE classification from their signals. As discussed in Section **2**, our data consists of 5 main classes, with sub-classes (groups), of UEs according to their mobility pattern. We first train our algorithm to detect all 5 classes. The results are presented in Fig. **3**. There is significant overlap between the different classes which may not be so surprising given how similar UEs must look for groups such as cars and busses or pedestrians and cyclists. This underscores the need to explore whether we can learn to identify larger groups such as for instance slow and fast movers. We will undertake this task next. Using the sama data from Fig. **3** a similar although perhaps more intriguing result emerges when we set $K = 2$ groups. Specifically we observe in Fig. **4** that some separation exist between fast (cars, busses, motorcycles)

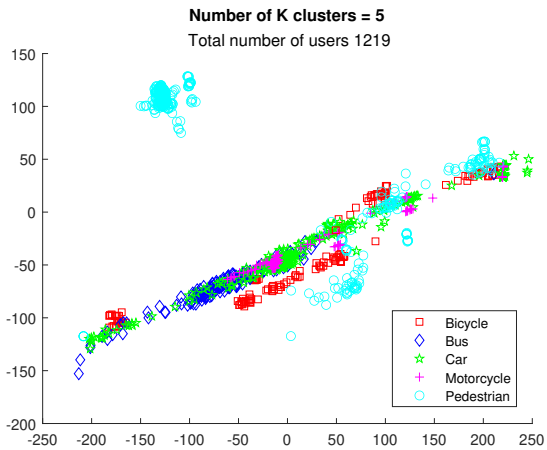


Figure 3: K-means clustering results for all 5 classes of UEs. Training on 5 groups: 165 bicycles, 180 buses, 320 cars, 130 motorcycles and 424 pedestrians. Class separation is not obvious at this scale but shows pattern similarities may exist between classes of slow movers or classes of fast movers.

and slow (bicycles and pedestrians) moving UEs. The overlap has now reduced significantly when only grouping into 2 groups. We explore this initial result further below.

A. *Slow-moving UEs - pedestrians vs bicycles*

It is not surprising, especially based on the baseline results of Fig.**4**, that velocity of a given UE can be a good indicator as to whether it should be classified as a car or a pedestrian. Differentiating however between similarly moving UEs should be more challenging. We undertake this task in an unsupervised setting for now by examining a mixed group of 589 UEs consisting of 424 pedestrians and 165 bicycles. We note that UEs in that dataset move along identical trajectories on the pavement with either small or no variation and ending/starting locations unrelated to their group type. We also note that some of the pedestrians will also cross the street while the bicycles do not. We train on this data with a PCA method followed by K-means as discussed earlier. The results in Fig. **5** show a rather clear separation

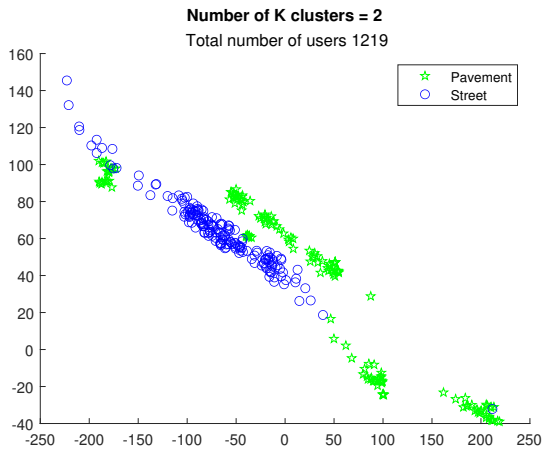


Figure 4: K-means clustering results for all 5 classes of UEs. Training on 2 groups: fast 630 (cars, busses, motorcycles) versus 589 slow-moving (bicycles, pedestrians) UEs.

between the two classes of users with a surprisingly small overlap even though the 2 groups have almost similar moving velocities. Specifically, we observe a clear diagonal linear yet separated trend in the majority of the data - we should point out, that the clustering space is non-dimensional and has no physical meaning. We also note an interesting cluster of pedestrians at the top left corner of that figure which seems to defy the overall pattern. To better understand this cluster we then train our algorithm on the single group of 424 pedestrians from Fig. **5** while requiring that $K = 2$ once again. The results presented in Fig. **6** verify that the majority of that sub-group is attributed to the street-crossing pedestrians. The results therefore show that identification and separation among all classes and even sub-group of the slow UEs could be possible.

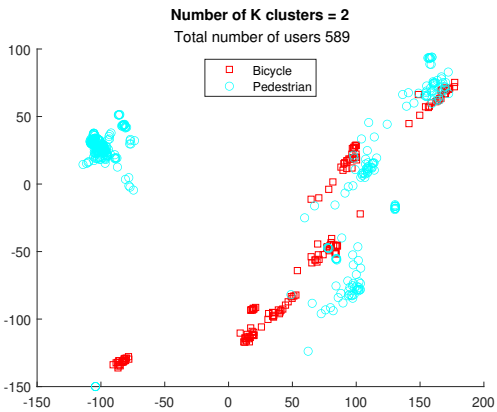


Figure 5: K-means clustering results for slow-moving UEs. Training on 2 groups: 165 bicycles versus 424 pedestrians. Compare with Fig. **6** where pedestrians are further distinguished into crossing/non-crossing the street.

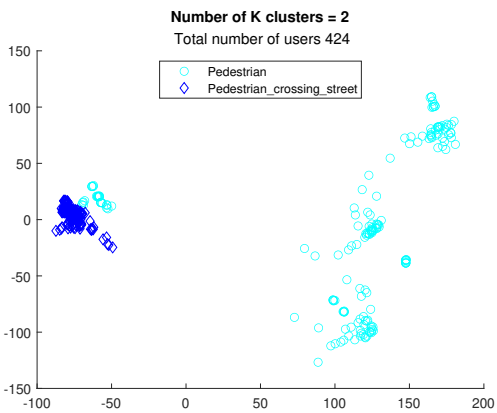


Figure 6: K-means clustering results after training on just the class of pedestrian UEs in Fig. **5**. Training on 2 groups: 182 pedestrians non-crossing the street versus 242 crossing the street. Pedestrians crossing the street seem to be mainly responsible for that single cluster.

B. *Classification*

We now train a number of different decision tree type (supervised) classifiers in order to identify specific UE mobility classes based on this data. Furthermore we task the classifiers to distinguish modality with high probability (see Table **1**) based on only a short (40 ms) sequence of RSRP and narrow-beam numbers reported from each UE. We present a list of the most successful of those and the respective type of UE data used for their training in Table **1**. Our main metric for success or failure here is the miss-classification rate on unseen data. The best classifier, the Extra Trees Regressor, achieves a miss-classification rate of 2%. Overall we found that training

Classifier	Data used	Miss-Classif %.
Extra Trees Regres.	Ped-nc, Car, MC	1.8
Extra Trees Regres.	Ped-cr, Car, MC	2.2
Extra Trees Regres.	Ped, Car, MC	5.2
Ada Boost Regres.	Ped, Car, MC	8.1

Table 1: Decision trees and respective miss-classification rates. Data distribution used: All pedestrians 424 (Ped), 242 Pedestrians crossing (Ped-cr) and 182 non-crossing (Ped-nc). 320 Cars and 130 MC=Motorcycles.

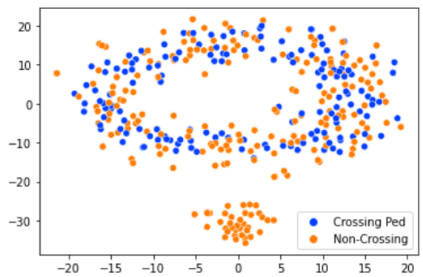


Figure 7: tSNE clustering results. 242 pedestrians crossing street vs 182 non-crossing. As expected, due to short (40ms) reporting the non-crossing are miss-identified.

a classifier to distinguish the pedestrians crossing the road from those not crossing (see also clustering Fig. **7**) was the most difficult task. That however was expected since we essentially requested an impossible task from the classifier. The fraction of

pedestrians who eventually cross the street do so only for a brief part of their motion and until that moment they would be indistinguishable from pedestrians intending not to cross the street. Thus the classifier, based on the information it receives, correctly identifies those pedestrians as part of the non-crossing group. It is therefore not surprising that the classifier was not able to detect that particular sub-group of UEs based on the short (40 ms) history of the training sequence provided.

6. Conclusions and Future Work

We identified ways of classifying UEs in dynamic millimeter wave scenarios by employing conventional ML techniques on CSI-RS measurements. This initial study in clustering and classification of UEs based on their network measurement reports alone, without considering physical positioning or other supporting information, shows that it is possible to infer the mobility mode of the UEs with some success. The work presented here offers insights towards new beam prediction mechanisms in mobility-aware MIMO scenarios. These results together with trajectory forecasting (future research) could in turn provide useful information when preparing hand-over and expected resource demands in order to ease and avoid operational bottlenecks.

References

[1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166-1199, Jul. 2021.

[2] M. Shafi, J. Zhang, H. Tataria, A. F. Molisch, S. Sun, T. S. Rappaport, F. Tufvesson, S. Wu, and K. Kitao, "Microwave vs. millimeter-wave propagation channels: Key differences and impact on 5G cellular systems," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 14-20, Dec. 2018.

[3] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder,"5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201-1221, Apr. 2017.

[4] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data", *IEEE Transactions on Communications*, vol. 14, no. 7, pp. 3899-3911. Jul. 2015.

[5] X. Li, K. Batstone, K. Åstrom, M. Oskarsson, C. Gustafson, and F. Tufvesson, "Robust phase-based positioning using massive MIMO with limited bandwidth", *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct. 2017.

[6] J. Vieira, E. Leitinger, M. Sarajlic, X. Li, and F. Tufvesson, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," *IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct. 2017.

[7] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using sub-6 GHz channels," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5504-5518, Sep. 2020.

[8] D. Burghal, R. Wang, and A. F. Molisch, "Band assignment in dual band Systems: A learning-based approach," *2018 IEEE Military Communications Conference (MILCOM)*, Oct. 2018, pp. 7-13.

[9] D. Burghal, A. Ravi, V. Rao, A. Alghafis, and A. F. Molisch, "A comprehensive survey of machine learning based localization with wireless signals," *arXiv:2012.11171*, Dec. 2020.

[10] B. Ekman, "Machine learning for beam based mobility optimization in NR", *Master of Science Thesis*, Department of Electrical Engineering, Linköping University, Sweden, 2017.

[11] H. Patel, "Beam refinement and beam tracking using machine learning techniques in 5G NR RAN," *Master of Science Thesis*, Faculty of Computing, Blekinge Institute of Technology, Sweden, 2021.

[12] A. Courville, R. Fergus and C. Manning, "Accelerating t-SNE using tree-based algorithms," *Journal of Machine Learning Research*, vol. 15, no. 93, pp. 3221-3245, Oct. 2014.

[13] 3GPP TR 38.104, "Base Station (BS) radio transmission and reception", *Third Generation Partnership Project (3GPP)*, Dec. 2020.

[14] 3GPP TR 36.873, "Study on 3D channel model for LTE", *Third Generation Partnership Project (3GPP)*, Nov. 2017.

[15] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies", *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173-196, Sep. 2018.

[16] 3GPP TR 38.802, "Study on new radio access technology - physical layer aspects", *Third Generation Partnership Project (3GPP)*, Nov. 2017.

[17] 3GPP TR 36.133, "5G NR – Requirements for support of radio resource management", *Third Generation Partnership Project (3GPP)*, Dec. 2020.

[18] T. Kuber, "Learning and prediction using radio and non-radio attributes in wireless systems," *Ph.D. Thesis*, Rutgers University, USA, 2021.

[19] S. Qiu, D. Gesbert, and T. Jiang, "Enabling covariance-based feedback in massive MIMO: A user classification approach," *52nd Asilomar Conference on Signals, Systems, and Computers*, Oct. 2018.

[20] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 3221-3245, Nov. 1987.

Paper II

Paper II

Reproduced, with permission from IEEE

A. Ráth, D. Pjanić, B. Bernhardsson, and F. Tufvesson, "ML-Enabled Outdoor User Positioning in 5G NR Systems via Uplink SRS Channel Estimates," *IEEE International Conference on Communications*, May 2023, Rome, Italy, doi: 10.1109/ICC45041.2023.10279249.

ML-Enabled Outdoor User Positioning in 5G NR Systems via Uplink SRS Channel Estimates

Andre Ráth[1], Dino Pjanić[1,2],
Bo Bernhardsson[3], Fredrik Tufvesson[1],
[1] Dept. of Electrical and Information Technology, Lund University, Sweden
[2] Ericsson AB, Sweden
[3] Dept. of Automatic Control, Lund University, Sweden

Abstract

Cellular user positioning is a promising service provided by Fifth Generation New Radio (5G NR) networks. Besides, Machine Learning (ML) techniques are foreseen to become an integrated part of 5G NR systems improving radio performance and reducing complexity. In this paper, we investigate ML techniques for positioning using 5G NR fingerprints consisting of uplink channel estimates from the physical layer channel. We show that it is possible to use Sounding Reference Signals (SRS) channel fingerprints to provide sufficient data to infer user position. Furthermore, we show that small fully-connected moderately Deep Neural Networks, even when applied to very sparse SRS data, can achieve successful outdoor user positioning with meter-level accuracy in a commercial 5G environment.

Index Terms

5G, beamforming, deep neural network, machine learning, positioning, sounding reference signal, radio access network, localization

I. Introduction

For many years, User Equipment (UE) positioning has been accomplished with Global Navigation Satellite Systems (GNSS), assisted by cellular networks. Besides aiming to achieve reliable and low-latency wireless connectivity, high-accuracy positioning enabled through 5G could coexist and complement existing GNSS-based systems on 5G-capable smart devices. However, GNSS technology is based on unicast transmission and user position is not directly accessible by cellular networks. The latest features within 5G

This work is partially sponsored by the Swedish Foundation for Strategic Research and Ericsson AB.

beam forming technologies drive a distinctive need to acquire accurate user location via radio access interface for location-dependent network functionalities such as beam forming algorithms etc. It is expected that in dense urban area deployments, sub-meter mean positioning accuracy can be achieved [1], [2]. New 3GPP releases are expected to further specify methods for sub-meter accuracy [3]. A range of positioning methods, both downlink (DL)-based and uplink (UL)-based, are used. For radio-based positioning, there is typically a need for specific signals on which a receiver can measure/estimate channel characteristics of interest. This is often expressed as *channel sounding*. Channel State Information (CSI) for the operation of massive multi-antenna schemes can be obtained by the feedback of CSI reports. In a TDD system, the UL channel can be estimated based on SRS transmitted from each UE for which the base station (BS) estimates the DL channel by exploiting channel reciprocity [4], [5]. UL channel estimation includes estimating the Time of Arrival (ToA), the received power, and the Angle of Arrival (AoA) - all being parameters from which the position of the User Equipment (UE) can be estimated. As defined in 3GPP [5], the SRS is a UL Orthogonal Frequency Division Multiplexing (OFDM) symbol with a Zadoff–Chu sequence on its subcarriers, known by both the UE and BS.

Positioning by radio signals is enabled through methods such as fingerprinting or model-based estimation using signal features [6]. The multi-path information of the environment is embedded in the CSI data, and hence the CSI can be used to characterize the radio environment. Examples of CSI-based indoor positioning were presented in [7], [8] while researchers in [9] and [10] demonstrated UE positioning via beam information from Reference Signal Received Power (RSRP). The work presented in [11] shows that the statistics of the wireless channel in Long Term Evolution (LTE) can be used to create a positioning solution even in non Line-of-Sight (NLoS) conditions through an azimuthal-delay representation of the wireless channel. Another LTE DL reference signal-based approach [12] demonstrates that multipath effects can be utilized advantageously to estimate not only user position but also orientation through wireless fingerprinting. In related literature, spatial fingerprinting in conjunction with classical machine learning (ML) methods enables UE localization via learned features of the environment [13]. Recent positioning-related results in [14], [15], also applicable to mmWave networks, target localization accuracy in cases where either the network is optimized for positioning applications or the positioning algorithm is tailored to the particular network geometry. One of the few studies exploring a UL-based method is [16] where simulated UL SRS channel estimates are utilized to investigate the feasibility of SRS estimates for 3D positioning based on joint angle-time estimation and expectation-maximization. Another UL-based method was presented in [17], where indoor positioning through simulated UL SRS signals in LTE-FDD was presented. While the vast majority of the studies above rely on various DL-based

methods for user positioning, we opt to demonstrate a novel ML-powered method using UL channel estimates from SRS transmissions generated in a real-world 5G base station (gNodeB). To the best of the authors' knowledge, there is no prior work on this matter. The main contributions of this paper can be summarized as:

We demonstrate that UL SRS-obtained channel estimation in the BS provides sufficient information to regress for UE position through Deep Neural Networks (DNN). In this study, we consider sparsity to be defined as using channel estimate information from only a small fraction of the available Physical Resource Blocks (PRBs). Data sparsity enables minimal data processing overhead and the use of DNNs that are both low-power and moderately shallow, which reduce the risk of causing potential delays and capacity overloads, necessary for real-time Layer 1 processing where decisions must be made in milliseconds.

Furthermore, in contrast to the majority of prior studies, we prove the viability of SRSs collected in a real commercial 5G NR network setup instead of a simulated environment or non-commercial setup. From a technology perspective, ML is about improving network decision-making capability and allowing it to learn from patterns [19]. The latter urges designing ML-powered methods for real-time operations with the capability to solve complex and unstructured problems using data collected at L1-L2 interaction between BS and UE. Decisions need to be made near where data is generated [20].

II. Data Collection And Methodology

To establish an ML-driven proof-of-concept (POC) for positioning using channel estimates from UL physical layer (L1) channel SRSs, we employ a commercial-grade 5G radio system compliant with 5G NR 3GPP 38.104 Rel15 [18]. A commercial-grade Phased Array Antenna Module (PAAM) is utilized in a BS operating at the center frequency of 3.85 GHz with a 100 MHz bandwidth. We used a proprietary 5G-capable, Android-based test UE, with user motion in different mobility patterns at a distance of approximately 70 m from the roof-top antenna. We opt to extract the channel estimates from the BS baseband unit, which processes the time-varying SRS reports as per the SRS feedback loop structure depicted in Fig. 1. The general thought behind positioning with UL channel estimates is that a physical location under similar network conditions roughly corresponds to a specific SRS-generated channel matrix estimate. In other words, different locations in space have distinct channel fingerprints. Continuous data collection was specifically chosen for this study to mirror realistic navigation conditions. The SRSs are designed to cover the full bandwidth, where the resource elements are spread across the different symbols to cover all subcarriers. In the proprietary baseband hardware unit, the internal beam-space representation of the channel can be extracted and post-processed after ensuring that the UE had high data-rate signalling throughout the channel measurements through 4K video streaming.

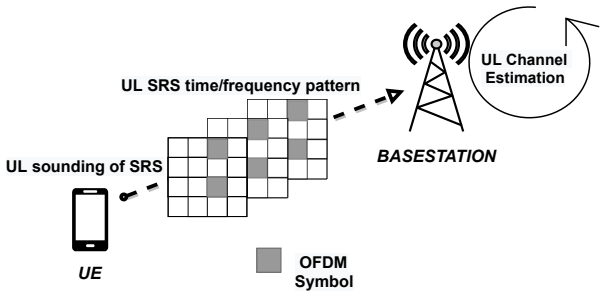


Fig. 1. UL SRS transmission from a UE; The BS obtains Sounding Reference Signals (SRS) containing channel information data from the UE. The SRS is designed to cover the full bandwidth, where the resource elements are spread across the different symbols to cover all sub-carriers. Therefore, SRS is designed with a comb-based pattern.

The SRS-derived channel estimates are stored in a complex-valued matrix structure, which henceforth is referred to as a *channel matrix*. For every SRS sampled from a specific UE, the BS channel matrix contains channel estimates of 64 directional BS antenna elements (directions) for each UE antenna and PRB container. The test UE supports a 1/2/4-antenna configuration, of which the 4-antenna configuration was used during our testing. Furthermore, the 100 MHz Time-Division Multiplexing (TDD) configuration supports 273 PRBs, which in the BS are then allocated to containers with a configurable number of 2, 4, or 8 PRBs per container. In our setup, 2 PRBs were enabled per container. Therefore, there were 137 frequency channels configured, each containing two adjacent PRBs. The channel matrices retrievable from the BS thus contained one complex value/antenna direction for every PRB container and UE/BS antenna pair. In summary, the retrievable channel estimate $\mathbf{H}$ from our experimental setup consists of a complex-valued matrix with up to 137 frequency channels, 64 BS antennas and 4 UE antennas. Together, the upper limit for data extraction in our experimental setup consists of 35072 complex values per SRS transmission.

$$Max\left[\mathrm{Cap}\left(\mathbf{H}\right)\right] = Max\left[N_{Ch}N_{tx}N_{Dir}\right] = 137 \cdot 4 \cdot 64, \tag{1}$$

where N_{Ch} is the number of channels, N_{tx} is the number of UE antennas and N_{Dir} number of BS antenna elements. With 35072 complex values extracted potentially every few milliseconds, the internal data amount handled becomes a major concern. Due to data rate constraints during data logging, we first aimed to explore how a small data amount is

sufficient for meter-level positioning. Only 3 containers with 2 PRBs each were retrieved, hence 768 input features, or in other words, 12 of what we term *sub-channel matrices*, denoted as $\mathbf{H}_{8\times 8}$. Using only a sparse 768 of the potential 35072 values does not prevent positional information from being extractable from the SRS-obtained channel estimate. As shown later in the paper, 768 features still present a unique opportunity for ML frameworks to learn and later regress for the UE position. Furthermore, since SRS measurements are periodic for a given 5G NR waveform numerology, they present an ideal opportunity for ML frameworks to utilize for UE localization. The resulting single-measurement data matrix had a dimensionality as follows:

$$\mathbf{H}_{N_{Ch}\times N_{TX}\times N_{\psi,h}\times N_{\psi,v}} = \mathbf{H}_{3,4,8,8}. \tag{2}$$

Channel matrices may also be expressed as in (3),

$$\mathbf{H}_{(N_{Ch}\cdot N_{TX})\times N_{\psi,h}\times N_{\psi,v}} = \mathbf{H}_{12,8,8}. \tag{3}$$

To utilize the entire 100 MHz bandwidth, the three PRB containers chosen were the lowest, middle, and highest sub-channels. This represents solely 2 $\%$ of the total number of possible sub-carriers. The final logging aspect is that of time. In this case, two timestamps are logged: the frame number corresponding to the actual network time, and the UTC time-stamp corresponding to the time a given SRS dataset was collected and written to the log, used to regress UTC-timestamped GNSS position to channel fingerprints. To decide the geographic area to conduct the measurement campaign in, a few aspects were considered. First and foremost, as the training process is based on GNSS data in our setup, the GPS signal had to be preferably unobstructed throughout the route. To investigate the validity of the proposed ML approach, both LoS and NLoS scenarios were investigated. Three predefined routes were used as the baseline for positioning: a square-shaped area of the dense walk for the training set with a natural random-walk validation and test dataset, an LoS path training set for positioning in the larger area along a predictable path, and an NLoS path data nearby to compare LoS and NLoS effects, as depicted in Fig. 2. Both the rooftop LoS scenario and the ground-level NLoS scenario, respectively henceforth referred to as LoS-A and NLoS-A. The square-shaped area of the dense walk data will be termed LoS-D. The UE moved at the standardized pedestrian velocity of 3 km/h in all the scenarios. We remark here that our target was to study pedestrian velocities, velocities higher than pedestrian ones were not within the direct scope of this study hence conclusions made here shall not be extrapolated to those.

We collected three distinct datasets for each scenario: training, validation, and test.

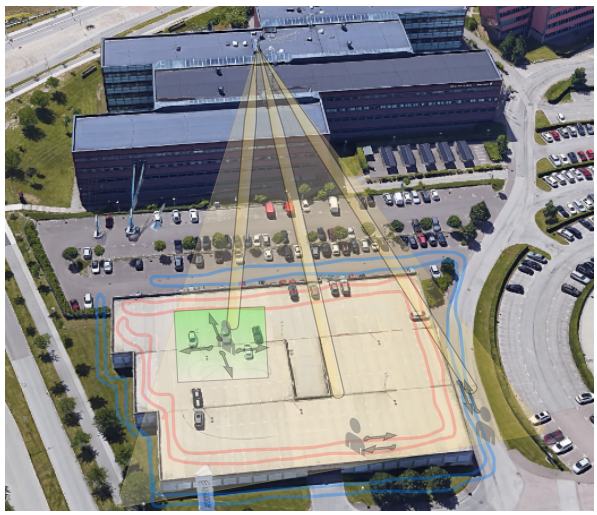


Fig. 2. The pre-defined measurement routes in a SU-MIMO scenario: A 2-story, 10 m high garage building where the red line on the top of the building is the LoS route. The blue line is representing the ground-level route where the surrounding buildings block and reflect the signal from a 20 m high rooftop antenna causing NLoS propagation. The vivid green square shows the region for the LoS dense-walk route. The general coverage area is illustrated in light yellow color whereas yellow-colored narrow beams were generated by the BS equipped with a 64-antenna element array.

These datasets were collected in different acquisition sessions with identical measurement setups. In the LoS-D scenario, the training, validation and test dataset collection happened on the same day; for the LoS-A scenario the training dataset collection was done a week before the test and validation dataset collection, which happened on the same day. For the NLoS-A datasets, the training and validation datasets were collected on separate days of the same week, while the test dataset was collected a month later. We would like to emphasize that the test datasets were not touched by either the model or the data processing pipeline before results were evaluated, and were *not* factored in during model selection either. However, manual data analysis and processing were conducted by the authors on the test data *before* model evaluation to examine data validity.

III. Data Processing Pipeline

A. SRS Data

With the SRS-derived raw dataset obtained from the baseband module in BS, the next step was retrieving the $\mathbf{H}_{12\times8\times8}$ matrices from the data logs. The SRS-derived channel

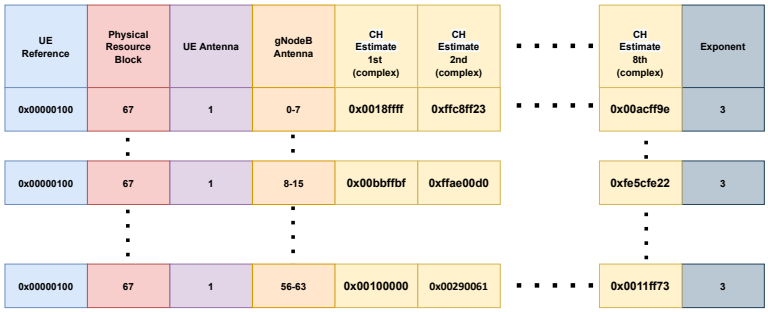


Fig. 3. The format of a single dataset instance, containing the SRS channel estimate for a single frequency channel and UE antenna.

estimate dataset is generated per SRS measurement occasion, meaning per cell, symbol, and UE antenna. It is stored in 1-12 subchannel/UE antenna pair order for all BS antenna directions, down to millisecond intervals. Channel estimates are represented as 4 hex digits for the real and the imaginary components. A representation of the dataset format can be seen in Fig. 3. We consider the System Frame Numbers (SFN) numbers from a repeating sequence of 0 to 1023 throughout the measurement. The series of channel matrices $\mathbf{H}_{12\times8\times8}[SFN]$ are then stored in a 4D matrix $[\mathbf{H}_{12\times8\times8}]_{N_{data}}$, where N_{data} is the number of unique SFN during which at least one sub-channel matrix was measured. The first step in feature selection is then separating the phase and amplitude of the complex channel matrices.

1) Phase component: Comparing the phase in the extracted data to the phase in the raw channel matrix in our experimental setup, we find that the gNodeB's built-in beam-domain transformation uses the received signal phase. The ML system then obtains data with the angle of departure from the gNodeB already utilized. The remaining information in the phase of the complex numbers in the ML input data is discarded in this work due to its dynamic nature.

2) Amplitude component: The extracted data amplitude should contain meter-scale positional information arising from large-scale fading. An underlying assumption is that in both LoS and NLoS cases, the amplitude transfer function of a radio signal depends on environmental geometry, with amplitude thereby acting as a slowly-varying correlate to a position. This will then be visible in the extracted data from the BS - the periodicity of position as the path is walked back-and-forth on is expected to result in a similar periodicity in the recorded complex amplitude. On the NLoS-A dataset, for example, there are 5 back-

and-forth cycles on the NLoS path seen in Fig.2. The expectation is then that certain outputs will have very visible periodicity, e.g. as confirmed in Fig.4. This becomes more clear when the amplitude for all the outputs in a single sub-channel matrix is examined, as per Fig.5.

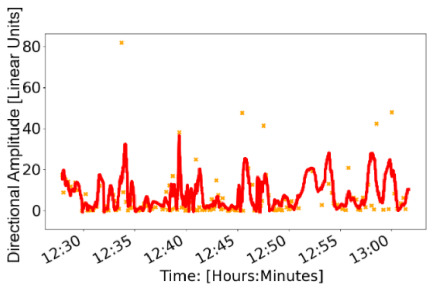


Fig. 4. A snapshot of the complex amplitude output of a single direction $h_{[8,5]}$ in a sub-channel matrix $\mathbf{H}_i$ of the NLoS-A1 database. The red curve shows the amplitude smoothed over 100 samples.

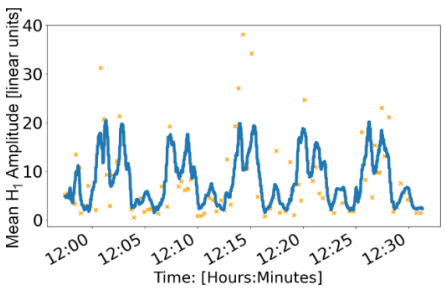


Fig. 5. The mean complex amplitude of all the outputs in a sub-channel matrix $\mathbf{H}_i$ of the NLoS-A2 database. The amplitude of the outputs varies over 5 periods with the periodicity expected from the path dataset. The blue curve shows the moving average of the amplitude over 100 samples.

B. GNSS Data

A commercial UE was used to record GNSS data with an open-source android app to interface with the Android GNSS API. The app obtained GNSS-INS (Inertial Navigation System *Position Navigation Timing* estimates at a 1 Hz sample rate. The UE model was running OxygenOS 11 with Dual-band Multi-Constellation GNSS rated at 3.5 ± 0.5 (m) horizontal accuracy.

C. Combined Data

To use the extracted complex-amplitudes of the sparse channel matrices $[\mathbf{H}_{12\times8\times8}]_{N_{data}}$ as input for position regression with ML, further processing is required. As not all sub-channel matrices are updated during SRS transmission, the missing channel estimate values for any given sample time must be somehow represented for the DNN. Filling in the missing channel-matrix values for any given $\mathbf{H}[\tau]$ extracted channel matrix at sample-time 'τ' is the first step in our data pre-processing pipeline. The simplest method for filling in missing data without using future values or known priors is forward-filling the latest known values. Forward-filling for the channel matrix $\mathbf{H}$ is visualized in Fig. 6. During our measurements, on average 6 of the 12 channel sub-matrices were refreshed every SRS sample, and we observed a variable delay between samples of approx. 35-110 ms, with a higher sample rate in NLoS conditions and a lower sample rate in ideal LoS conditions. This refresh rate was high-enough that unless a connection drop is observed, most sub-channel matrices only persisted for under half a second. For data normalization in this study, only linear scaling was utilized, with improvements in this area left for future work. This was done using min-0 max-1 scaling of the datasets by simple division. The normalization factor was determined by obtaining the maximum amplitude present in the training data, thereby preventing the contamination of the validation and test datasets. Furthermore, we found that taking the square root and fourth root of the channel matrices substantially improved validation positioning results. The exact cause of this performance improvement is unclear and may be the topic of future investigation. However, the maximal benefit was achieved for NLoS scenarios when both square- and fourth root of the input data was used. For the

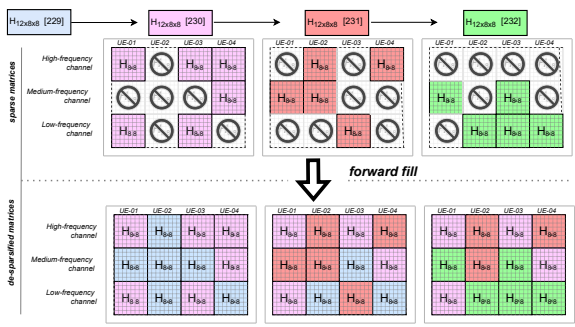


Fig. 6. Using forward-filling on channel matrices $\mathbf{H}_{12\times8\times8}[SFN]$. For any sub-channel matrix $\mathbf{H}$ at SRS sample-time τ, if a value is not given by the current SRS then the most recent known value for that sub-channel matrix is used instead.

LoS scenarios, the fourth root was of unclear benefit.

The square-root and fourth-root concatenation came at the cost of doubling the number of input parameters to the network, to 1524 in total. However, this many parameters as input for a fully-connected network could lead to overfitting. From the assumption that location is mostly independent of UE orientation, reducing the number of input parameters can be achieved with low performance penalty by only taking one $\mathbf{H}_i$ sub-channel-matrix-equivalent as input for every sampled frequency channel. With an eye for future scalability w.r.t. different configured UE antenna numbers, this would also enable ML systems input parameter counts to be independent of different UE antenna configurations and models.

For this reason, we take the average per direction of the sub-channel-matrices $\mathbf{H}_i$ belonging to the same frequency channel. This enables dimensionality reduction without losing frequency-channel information: (4).

$$\mathbf{H}_{3\times 8\times 8}^{mUE} = \frac{1}{N_{TX}} \sum_{i=1}^{N_{TX}} \mathbf{H}_{N_{Ch}\times N_{\psi,h}\times N_{\psi,v}}^{\mathrm{UE_{antenna}}=i} \tag{4}$$

To assign positioning *ground truths* to the channel matrix data $\mathbf{H}_{3\times 8\times 8}^{mUE}$, the UTC timestamp of both the SRS Channel matrix data and the UE position output is used. First, the two datasets are synchronized. Linear time-interpolation from the GNSS-position data is used to create interpolated trajectories, through which the 'ground truth' P_{EN} coordinate pairs for each channel matrix is generated, which can be converted to local P_{XY} coordinates. Finally, all channel matrices that fall outside the bounds of the GPS measurement are discarded. The

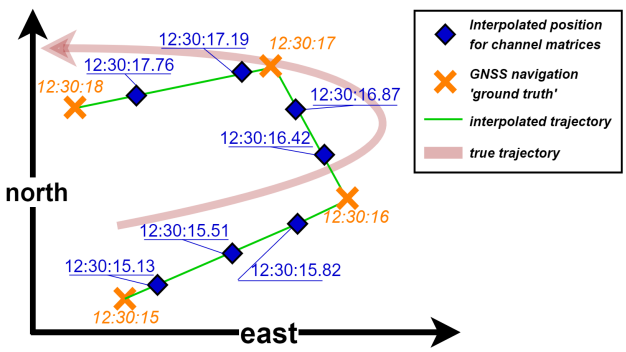


Fig. 7. Assigning position to channel matrices $\mathbf{H}_{12\times 8\times 8}[SFN]$ using shared UTC timestamps with the GNSS dataset and simple linear interpolation. Also shown is the GNSS-based navigation position fix deviation from the 'true' pedestrian trajectory.

position interpolation process is shown in Fig. 7. GNSS inaccuracy is partially modelled during the training process by injecting Gaussian noise of similar magnitude as the GNSS measurement onto the training data P_{train} every epoch during the DNN training process. This also functions as output regularization.

IV. Proposed Machine Learning Framework

In this section, we describe the network architecture used for learning and discuss some design aspects. The proposed system architecture is illustrated in Fig. 8, where the fully connected Deep Neural Network (DNN) uses features from the SRS dataset as input to regress for local P_{XY} position. To demonstrate a real-time POC ML-driven positioning

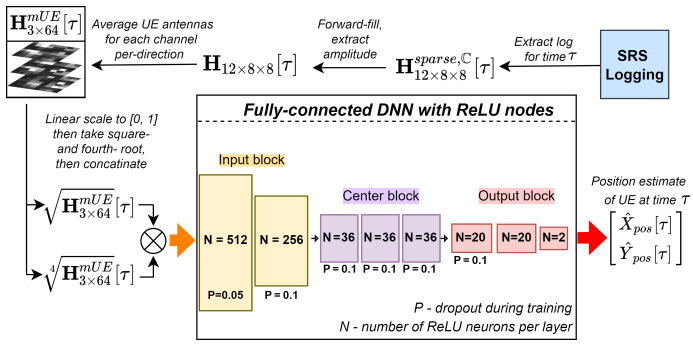


Fig. 8. The 3-block fully connected DNN used in this study along with the input pipeline and intended output. The per-layer dropout used while training is also shown.

with minimal computing overhead, only small moderately deep fully-connected DNNs were tested. Architectures with up to 15 layers at a maximum of 128 artificial neurons (ANs) per layer after the input and first hidden layers were non-comprehensively searched. Of the tested DNNs, the best-performing architecture on the validation data was selected, with no tuning or selection done using test data. Network architecture was unchanged between the ML models for the different datasets. For the final hyperparameter search, three discrete blocks of fully-connected layers were defined, each with the rectified linear unit (ReLU) activation functions and varying AN and layer counts within a limited range.

1) **Input block:** The *input block* serves to take the input data through subsequent shrinking layers into a parameter bottleneck, compressing the data.

2) **Center block:** The expectation is that the process block takes the reduced dimensions from the input block and feeds it through identical fully-connected layers, processing the lower-dimensional representation further.
3) **Positioning block:** The expectation is that the positioning block takes the center block's output and finally narrows it down to two dimensions to regress for a position. Note that the last layer has two outputs corresponding to a position's local X and Y coordinate pair (P_{XY}).

This design choice originated from our testing where introducing a bottleneck of 20-40 fully-connected ANs per layer for all but the first two layers reduced overfitting while having a wide input and first hidden layer improved general performance. The overall number of hidden layers was also kept low, as increasing layer counts over 7 did not discernibly affect validation performance.

Finally, the minimized loss for the network is the Mean Euclidean Distance Loss (MEDL), which can be expressed as:

$$\text{MEDL} = \frac{1}{N_s}\sum_{\tau=1}^{N_s}\|P_{XY}[\tau] - f_\theta\left(\mathbf{H}^{mUE}[\tau]\right)\|_1, \tag{5}$$

where f_θ is the ML model with θ optimizable parameters, $P_{XY}[\tau]$ the interpolated GNSS coordinates for sample-time τ, and N_s is the number of time-samples in a batch. The loss was minimized with an ADAM [21] optimizer using PyTorch on a CUDA-capable GTX 2060.

V. Results And Discussion

To summarize our results shown in Table I, we obtain an approximate mean euclidean distance of 3-9 m as compared to the GNSS data when evaluated on test data, with accuracy depending mostly on data conditions. We note again that model selection or parameter optimization was not done on test data. During validation, it became apparent that data character changed between the training datasets in LoS-A and the validation datasets, potentially explaining the degraded performance compared to LoS-D, where no domain change was observed.

TABLE I
Mean euclidean error in meters for each dataset

dataset name	validation dataset	test dataset
LoS-D	2.8 (m)	3.3 (m)
LoS-A	9.2 (m)	9.7 (m)
NLoS-A	7.3 (m)	8.1 (m)

These results compare favourably to results found in the literature on most outdoor positioning systems using similar DL/UL-based positioning approaches and density real-world data. The NLoS data accuracy indicates this method's viability for positioning in a real-world environment. We note that the precise effects on the SRS channel matrices of non-pedestrian tracking at high velocity and, e.g., users in a vehicle have not been tested. As an example, forward-filling introduces data from previous sample times. For scenarios where significant distance may be travelled between SRS samples, alternatives to forward-filling might be needed e.g. using only the latest SRS as a partial data point. As an extension of this work, we show a proof-of-concept in [22] where the same datasets and ML pipeline introduced in this paper are extended with simulating pedestrian motion through particle filtering, improving mean accuracy to around 5-6 m for NLoS scenarios.

To summarize, this study demonstrates the practical viability of UL SRS channel estimates in a realistic outdoor NLoS propagation environment. In contrast to other studies employing multi-antenna arrays at the receiver side, we use a commercial-grade, 4-antenna-equipped UE.

VI. Conclusions And Future Work

We have shown some of the potentials of DNNs for outdoor user positioning in 5G NR systems using UL SRS channel estimates in a very sparse data sampling regime. The results presented show sub-10 m of mean accuracy for all test scenarios, despite an already inherent ground-truth horizontal positioning inaccuracy of 3.5 ± 0.5 m in the GNSS dataset. A more accurate GNSS positioning setup for training data should substantially improve results. Similarly, a higher SRS sampling rate should also improve the positioning results significantly. For future research, the phase of the SRS channel estimates could be a possible feature source to explore. Finally, considering the simplicity of the DNN model we used could also be interesting, as more sophisticated models may further improve accuracy.

References

[1] H. Wymeersch, G. Seco-Granados, G. Destino, D. Dardari and F. Tufvesson, "5G mmWave Positioning for Vehicular Networks," IEEE Wireless Communications, vol. 24, no. 6, pp. 80-86, Dec. 2017.

[2] S. Dwivedi *et al.*, "Positioning in 5G Networks," IEEE Communications Magazine, vol. 59, no. 11, pp. 38-44, Nov. 2021.

[3] R. Keating, M. Säily; J. Hulkkonen and J. Karjalainen, "Overview of Positioning in 5G New Radio," 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland, pp. 320-324, Aug. 2019.

[4] T. L. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," IEEE Trans. on Wireless Communications, vol. 9, Issue: 11, pp. 3590-3600, Nov. 2010.

[5] 3GPP TR 38.211, "5G NR; Physical Channels and Modulation," Third Generation Partnership Project (3GPP), Dec. 2020.

[6] M. M. Butt, A. Rao and D. Yoon, “RF Fingerprinting and Deep Learning Assisted UE Positioning in 5G,” 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, pp. 1-7, May 2020.

[7] G. -S. Wu and P. -H. Tseng, “A Deep Neural Network-Based Indoor Positioning Method using Channel State Information,” 2018 International Conference on Computing, Networking and Communications (ICNC), Maui, HI, USA, pp. 290-294, 2018.

[8] X. Wang, X. Wange, and S. Mao, “Indoor Fingerprinting With Bimodal CSI Tensors: A Deep Residual Sharing Learning Approach,” IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4498-4513, March 2021.

[9] E. Rastorgueva-Foi et al.,“Networking and Positioning Co-Design in Multi-Connectivity Industrial mmW Systems,” IEEE Trans. on Vehicular Technology, vol. 69, no. 12, pp. 15842-15856, Dec. 2020.

[10] M. M. Butt, A. Pantelidou and I. Z. Kovács, “ML-Assisted UE Positioning: Performance Analysis and 5G Architecture Enhancements,” IEEE Open Journal of Vehicular Technology, vol. 2, pp. 377-388, Sep. 2021.

[11] R. Whiton, J. Chen, T. Johansson and F. Tufvesson, “Urban Navigation with LTE using a Large Antenna Array and Machine Learning,” IEEE 95th Vehicular Technology Conference (VTC2022-Spring), Helsinki, Finland, pp. 1-5, June 2022.

[12] R. Whiton, J. Chen and F. Tufvesson, “LTE NLOS Navigation and Channel Characterization,” Proceedings of the 35th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2022), Denver, Colorado, pp. 2398-2408, Sept. 2022.

[13] X. Li, K. Batstone, K. Åstrom, M. Oskarsson, C. Gustafson and F. Tufvesson, “Robust phase-based positioning using massive MIMO with limited bandwidth,” IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, pp. 1-7, Oct. 2017.

[14] J. D. Roth, M. Tummala and J. C. McEachen, “Fundamental implications for location accuracy in ultra-dense 5G cellular networks,” IEEE Trans. on Vehicular Technology, vol. 68, no. 2, pp. 1784-1795, Feb. 2019.

[15] J. A. del Peral-Rosado, G. Seco-Granados, S. Kim and J. A. López-Salcedo, “Network design for accurate vehicle localization,” IEEE Trans. on Vehicular Technology, vol. 68, no. 5, pp. 4316-4327, May 2019.

[16] B. Sun, B. Tan, W. Wang and E. S. Lohan, “A Comparative Study of 3D UE Positioning in 5G New Radio with a Single Station,” MDPI Journal, Sensors, vol. 21, no. 4, 1178, Jan. 2021.

[17] F. Wang, J. Chen and Q. Liu, “SRS-based LTE indoor wireless positioning system,” 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, pp. 2356-2359, March 2017.

[18] 3GPP TR 38.104, “Base Station (BS) radio transmission and reception,” Third Generation Partnership Project (3GPP), Dec. 2020.

[19] S. Axelsson, “The how and why of AI: enabling the intelligent network,” Ericsson tech blog, https://www.ericsson.com/en/blog/2019/11/ai-enabling-the-intelligent-network, Nov. 2019.

[20] A. Platek and J. You, “What is the relationship between AI and 5G,” Ericsson tech blog, https://www.ericsson.com/en/blog/2022/1/whats-the-relationship-between-ai-5g, Jan. 2022.

[21] Kingma, Diederik P. and Ba, Jimmy, “Adam: A Method for Stochastic Optimization”, 3rd International Conference on Learning Representations (ICLR), May 2015.

[22] A. Rath “Beamformed Channel Matrix Positioning Using 5G Testbench CSI data With a Deep-Learning Pipeline,” Master’s thesis, Lund University, http://lup.lub.lu.se/student-papers/record/9100498, Sept. 2022.

Paper III

Paper III

Reproduced, with permission from IEEE

D. Pjanić, A. Sopasakis, A. Reial, F. Tufvesson, "Early-Scheduled Handover Preparation in 5G NR Millimeter-Wave Systems," *IEEE Open Journal of The Communications Society*, Vol. 5, Oct. 2024, doi: 10.1109/OJCOMS.2024.3488594.

Early-Scheduled Handover Preparation in 5G NR Millimeter-Wave Systems

Dino Pjanić[1,*], Member, IEEE, Alexandros Sopasakis[1,3,*], Member, IEEE, Andres Reial[1], Senior Member, IEEE, and Fredrik Tufvesson[1], Fellow, IEEE
[1] Department of Electrical and Information Technology, Lund University, Sweden
[2] Ericsson AB, Sweden
[3] Department of Mathematics, Lund University, Sweden

Abstract

The handover (HO) procedure is one of the most critical functions in a cellular network driven by measurements of the user channel of the serving and neighboring cells. The success rate of the entire HO procedure is significantly affected by the preparation stage. As massive Multiple-Input Multiple-Output (MIMO) systems with large antenna arrays allow resolving finer details of channel behavior, we investigate how machine learning can be applied to time series data of beam measurements in the Fifth Generation (5G) New Radio (NR) system to improve the HO procedure. This paper introduces the Early-Scheduled Handover Preparation scheme designed to enhance the robustness and efficiency of the HO procedure, particularly in scenarios involving high mobility and dense small cell deployments. Early-Scheduled Handover Preparation focuses on optimizing the timing of the HO preparation phase by leveraging machine learning techniques to predict the earliest possible trigger points for HO events. We identify a new early trigger for HO preparation and demonstrate how it can beneficially reduce the required time for HO execution reducing channel quality degradation. These insights enable a new HO preparation scheme that offers a novel, user-aware, and proactive HO decision making in MIMO scenarios incorporating mobility.

The work is partially sponsored by the Swedish Foundation for Strategic Research and Ericsson AB, Sweden, as well as by grants from eSSENCE no. 138227, Vinnova 2020-033375, Formas 2022-00757 and Swedish National Space Board. The training and data handling was enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Index Terms

Beam Management, Handover Control Parameters, Measurement Event A3, Handover Preparation, ML, mmWave, Mobility Robustness Optimization.

I. INTRODUCTION

TO ensure seamless user mobility between neighboring cells, the handover (HO) mechanism is defined in the 3GPP specification 38.300 [1], from the First Generation (1G) onward. Reliable communication during the mobility of user equipment (UE) is crucial, and HO management is a key capability [2]. During HO, control messages are exchanged between the UE and the serving Base Station (BS) under predefined conditions. However, since these messages are sent over the air interface, they may be initiated when the radio link faces severe attenuation and various propagation issues such as noise and interference. A robust HO mechanism is essential to maintain user mobility under these conditions; otherwise, user mobility is compromised.

To address these challenges, each generation of cellular networks has refined the HO procedure while maintaining its core functionality, which consists of three phases: *preparation*, *execution*, and *completion*. The preparation phase, as the initial step of the HO procedure, typically occurs when the signal quality of the serving cell is low and interference from neighboring cells is high. This makes the UE exposed to Handover Failure (HOF) and Radio Link Failure (RLF), therefore, among the three phases, HO preparation is the most vulnerable [3].

The existing event-driven 5G HO procedure requires the participation of both UE and BS during its preparation phase. In the initial part of this phase, the UE is primarily responsible for measuring the quality of the channel of the serving and neighboring cells and reporting when a measurement event is fulfilled. More precisely, an offset value and a hysteresis value, jointly called the HO margin (HOM), determine when an *entry criterion* of a measurement event is fulfilled, depicted as Step 2 in Fig. 1, where the intrinsic delay of the Time-to-Trigger (TTT) timer bridges Steps 2 and 3. The HOM is the most significant parameter to control the HO decision [4]. The TTT timer and HOM comprise a tightly coupled setting named HO Control Parameters (HCP) which determine when an HO *event* (HE) is *fulfilled*, depicted as step 3 in Fig. 1, and thus impact the initial timing of an HO preparation phase. For optimal initiation of the HO preparation phase, it is essential to adjust the HO timing to each user's specific mobility pattern and current radio conditions. Fig. 1 also illustrates how the traditional HO preparation mechanism assigns a passive and disadvantageous role to the BS, making it unaware of imminent HO events and thus prone to initiate an HO too late.

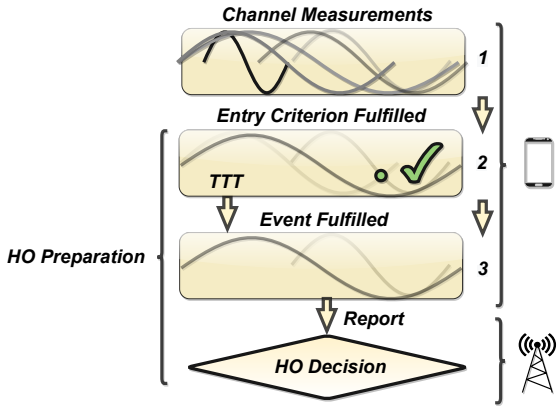


Fig. 1. Interplay between UE and NW during the handover HO preparation phase.

A. A3 Handover Event

In this section, we examine the core components of an HO event-triggered mechanism, as specified in 3GPP 36.133 [4], and clarify how HCPs interact.

1) Handover Control Parameters: The A3 and A5 events, illustrated in Fig. 2 and 3, embody the signal quality of the serving cell and neighboring cells using the reference signal received power (RSRP) metric. Event A5 provides a handover triggering mechanism based on *absolute* measurement results. Only the A3 event evaluates a *relative* comparison between the signal quality of the serving cell and that of neighboring cells, making it adaptable to varying network conditions. As we focus on an intra-frequency HO scenario, we chose the most widely used A3 event whose entry criterion fulfillment, hereafter referred

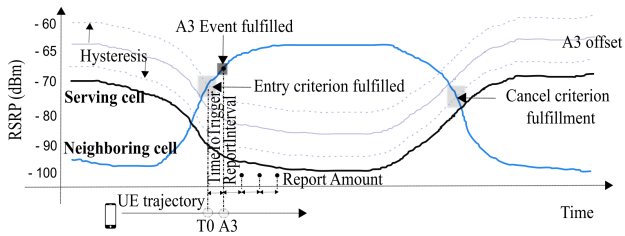


Fig. 2. A3 Event. The quality of neighboring cells exceeds the quality of the serving cell by an offset value. A3 event entry criterion fulfillment (T0) throughout the TTT duration (A3).

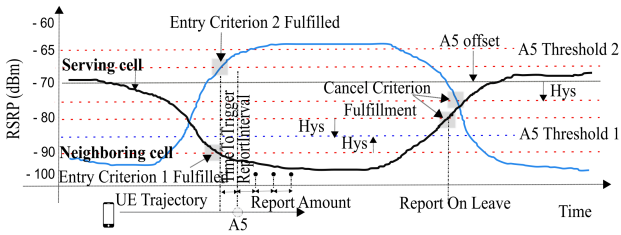


Fig. 3. A5 Event. Only when both entry criteria are satisfied, the UE reports event A5 to gNB.

to as **T0**, is given by the inequality (1) where $RSRP_{Target}$ and $RSRP_{Serving}$ are long-term averaged Layer3-filtered measurements from the serving and neighboring cells, respectively.

$$RSRP_{Target} > RSRP_{Serving} + HOM. \tag{1}$$

2) Handover Margin: Any inappropriate HOM settings between low and high may lead to a ping-pong effect or high Radio Link Failure rate. The HOM setting is set at the cell level, which means that all users within the cell will apply the same HOM. Preferably, the adjustment of the HOM settings shall be adapted individually concerning each user's context such as velocity, mobility pattern etc [5]. Even though the HOM determines the **T0** fulfillment it can not be entirely decoupled from the TTT functionality in the context of event-triggered HO optimizations.

3) Time-To-Trigger: Upon **T0** fulfillment, the UE awaits TTT expiry before reporting HE fulfillment to the BS, hereafter referred to as **A3**, and as illustrated in Fig. 2. The TTT timer has been introduced in previous generations of cellular networks and inherently introduces a time delay. If the TTT value is too large, it may cause connection interruption and an HOF. Conversely, too small a value can prevent long delays but lead to an increased HO ping-pong or unnecessary HO.

Given the discussed background, sub-optimal HCP settings can negatively impact the optimal timing for HE and reduce the overall HO success rate.

B. Related Work

The HCP parameters heavily influence the timing of the HO preparation phase, and numerous techniques have been developed to ensure that the HO is initiated at the most optimal moment. The number of potential HO regions inevitably increases in dynamic mmWave environments characterized by reduced cell coverage and multi-beam architecture requiring smaller cell sizes. An HO region is the distance between the HO event trigger point and the Physical Downlink Control Channel (PDCCH) outage point [6]. The handover failure (HOF) rate is directly proportional to the UE mobility speed and inversely proportional to

the size of the HO region. The HOF rate can be reduced by expanding the hypothetical HO region through careful tuning of HCP parameters, which must account for varying network deployments, cell sizes, user velocities, and mobility patterns.

Unlike previous research, we dissect the event-triggered mechanism and explain how to distinguish it into two chronological occurrences, advancing the timing of the HO preparation phase. Our machine learning (ML)-assisted method decouples these events by predicting the earliest **T0** based on changes in the signal patterns of the UE beam measurements. From a network perspective, our solution claims insights into steps 2-3 illustrated in Fig. 1.

We briefly shed light on the strengths and limitations of two key optimization techniques that represent the most relevant related research, namely Conditional Handover and Mobility Robustness Optimization.

Conditional Handover (Conditional HO), introduced by 3GPP in 5G Release 16 [1] decouples the base station (BS) preparation and HO execution phases, reducing the number of HOFs by allowing the UE to decide when to initiate the HO. Unlike baseline 5G HO schemes, Conditional HO employs early HO preparation to mitigate the risk of a critical signal quality drop between the UE and the BS. The authors of [7] propose an improved conditional HO scheme that uses trajectory prediction to prepare the BSs along the path of the UE. In contrast, [8] explores ways to improve early preparation success by predicting the next BS during Conditional HO. These techniques aim to optimize the timing of the HO preparation phase by shifting the responsibility entirely to the UE. However, Conditional HO introduces significant signaling overhead during the HO preparation phase, particularly in dense cell deployments with high HO frequency [9]. However, the Conditional HO technique has some disadvantages and challenges that must be addressed.

Signalling Overhead: Conditional HO requires the network to pre-configure multiple target cells for a potential handover, which adds complexity to network management.

HO Decision-Making: The decision logic for triggering a handover becomes more complex, as the UE has to monitor multiple candidate cells and decide which one is optimal under changing conditions.

Mobility Robustness Optimization (MRO) approaches fall under the HO self-optimization technique family, which aims to automate HCP settings with minimal human intervention. Approaches include optimizing HCP parameters individually, considering trade-offs, or treating them as a unified entity [11] - [30]. Studies like [31], [32] emphasize the need to adapt HCPs in millimeter-wave (mmWave) deployments with dense small cells. These studies propose algorithms to adjust HCPs based on RSRP and UE velocity, continuously refining these parameters after each measurement report. However, despite improvements in performance metrics, these solutions present notable challenges.

Signalling Overhead: MRO requires ongoing adjustments to HCPs based on real-time network conditions, which can occasionally result in HO failures or unnecessary HOs. Additionally, MRO solutions often rely on generalized approaches that may overlook the specific UE context such as mobility patterns or velocity, leading to suboptimal performance in certain scenarios.
HO Decision-Making: The self-optimization process could lead to either too aggressive or too conservative HO decisions, further contributing to handover failures or an increase in unnecessary handovers.
Inaccurate Handover Predictions: MRO algorithms rely on predictive models to optimize handovers. If user mobility patterns or network conditions change suddenly or unpredictably, the system might make inaccurate predictions.

It is evident that optimizing the timing of the HO preparation phase is a recurring focus in much of the research conducted in this area. In the following sections, we explain how the proposed *Early-Scheduled Handover Preparation* (ESHOP) scheme addresses the advantages above and limitations identified in related research, as well as those within the ESHOP framework itself. Regarding MRO techniques, continuous adjustment of HCPs requires frequent signaling in the downlink via measurement radio resource reconfiguration, which significantly increases power consumption on the UE side [33]. This issue becomes particularly pronounced at high UE velocities and in small cell deployments with frequent HOs. Furthermore, these solutions assume that the UE's velocity is known, a parameter that is typically not tracked by cellular networks. Our solution also relies on dedicated signaling toward the UE and is sensitive to sudden and unpredictable changes in UE mobility patterns. Unlike traditional approaches, however, our solution can learn from measurement data, improving handover robustness even in the face of unexpected events.
When optimizing the HO preparation phase, Conditional HO complicates decision-making by triggering multiple target cells. In contrast, our study employs a different approach to optimize the timing of the HO preparation phase and reduce signaling overhead. Instead of explicitly estimating individual UE paths or velocities using conventional wireless channel modeling, we utilize a technique that associates a series of radio channel measurements with physical locations through channel fingerprinting. These fingerprinted features, based on each user's trajectory and velocity, enable us to analyze the time series of relationships between these variables, forming the foundation of our study. This approach allows the ESHOP scheme to trigger HO preparation in a just-in-time manner.

C. Contributions

- *Predictive Timing*: By accurately predicting the timing of the **T0** fulfillment, the ESHOP scheme allows the network to initiate HO preparation in advance and ensures that the preparation phase begins at the most appropriate time. This proactive approach

contrasts with traditional reactive methods that wait for the subsequent **A3** fulfillment to be met and reported by UE before initiating HO preparation.

- *Enhanced HO Regions*: The proposed ESHOP scheme proactively expands the hypothetical HO region by initiating the preparation phase earlier. This expansion allows for more time to manage the HO process, thereby minimizing the risk of users experiencing signal degradation or loss of connectivity during the HO. This is particularly beneficial in dynamic mmWave environments characterized by small-cell deployments.

- *Dynamic HO Preparation*: The ESHOP scheme dynamically adjusts the timing of the HO preparation phase. This user-centric approach enables flexibility in accommodating rapid changes in the radio environment, thereby enhancing the robustness of the HO process.

II. SIMULATED MODEL SETUP

To demonstrate the primary goal of this investigation, which is the feasibility of using beam measurements for HO predictions, this study employs an extremely simplified mobility model as a proof of concept. We acknowledge that the simulated system does not represent a typical urban 5G network deployment, such as a crowded metropolitan area. Despite its simplicity, the system model effectively presents fundamental HO-related issues in a small-cell network deployment, and the simulations provide a realistic HO procedure using a radio system that supports detailed beam management that is compliant with standardized 5G NR systems at mmWave frequencies [34]. We simulate a commercial-grade Phased Array Antenna Module in a BS operating at a center frequency of 28 GHz over a 100 MHz bandwidth. The 5G NR frame numerology is set to a 120 kHz subcarrier spacing with a slot duration of 125 μs (8000 slots per second). We simulate a three-dimensional area with a single site containing three cells, with the BS deployed in the center of three cells shaped as hexagons, as shown in Fig. 4 (b). Due to the small cell deployment, the coverage extends only over a fraction of each hexagon, since dense cell deployments typically allow limited freedom of movement within a cell. As depicted in Fig. 4 (a), the UE trajectory follows a circular but individually randomized path centered around the BS with a 50-meter radius, modeling a dense small-cell deployment that results in rapid and frequent HOs. The circular trajectory simultaneously presents a variety of instantaneous movement angles in relation to the base station position, emulating a mix of user movement patterns in a practical deployment. UEs appear along the circular trajectory at random starting points, causing variability in the timing of the next HO.

To reveal the necessity of initiating the HO preparation at its earliest stage, we simulate users traveling through multiple cells, spending only a short time within each cell along its trajectory. The circular trajectory ensures that the user crosses all three cell borders during

a single simulation duration, with each crossing occurring at slightly different angles and times, depending on the randomness of the UE's movement along the path. Cell border crossings are influenced by cell coverage overlap and potential HO opportunities. Users move at constant speeds of either 25 or 31 m/s, which significantly alters the dynamics between the BS and the user in terms of rapidly changing channel quality which is

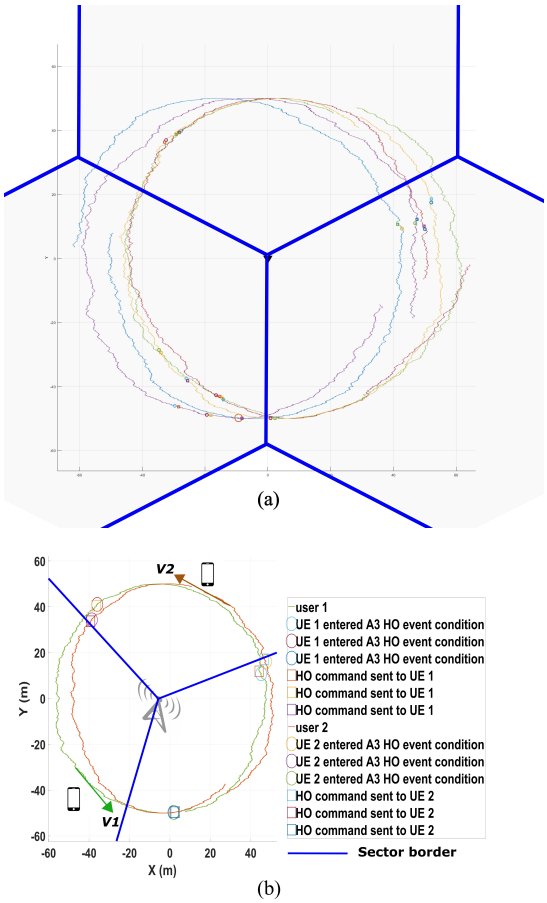


Fig. 4. Top view of random individual trajectories generated by five distinct users (a), 5G NR site deployment with the base station located at the corners of three adjacent cell sites, each shaped as a hexagon and depicted in dark blue (b). Users move along individually randomized circular trajectories at constant velocities of 25 m/s (v1) or 31 m/s (v2).

captured by the measurement reports sampled periodically at constant time intervals of 40 ms. During the simulations, the UEs are in radio resource control connected mode and engage in active UL-centric signaling towards the BS. The lifetime of each UE is recorded from the start to the end of the simulation. Parameter settings were evaluated for different UE loads using the HO-related parameters listed in Table I. Multiple seeds were used to ensure statistical confidence. Both Line-Of-Sight (LOS) and Non LOS propagation scenarios were included, as per the guidelines in [35]. As this study focuses on time predictability using historical RSRP values at standardized mmWave frequencies, bands outside these specified frequencies are beyond the scope of this study, and the conclusions should not be extrapolated to other frequencies.

From a UE perspective, an NR cell is defined by the physical transmission of a specific Synchronization Signal Block (SSB), which contains a physical cell identifier at a particular frequency. The BS transmits SSB beams periodically, with a T_{SS} of 10 ms. These beams are arranged in a grid of 3 in azimuth and 4 in elevation, resulting in $N_{SSB} = 12$ beams, as depicted in Fig. 5. The SSB beams are static and wide, always pointing in the same direction, forming a grid that covers the entire cell area. This makes them suitable for cell-level mobility evaluations. An HO occasion is dynamically defined by the rapidly changing mmWave radio environment conditions, fingerprinted in a time series of SSB beam measurements specific to each UE's movement pattern on UE trajectory, location and velocity. This is particularly notable as individual circular trajectories may be separated by tens of meters, as illustrated in Fig. 4 (a). Demonstrating robust behavior of the method at higher speeds also suggests corresponding or further improved performance at lower speeds, assuming that model is trained with corresponding data patterns at various speeds. As UEs move around the BS, they search for and measure the qualities of these beams, maintaining a set of candidate beams from multiple cells. A combination of a physical cell identifier (cell ID) and beam identifier (beam ID) differentiates beams from each other.

TABLE I. Simulated handover parameters

Parameter	Value
measurementType	A3
timeToTrigger	0.04 (s)
hysteresis	0 and 1 (dBm)
reportInterval	0 (s)
reportAmount	1
offset	3 (dBm)
measurementQuantity	RSRP

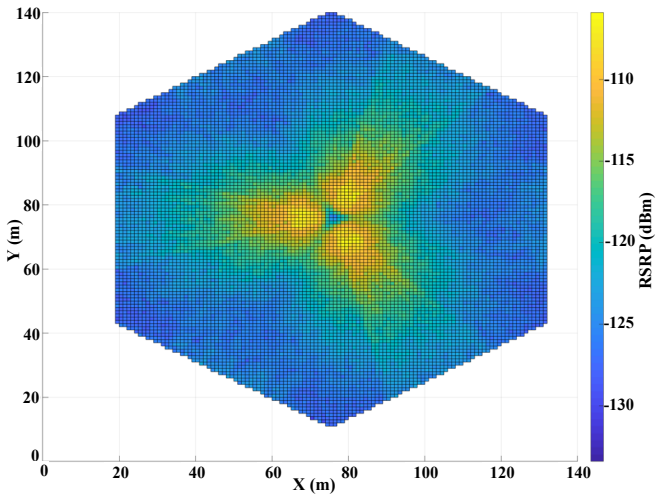


Fig. 5. 2D projection of RSRP distribution generated by the 4x3 SSB wide beam pattern.

III. PROPOSED ESHOP SCHEME

In this section, we examine the incentives underlying the introduction of ESHOP and explain the application area of our TCN-driven model, which offers several benefits worthy of close evaluation. Before the UE triggers an HE, it measures the signals of the serving cell and neighbor cells over the 5G NR air interface evaluating whether any measured signals satisfy the *entry criterion fulfillment* of a HE, **T0**, as highlighted in green in Fig. 6. The ESHOP aims to increase the robustness of the HO preparation phase by utilizing the time window between **T0** and **A3**, where **A3** marks the start of both the potential HO region and the legacy HO preparation phase. These two events are time-bridged by the TTT value, as highlighted in yellow in Fig. 6. The most significant component of the ESHOP scheme is the Time to Entry Criterion Fulfillment (**TEF**), i.e. the remaining time until **T0**. Being a predictive measure, **TEF** estimates the time at which the **T0** will be met allowing the network to initiate HO preparation well in advance, thereby reducing the likelihood of HOFs and improving overall network performance. The ESHOP scheme enables a user context-aware detection of incoming **T0** occasions. It establishes a new starting point for both the potential HO region and the HO preparation phase, scheduling them earlier by exactly the TTT duration. As developed in detail later in Section IV, we aim to contain the

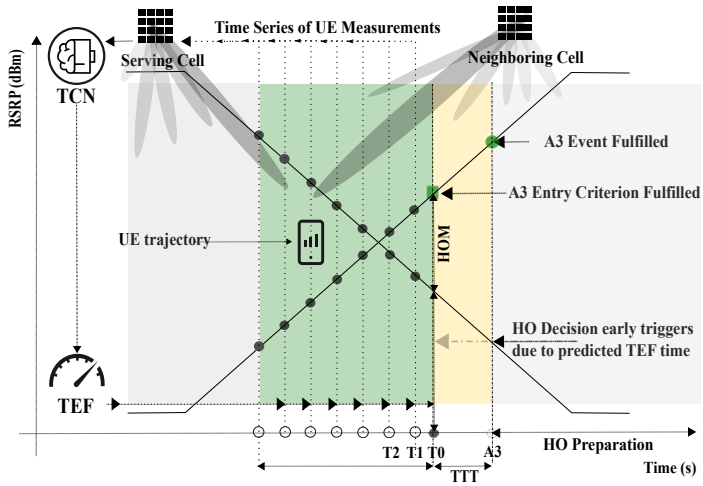


Fig. 6. Principals of ESHOP scheme and its most significant component, Time To Entry Criterion fulfillment (TEF).

HO decision procedure between source and target cells within the TTT window thereby early-triggering the HO preparation phase. The ESHOP relies on the capability of the data set to capture the fingerprinted features based on each user's trajectory and velocity and allows us to study the time series of relationships between these. As shown in Fig. 2 and 6, the HO hysteresis setting, as part of the HOM, impacts the timing of **T0** and pushes it along the time axis and the UE trajectory, e.g, when set negative it triggers HO preparation earlier risking a ping-pong effect. Conversely, a late-triggered HO preparation risks RLF and HOF. An obvious tradeoff between HO failures and ping-pongs is strongly related to the UE velocity whereas an expansion of the potential HO region is a possible solution [37]. Note that in the leading boundary of a traditional HO region, the **A3**, may be pushed in both directions, along the UE's trajectory depending on the hysteresis setting. The same is valid for the new leading boundary **T0** separated from **A3** via the TTT timer. The latter alludes to the essence of the ESHOP scheme and the underlying data set containing fingerprinted information about e.g.; hysteresis setting as well its co-relation with the UE's trajectory and velocity. This fact alleviates the need for additional optimizations of HOM parameters and allows for focusing on utilizing the TTT timer's duration for signalling time savings. Purposely, TTT duration was kept constant throughout this study. As a gNodeB remains unaware of an imminent **A3** occasion in the legacy HO procedure, the ESHOP scheme addresses this by assigning an *active* role to the gNodeB, enabling it to predict the remaining time to **TO** via the ML-inferred **TEF** metric.

IV. ENHANCEMENTS FOR HO ROBUSTNESS

This section demonstrates how the ESHOP scheme can be applied on the network side to enhance the robustness of the 5G HO. The HO procedure involves various levels of internal signaling between network nodes, with intra-gNB communication being the simplest in terms of signaling complexity. Figures 7 and 8 illustrate the general signaling order for both intra-gNB and inter-gNB communication. Upon fulfillment of the A3 event, the HO preparation is triggered. As part of the HO decision, the serving cell exchanges the UE's context with the candidate cells, and based on available resources, the target cell performs admission control to determine if it can accommodate the incoming UE. If admission is granted, an HO acknowledgment is sent to the source node, including an HO command conveyed in the downlink to the UE. As described in Section III, the ML model continuously predicts the remaining time approaching **T0** via **TEF** metric, represented as a countdown. The different stages of the proposed ESHOP scheme for intra-gNB and inter-gNB HO scenarios are illustrated in Figures 7 and 8. In the following, we exemplify the ESHOP scheme depicted in Figure 7; the same principles apply to Figure 8.

- **Step 1**: The BS continuously receives beam measurement reports for the serving cell and neighbor cells for each active UE. With the aid of the incoming measurement reports, the ML model predicts the remaining **TEF** time based on its training experience and evaluates whether to proceed with an early triggering of the HO preparation phase. The incoming HO event criterion fulfillment ideally results in a continuously decreasing **TEF** value, approaching **T0**. Once **T0** is reached, the next step is initiated.
- **Step 2**: An **A3** event is expected to be reported by the UE after TTT expiry. The BS utilizes the TTT duration for early activation of the HO preparation phase, starting with the HO decision-related procedures such as steps 3b, 4, and 5. At this stage the source gNB decides whether to prepare the target cell in advance. If the target cell is prepared, the next step is initiated.
- **Step 3**: Upon TTT expiry, an A3 event is expected to be reported by the UE.
  Note: In case the ESHOP prediction fails suggesting that UE does not report an A3 event fulfillment, the subsequent steps are not executed, and the current ESHOP scheme is aborted, allowing the BS to fall back to the legacy HO preparation procedure.
- **Step 6-8**: Based on already prepared HO decision-related signalling in the previous step, the remaining signalling of the HO procedure is carried out.

The ESHOP demonstrates the capability to mitigate HO failures induced by often sub-optimal HOM settings. Considerable time savings are achieved during the HO preparation phase by performing the most time-consuming signalling during the TTT time window. The remainder of this paper presents the feasible time savings achieved through this approach, paving the way for a higher success rate in the HO preparation phase, especially in the

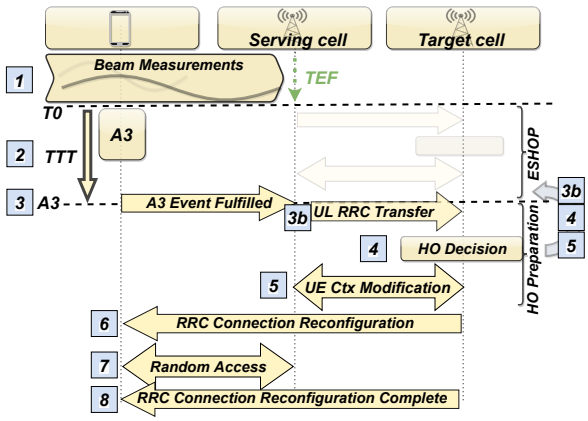


Fig. 7. Legacy intra gNodeB handover procedure and the proposed ESHOP scheme.

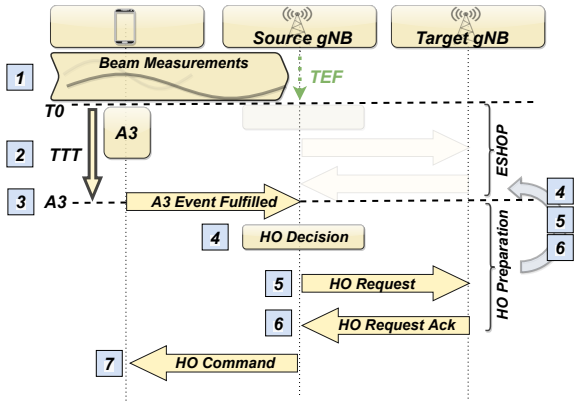


Fig. 8. Legacy inter gNodeB handover procedure and the proposed ESHOP scheme.

mmWave frequency bands [44]. For the ESHOP scheme in Fig. 7, we note that including the signaling step 6 within the scope of the ESHOP scheme would cause the TTT abortion according to [33]. The same applies to Step 7 in Fig. 8. Therefore, we confine the signalling optimizations within the proposed ESHOP scheme signalling scopes. Nevertheless, the BS retains the flexibility to cancel the ESHOP scheme at any time and revert to the legacy HO preparation procedure if necessary. In addition, when ESHOP fails to predict the fulfillment of the A3 event (Step 3), even though the BS can fall back to the legacy procedure, this

results in unnecessary HO preparation signaling which is the most significant disadvantage of the proposed HO scheme.

Finally, the findings of this study are limited to the simulated dataset generated based on the system described in Section II. In contrast, real network deployment would necessitate a larger dataset encompassing various mobility patterns adding more complexity when training the proposed ML model. However, the small-cell network deployments allow only for limited mobility patterns, confined to a specific coverage area, lending credibility to our model setup despite its simplicity.

V. MACHINE LEARNING FRAMEWORK

This section clarifies how ML algorithms can beneficially discover cell relationships between serving and neighboring cells to improve the robustness of the 5G HO preparation phase. Robust mobility management in advanced 5G deployments is challenging due to the unpredictable nature of user mobility patterns. To address this, machine learning (ML) algorithms have proven efficient in analyzing traffic and network data, and they are expected to be essential for improving 5G performance and robustness. ML-based technologies' ability to optimize parameters across multiple layers and identify patterns over complex time series has garnered significant attention in the wireless industry, as they hold the potential to revolutionize wireless network design and deployment. In realistic scenarios involving mobility, either in the propagation channel or due to UE movement, a massive number of signal observations are generated at each port of the gNodeB's MIMO antenna array. Consequently, the radio access network acquires, computes, and processes substantial amounts of data between layers 1 and 3. This implies that ML is ideally applied at higher system layers, utilizing signal observations from layer 1. For real-time HO prediction, it is preferable to select ML algorithms that can handle compute-intensive problems without compromising the baseband processing capacity. Thus, we chose to explore Temporal Convolutional Networks (TCN).

A. TEMPORAL CONVOLUTIONAL NETWORK

One significant application of neural networks is sequence modeling, specifically time series analysis, which involves capturing temporal structures in data for the purpose of making time-series predictions. Temporal Convolutional Networks (TCNs) excel in prediction tasks that involve time series data with complex patterns, making them an ideal choice over recurrent neural networks (RNNs). TCNs offer several advantages: they avoid the common drawbacks of RNNs, such as the exploding/vanishing gradient problem and inadequate memory retention. Additionally, TCNs enhance performance by enabling parallel computation of outputs, unlike RNNs. A key feature of TCNs is their causal nature, which ensures that an element in the output sequence relies only on preceding elements in the input sequence. This causality allows for direct conclusions between input and output,

something typically not achievable with most machine learning architectures. TCNs are implemented as residual blocks, which further boosts their learning capabilities and enables them to outperform other deep learning networks.
In the following, we explore the basic building blocks that a TCN consists of, and how they all interplay.
1) Sequence modeling: A sequence modeling network is any function given by

$$\boldsymbol{f} : \mathbf{X}^n \rightarrow \mathbf{Y}^n \tag{2}$$

that can be described as a function $\boldsymbol{f}$ that maps a given input sequence $\mathbf{x} \in \mathbb{R}^n = [x_0, x_1, ..., x_{n-1}]$ to a corresponding output sequence $\tilde{\mathbf{y}} \in \mathbb{R}^n = [y_0, y_1, ..., y_{n-1}]$ such that

$$\tilde{\mathbf{y}} = \mathbf{f}(\mathbf{x}) \tag{3}$$

In predicting task, each element y_t for a specific time index t ($0 \leq t \leq n$) can be calculated based on the input vector $\hat{\mathbf{x}} \in \mathbb{R}^t = [x_0, x_1, ..., x_{t-1}]$ that collects data that has been previously observed. In other words, the function $\boldsymbol{f}$ is *causal* which means that it does not depend on any future input x_{t+1}. In this paper, a neural network aims to solve a sequence modelling task so that the predicted output $\tilde{\mathbf{y}}$ approaches its ground truth $\mathbf{y} \in \mathbb{R}^n$. To measure the prediction quality, the following loss function L_{RMSE} is applied

$$L_{RMSE}(\mathbf{y}_t, \tilde{\mathbf{y}}_t)) = \sqrt{\sum_{i=1}^{t} \frac{(y_i - \tilde{y}_i)^2}{t}}, \tag{4}$$

where RMSE is defined as in (13).
2) Causal Convolutions: The TCN generates an output of the same length as the input and no data exposure from the future into the past time steps is allowed. The basic convolution relies on a causal input and filter, which makes it inappropriate for sequence modeling tasks, since the convolution operation depends on future time steps. The architecture of TCN is an extended model of a one-dimensional (1D) convolutional neural network (CNN) consisting of stacked convolutional layers and can be described as

$$F(x_t) = \sum_{p=0}^{k} f[p]\, x[t-p], \tag{5}$$

where k is the size of the filter with an input sequence of length n that returns a sequence of length $n - k + 1$ applied at time t. The zero padding of length $k - 1$ appended at the beginning of the sequence ensures the equal length of input and output. In other words, a casual convolution is used to prevent leakage from the past into future steps.

3) Dilated Convolutions: A dilated convolution is based on the causal convolution model with a slight modification via the so-called dilation factor d as defined below.

$$F'(x_t) = \sum_{p=0}^{k} f[p]\, x[t - (d \cdot p)]. \tag{6}$$

A dilated convolution, as shown in Fig. 9, allows a network to understand the dependencies of previous steps and exponentially increase the receptive field by expanding the dilation factor over all layers.

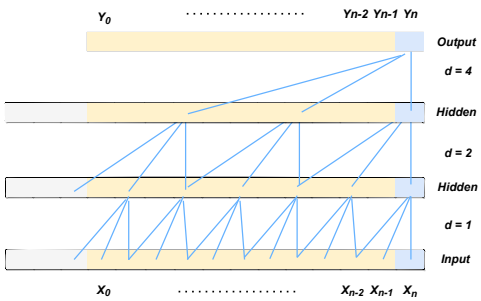


Fig. 9. A dilated causal convolution with dilation factors d = 1, 2, 4 and filter size k = 3.

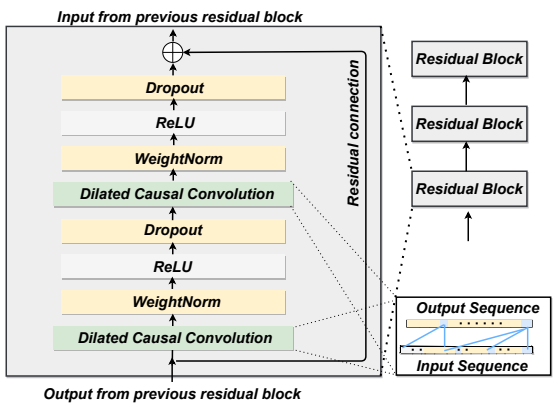


Fig. 10. Residual Convolutional Block.

4) Residual Connections: The mechanism of a residual block is illustrated by Fig. 10. In accordance with Fig. 10, the output of the network $\tilde{\mathbf{y}} \in \mathbb{R}^n$ can be expressed as

$$\tilde{\mathbf{y}} = f_{relu}(\boldsymbol{x} + \mathbf{F}(\boldsymbol{x})), \tag{7}$$

By applying the activation function f_{relu} to the each element x_i in $\mathbf{x}$, we get

$$f_{relu}(x_i) = max(0, x_i) \tag{8}$$

VI. RESULTS AND EVALUATION

We explore the potential of the ESHOP scheme using data generated from extensive simulations conducted within the system framework described in Section II. The simulated channel quality estimator for Layer 3 RSRP incorporates filtering based on the SSB beam-specific configuration, as part of the higher-layer radio resource management in 5G NR, following 3GPP standards [1] [33]. This filtering aims to eliminate the effects of fast fading and disregard short-term variations. The resulting measurement reports are sampled periodically at predefined time intervals of 40 ms resolution, providing averaged long-term RSRP measurements. Each beam measurement report comprises 12 pairwise samples of

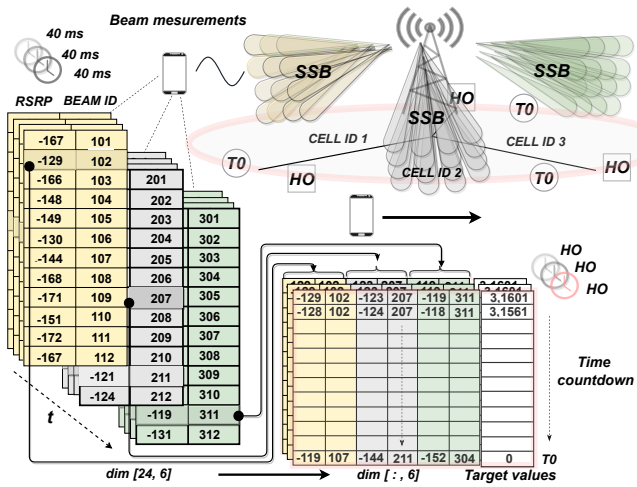


Fig. 11. Data set structure. L3-filtered RSRP beam measurements collected with a 40 ms reporting periodicity contain 12 RSRP values for each SSB beam. Between two HO occasions, these 40 ms measurement snapshots are reduced to the strongest beam for each cell, forming a time series of channel measurements including the corresponding time to the next HO occasion as the prediction target.

the best RSRP and beam ID per cell. Initially, the time series of measurement reports

yields a dimension of 3 cells x 12 beam ID/RSRP pairs, collected from the serving and the two neighboring cells. Subsequently, the dataset is reduced to an input vector containing only 6 features, such as the best RSRP with the corresponding SSB beam ID for each of the three cells, as illustrated in Fig. 11. All UE measurements are logged with the UE's simulated lifetime timestamps. Throughout each UE's simulated lifespan, spanning from the simulation's start to its end, we document the occurrence of **T0**, which is then converted into a countdown format. During ML model training, these timestamps serve as the prediction target, indicating the time remaining until the next **T0** event, presented as a countdown. Our evaluation is centered on enhancing the robustness of the HO preparation phase. Consequently, we restrict data collection to the moment of HO command transmission, disregarding the actual HO outcome. Additionally, if **T0** is not maintained for the entire duration of the TTT timer, leading to an A3 event, the associated measurements are excluded. Only measurements preceding the actual start of the HO procedure are considered. It's important to note that our approach, which relies on historical RSRP data, should not be seen as a limitation of our system or the proposed contribution. By incorporating historical RSRP data alongside the corresponding beam ID, our method ensures the effective linkage of the UE's trajectory and velocity with the measured attributes of the received signal. Given that the SSB beams remain static in terms of horizontal and vertical radio coverage, this approach inherently captures fingerprinted HCPs within the dataset, particularly those settings that have successfully satisfied the **T0** occasion.

A. TCN MODEL PERFORMANCE

The employed, relatively simple ML model, consists of a single TCN layer with a kernel size of 11 and incremental dilations with sizes $1, 2, 4, 8, 16, 32, 64$. This TCN layer is followed by three dense neural network layers of size 32, 16 and 8. There are only 590209 total parameters that need to be trained in this network. To evaluate the performance of the ML model the following metrics were used. Residual Mean Square Error R-Squared R^2 is calculated by comparing the Sum of Squares of Errors (SSE) to the Total Sum of Squares (SST) (9), Explained Variance Score (EVS) (10), Mean Absolute Percentage Error (MAPE) (11), Mean Absolute Error (MAE) (12) and Root Mean Squared Error (RMSE) (13).

$$R^2 = 1 - \frac{SSE}{SST} \tag{9}$$

$$\text{EVS} = 1 - \frac{Var(y - \hat{y})}{Var(y)} \tag{10}$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{\mid y_i - \hat{y} \mid}{y_i} \tag{11}$$

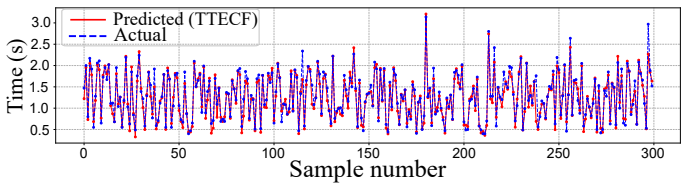
$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y} \mid \tag{12}$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y})^2}{n}}, \tag{13}$$

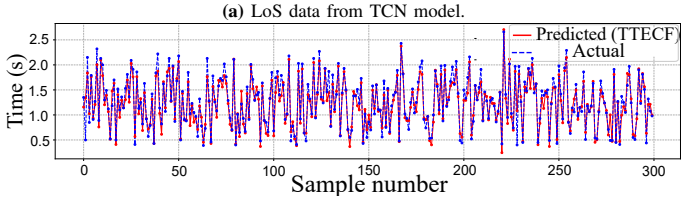
with y being predicted values and $\hat{y}$ the observed ones with results presented in Table II.

TABLE II. Averaged performance metrics for Fig. 12.

Metrics \ Scenario	LoS TCN	NLoS TCN
Explained Variance Score	0.934	0.897
Mean Absolute Percentage Error (%)	9.27	9.96
Mean Absolute Error	0.134	0.113
Root Mean Squared Error (s)	0.142	0.158
R-Squared (R^2)	0.929	0.886



(a) LoS data from TCN model.



(b) Non-LoS data from TCN model.

Fig. 12. Predicted vs. actual time to A3 event entry criterion fulfillment. 5000 epochs, UE velocity 25 m/s and Hysteresis = 1 dBm.

The results demonstrate that ML algorithms effectively capture the characteristics and corre-

lations among UE velocity, trajectory, and historical RSRP measurements while integrating the HO parameters with relatively low error rates, as depicted in Fig. 12 (a) and 12 (b) and summarized in Table II. Notably, while we opted to present results for a single velocity, the TCN framework yields comparable outcomes for both user velocities across accuracy and loss metrics.

The primary goal of this work was not to identify the fastest ML method but to demonstrate the feasibility of the ESHOP scheme, however, we acknowledge that in scenarios requiring extremely low latency, such as millisecond-level handover processes, the computational complexity of TCNs could pose a limitation. For highly time-sensitive applications, a thorough assessment of TCNs' real-time performance is essential. As deploying TCNs in real-time systems might be challenging, the use of fast CPUs, GPUs, or specialized hardware can help avoid significant delays. Additionally, TCNs might be more feasible for systems that make predictions based on shorter data sequences, such as those found in small-cell deployments with high UE velocity, where short but frequent handovers occur. Notably, although not presented in this study, we also explored other machine learning methods, such as Decision Trees [45], which can deliver comparable results to TCNs while offering much faster deployment.

When faced with challenges in obtaining sufficient and diverse data, the model's behavior might not be reliable in unique scenarios where data is lacking. However, by collecting data across various cell deployment classes and user movement patterns, the model's ability to generalize across different conditions would be enhanced.

B. ESHOP SCHEME PERFORMANCE

This section evaluates how the ESHOP approach can reduce the likelihood of UE encountering potential PDCCH outage areas at the far end of the HO region by initiating the early parts of the HO preparation stage. Our fingerprint-based approach indirectly incorporates HCP configuration, enabling user context-aware HO optimization. As a result, we refrain from comparing multiple HCP configuration settings or determining the optimality of one approach over another. Instead, our focus is on understanding the behavior and impact of the ESHOP scheme in various scenarios. Even without instantly optimized HCP parameters, we can demonstrate the viability of our innovative approach. Due to their relatively high velocity, users enter the HO region where radio conditions may deteriorate significantly, increasing the likelihood of transmission failure for measurement reports or HO commands. This situation is especially likely in denser cell deployments with reduced HO region size. In such mobility scenarios, it is crucial to anticipate impending HOs, as existing HO robustness mechanisms are prone to failure due to their reactive nature, partly caused by the inherent TTT delay. Figure 13 highlights the importance of user context-aware HO optimization, given that the network lacks control over UE-triggered events. The horizontal dashed lines indicate the average times between **A3** fulfillments for two different hysteresis settings, which are, as expected, highly dependent on UE velocity. As velocity increases, the time

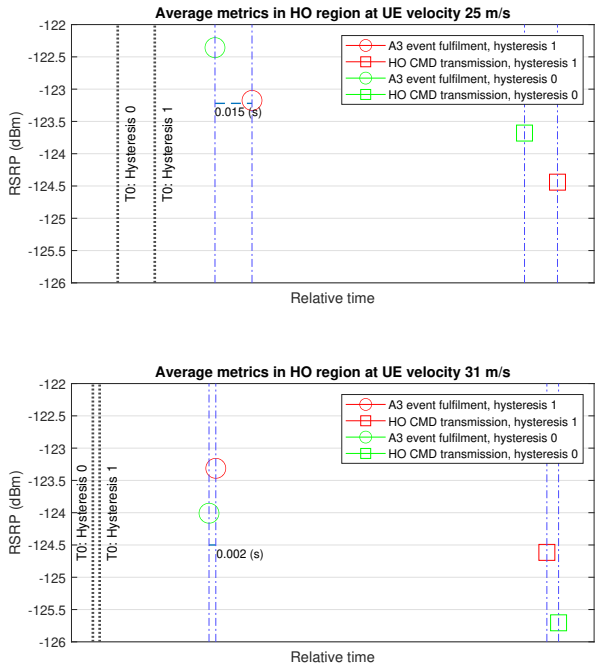


Fig. 13. Expanded handover region and the A3 event dynamics associated with different UE velocities. Due to ML-inferred A3 event *entry criterion fulfillment*, T0 is the new earliest HO trigger instead of the traditional A3 event fulfillment.

difference between the two events decreases, nearly merging into a single point in time. This observation provides crucial insights into the impact of UE velocity on the distance covered between consecutive UE measurement report intervals. High velocity renders static HCP settings, such as HO hysteresis, ineffective. This underscores the need to advance the start of the handover preparation phase and emphasizes the urgency for context-aware mobility. The value of the ESHOP scheme's ability to pinpoint the **T0** occasion, regardless of UE velocity, is evident.

The time spent in the HO region, by definition, does not include the TTT duration. In the legacy HO procedure, the TTT interval is often a period of inactivity, during which no signaling occurs until a potential HO event is triggered. However, during this time,

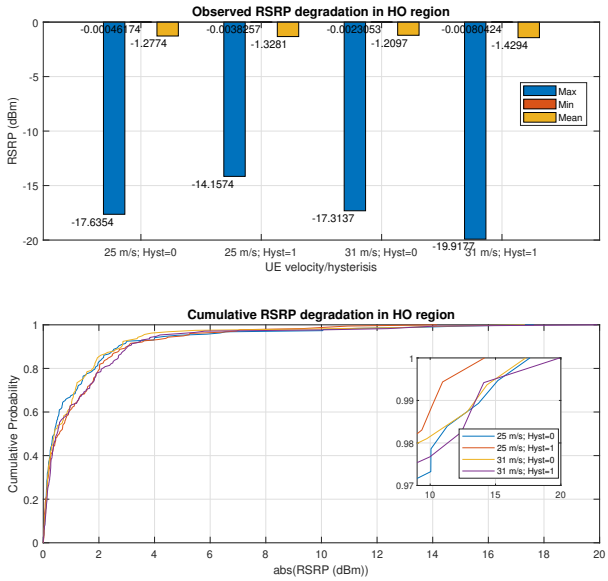


Fig. 14. Serving cell RSRP degradation mitigated by the ESHOP 40 ms after A3 fulfillment.

signal quality indicators such as the RSRP metric can significantly degrade, potentially leading to RLF if the UE cannot maintain a stable connection with the serving cell long enough to complete the handover to the target cell. As illustrated in Figures 7 and 8, the ESHOP scheme effectively anticipates the HO preparation by exactly the TTT duration and integrates HO decision-related signaling into the TTT interval. This approach is the cornerstone of the ESHOP scheme as it parallelizes the most time-consuming part of the HO preparation process with the TTT execution. Consequently, the HO command can be sent immediately upon **A3** event fulfillment, whereas in conventional HO procedures, this would be delayed by HO decision-related signaling. By reducing the overall HO signaling time, ESHOP helps maintain higher RSRP levels, which is particularly crucial in high-speed UE mobility scenarios. Figure 14 shows the observed channel degradation from **A3** fulfillment until HO decision signaling is executed, approximately 35-40 ms. Even though the observed RSRP metrics are closely tied to the mobility scenarios modeled in Section II, they underscore that rapid UE movement can result in sudden RSRP degradation, causing the signal quality to deteriorate too quickly for the handover process to be completed successfully. By doing so, it helps maintain stable signal quality, preventing the rapid

deterioration that could otherwise compromise the handover process. In particular, the extreme RSRP degradation cases depicted in the upper part of Fig. 14, with a cumulative probability close to the 98th percentile, would likely result in RLF and HOF. Maintaining an adequate RSRP level is a primary and direct benefit of ESHOP's ability to parallelize HO-related signaling procedures.

We highlight that the mobility scenarios modeled in the system described in Section II concentrated on intra-gNodeB mobility and did not include HO decision-related signaling. To assess the performance of the ESHOP scheme, we used measurements from a commercial-grade 5G test network as a reference. The measured signaling times averaged between between 15-35 ms for legacy procedures as shown in steps 2-4, Fig. 7 and steps 2-5, Fig. 8 which fits entirely within the shortest specified TTT duration of 40 ms.

VII. CONCLUSIONS

This study demonstrates how the ESHOP scheme mitigates serving cell RSRP degradation within the HO region by initiating the HO preparation phase earlier. We focus on simple deployments and mobility scenarios to highlight the novel approach presented, specifically targeting small-cell deployments due to their frequent HOs, which are not as prevalent in large-cell networks. The dense, small-cell model results in rapid and frequent HOs, providing a robust test of the proposed algorithm. In contrast, larger network deployments would complicate the analysis due to their complexity and the challenges in generating, collecting, and processing measurement data, exceeding the scope of this study.

The proposed ESHOP scheme reduces or removes the forced inactivity by parallelizing the HO preparation and the TTT intervals, and reduces instances of severe link degradation and HO failures. Hence, a user experiences more stable connections and increases the 5G NR network capacity. Allthough our study focuses on A3 events, the deployed ML model can be trained on any predefined HO event due to its flexible implementation. The proposed ESHOP scheme is designed for the network side, however, its predictive capabilities for upcoming handover events can beneficially be integrated on the UE side, enhancing techniques such as Conditional HO.

Future studies should combine narrow-beam Channel State Information Reference Signals (L1-RSRP) with wide-beam SSB (L3-RSRP) measurements at the cell edge to enhance data resolution by capturing both levels of beam measurement information. We encourage future handover optimization studies to implement multiple instances of the ESHOP scheme across various network nodes, allowing them to share insights related to load balancing and handover coordination. Protocols like Xn facilitate communication between base stations, enabling these nodes to optimize both user experience and network efficiency. This approach would promote a more comprehensive and effective handover optimization strategy across the network.

Additionally, we believe that attention-enabled generative AI models could significantly

enhance the ESHOP framework. Attention models, particularly Transformer-based architectures, generally outperform TCNs for tasks that involve capturing long-range dependencies, which are likely to occur in more complex network environments. Their ability to model long-term dependencies and correlations makes Transformer-based models ideal for time series applications, such as forecasting. Since the proposed ESHOP model relies on the dataset's capacity to capture key fingerprinted features, such as a user's trajectory and velocity, incorporating a Transformer-based model for trajectory analysis could potentially improve the accuracy of the ESHOP framework. Although deploying Transformer-based models in commercial systems may pose challenges, GPUs or specialized hardware can help mitigate delays, particularly in complex commercial network environments.

REFERENCES

[1] 3GPP TR 38.300 Rel. 16, "5G NR – Overall description Stage-2, 9.2.4 Measurements," *Third Generation Partnership Project (3GPP)*, Dec. 2021.

[2] E. Dahlman, S. Parkvall, and J. Sköld, "5G NR - The Next Generation Wireless Access Technology," *Academic Press Inc*, 2018, pp. 16-19, 144-145.

[3] 3GPP TSG-RAN WG2 nr77, "Discussion on large-area HetNet simulations," *Third Generation Partnership Project (3GPP)*, Feb. 2012.

[4] 3GPP TR 36.133, "5G NR – Requirements for support of radio resource management," *Third Generation Partnership Project (3GPP)*, Dec. 2020.

[5] I. Shayea, M. Ergen, A. Azizan, M. Ismail, and Y. I. Daradkeh, "Individualistic Dynamic Handover Parameter Self-Optimization Algorithm for 5G Networks Based on Automatic Weight Function," *IEEE Access*, vol. 8, pp. 214392 - 214412, Nov. 2020.

[6] H. Chen, D. Lopez-Perez, Y. Yang, and F. Gunarsson, "Mobility and Handover Management, Heterogeneous Cellular Networks: Theory, Simulation and Deployment," *Cambridge Univ. Press*, pp. 245–83, 2013.

[7] A. Prado; H. Vijayaraghavan, and W. Kellerer, "Enhanced Conditional Handover boosted by Trajectory Prediction," *GLOBECOM - IEEE Global Communications Conference*, Dec. 2021.

[8] L. Changsung, C. Hyoungjun, S. Sooeun, and C. Jong-Moon, "Prediction-Based Conditional Handover for 5G mm-Wave Networks: A Deep-Learning Approach," *IEEE Vehicular Technology Magazine*, vol. 15, no. 1, pp. 54-62, Jan. 2020.

[9] S. B. Iqbal, A. Awada, U. Karabulut, I. Viering, P. Schulz, G. P. Fettweis, "On the Modeling and Analysis of Fast Conditional Handover for 5G-Advanced," *IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2022.

[10] A. Klein, C. Mannweiler, J. Schneider, F. Thillen, and D. S. Hans, "A concept for context-enhanced heterogeneous access management," *Proc. IEEE Globecom Workshops*, pp. 6-10, Dec. 2010.

[11] M. Manalastas, M. U. B. Farooq, S. M. A. Zaidi, A. Abu-Dayya, A. Imran, "A Data-Driven Framework for Inter-Frequency Handover Failure Prediction and Mitigation," *IEEE Transactions on Vehicular Technology*, vol. 71, Issue: 6, pp. 6158-6172, March 2022.

[12] M. U. B. Farooq, M. Manalastas, W. Raza, S. M. A. Zaidi, A. Rizwan, A. Abu-Dayya, "A Data-Driven Self-Optimization Solution for Inter-Frequency Mobility Parameters in Emerging Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, Issue: 2, pp. 570-583, Feb. 2022.

[13] M. T. Nguyen, S. Kwon, and H. Kim, "Mobility robustness optimization for handover failure reduction in LTE small-cell networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4672-4676, May 2018.

[14] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. Fettweis, "Evaluation of context-aware mobility robustness optimization and multi-connectivity in intra-frequency 5G ultra dense networks," *2016 IEEE Wireless Communications Letters*, vol. 5, no. 6, pp. 608-611, 2016., Dec. 2016.

[15] M. Nguyen, and S. Kwon, "Geometry-Based Analysis of Optimal Handover Parameters for Self-Organizing Networks," *IEEE Transactions on Wireless Communications* vol. 19, pp. 2670 - 2683, Jan. 2020.

[16] K. Vasudeva, M. Simsek, D. Lopez-Perez, and I. Guvenc, "Analysis of handover failures in heterogeneous networks with fading," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6060-6074, Jul. 2017.

[17] X. Xu, Z. Sun, X. Dai, T. Svensson, and X. Tao, "Modeling and analyzing the cross-tier handover in heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7859-7869, Dec. 2017.

[18] R. Karmakar, G. Kaddoum, and S. Chattopadhyay, "Mobility Management in 5G and Beyond: A Novel Smart Handover with Adaptive Time-to-Trigger and

Hysteresis Margin," *IEEE Transactions on Mobile Computing*, vol. 22, no. 10, pp. 5995-6010, Oct. 2023.

[19] G. Feng, S. Qin, Y. Liang, and T. Peter Yum, "The SMART Handoff Policy for Millimeter Wave Heterogeneous Cellular Networks," *IEEE Transactions On Mobile Computing*, vol. 17, no. 6, pp. 1456-1468, June. 2018.

[20] W. Huang, M. Wu, Z. Yang, and K. Sun, "Self-Adapting Handover Parameters Optimization for SDN-Enabled UDN," *IEEE Transactions on Wireless Communications*, Feb. 2022.

[21] V. Mishra, D. Das, and N. N. Singh, "Novel Algorithm to Reduce Handover Failure Rate in 5G Networks," *IEEE 3rd 5G World Forum (5GWF)*, Sept. 2020.

[22] H-S. Park, Y-S. Choi, B-C. Kim, and Jae-Yong Lee, "LTE Mobility Enhancements for Evolution into 5G," *ETRI Journal*, vol. 37, no. 6, Dec. 2015.

[23] A. Masri, T. Veijalainen, H. Martikainen, S. Mwanje, J. Ali-Tolppa, and M. Kajó, "Machine-Learning-Based Predictive Handover," *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, May 2021.

[24] F. Guidolin, I. Pappalardo, A Zanella, and Michele Zorzi, "Context-Aware Handover Policies in HetNets," *IEEE Transactions on Wireless Communications*, vol. 15, issue 3, pp 1895-1906, March 2016.

[25] A. Ö. Kaya, and H. Viswanathan, "Deep Learning-based Predictive Beam Management for 5G mmWave Systems," *IEEE Wireless Communications and Networking Conference (WCNC)*, 2021.

[26] M. U. B. Farooq, M. Manalastas, W. Raza, A. Ijaz, S. M. A. Zaidi, A. Abu-Dayya, and A. Imran, "Data Driven Optimization of Inter-Frequency Mobility Parameters for Emerging Multi-band Networks," *GLOBECOM - IEEE Global Communications Conference.*, Dec. 2020.

[27] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27-33, Nov. 2014.

[28] I. L. Da Silva, C. Eklöf, J. Muller, and R. Zhohov, "This is the key to mobility robustness in 5G networks," *Ericsson technology blog*, https://www.ericsson.com/en/blog/2020/5/the-key-to-mobility-robustness-5g-networks, May. 2020.

[29] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173-196, Sept. 2018.

[30] F. D. Calabrese, P. Frank, E. Ghadimi, U. Challita, and P. Soldati, "Enhancing RAN Performance with AI," *Ericsson technology review article*, https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/enhancing-ran-performance-with-ai, Jan. 2020.

[31] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, S. Alraih and K. S. Mohamed, "Auto Tuning Self-Optimization Algorithm for Mobility Management in LTE-A and 5G HetNets," *IEEE Access*, Dec. 2019.

[32] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, and S. Alraih, "Dynamic Handover Control Parameters for LTE-A/5G Mobile Communications," *IEEE Advances in Wireless and Optical Communications (RTUWO)*, Nov. 2018.

[33] 3GPP TR 38.331, "5G NR – Radio Resource Control; Protocol specification," *Third Generation Partnership Project (3GPP)*, Sep. 2021.

[34] 3GPP TR 38.104, "Base Station (BS) radio transmission and reception," *Third Generation Partnership Project (3GPP)*, Dec. 2020.

[35] 3GPP TR 38.901, "Study on channel model for frequencies from 0.5 to 100 GHz," *Third Generation Partnership Project (3GPP)*, Mars 2022.

[36] IST-4-027756 WINNER II D1.1.2 V1.2, "WINNER II Channel Models," *Third Generation Partnership Project (3GPP)*, Dec. 2007.

[37] H.-S. Park, Y. Lee, T.-J. Kim, B.-C. Kim, and J.-Y. Lee, "Handover mechanism in NR for ultra-reliable low-latency communications," *IEEE Network*, vol. 32, no. 2, pp. 41–47, Apr. 2018.

[38] M. E. Morocho Cayamcela, and W. Lim, "Artificial intelligence in 5g technology: A survey," *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 860-865, Oct. 2018.

[39] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine learning for 5G/beyond 5G mobile and wireless communications: Potential limitations and future directions," *IEEE Access*, vol. 7, pp. 137184-137206, Sept. 2019.

[40] 3GPP TR 38.401, "NG-RAN; Architecture description (Release 16)," *Third Generation Partnership Project (3GPP)*, Sept. 2022.

[41] 3GPP TR 36.331, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 16)," *Third Generation Partnership Project (3GPP)*, Sept. 2020.

[42] 3GPP TR 38.213, "NR; Physical layer procedures for control," *Third Generation Partnership Project (3GPP)*, April 2019.

[43] Y. Chen, Y. Kang, Y. Chen, and Z. Wang, "Probabilistic forecasting with temporal convolutional neural network," *2020 Neurocomputing*, vol. 399, pp. 491-501, July 2020.

[44] 3GPP TR 36.133, "Evaluation and design aspects for NR mobility enhancement," *Third Generation Partnership Project (3GPP)*, 3GPP R2-168852, Nov. 2016.

[45] L. Breiman, J. Friedman, C. J. Stone, and R.A. Olshen, "Classification And Regression Trees," *Taylor & Francis. Press*, 1984.

Paper IV

Paper IV

Reproduced, with permission from IEEE

D. Pjanić, K. E. Arslantürk, X. Cai, F. Tufvesson, "Dynamic User grouping based on Location and Heading in 5G NR System," *IEEE 100th Vehicular Technology Conference*, Oct. 2024, Washington DC, USA, doi: 10.1109/VTC2024-Fall63153.2024.10757679.

Dynamic User Grouping based on Location and Heading in 5G NR Systems

Dino Pjanić[1,2], Korkut Emre Arslantürk[2], Xuesong Cai[2], and Fredrik Tufvesson[2],
[1] Ericsson AB, Sweden
[2] Dept. of Electrical and Information Technology, Lund University, Sweden

Abstract

User grouping based on geographic location in fifth generation (5G) New Radio (NR) systems has several applications that can significantly improve network performance, user experience, and service delivery. We demonstrate how Sounding Reference Signals channel fingerprints can be used for dynamic user grouping in a 5G NR commercial deployment based on outdoor positions and heading direction employing machine learning methods such as neural networks combined with clustering methods.

Index Terms

5G, beamforming, localization, machine learning, positioning, radio access network, Sounding Reference Signal, user grouping,

I. Introduction

The need for User Equipment (UE) positioning in cellular networks dates back to their early generations, initially driven by requirements for emergency call localization. Precise geographical localization capabilities have been a subject of research for decades. Although most existing localization solutions are enabled by Global Navigation Satellite Systems (GNSS), there is an increasing need for standalone positioning capabilities within fifth-generation (5G) cellular systems. GNSS technology can be unreliable in dense urban environments due to shadowing, multipath propagation, and poor satellite coverage [1], [2]. Driven by various use cases such as smart factories, autonomous vehicles, and sensing, cellular UE positioning has emerged as a key service provided by 5G networks. Recent research has advanced 5G outdoor positioning to very accurate solutions, broadly classified into two categories: conventional signal processing [3]–[5], and Machine Learning (ML) based methods [6]–[8]. Signal processing methods, which use Time of Arrival (ToA), Angle of Arrival (AoA), and Time Difference of Arrival (TDoA), require the estimation of radio channel parameters between UE and base stations (BS). In contrast, ML-based methods rely on pre-processed data for training.

However, many essential 5G functionalities, such as mobility management, network planning, and data analytics [9] cannot depend on GNSS services and must rely on internal positioning

This work is partially sponsored by the Swedish Foundation for Strategic Research and Ericsson AB.

capabilities. As positioning techniques in 5G new radio (NR) systems evolve, position-based user grouping becomes the next logical step, enhancing the ability to capture spatial relationships alongside grouping based on channel conditions. This approach could significantly benefit key 5G functions, including:

- *Network Resource Optimization*: Load balancing can use geographical UE grouping to comprehend UE distribution across different cells, avoiding congestion. Spectrum efficiency improves by grouping UEs based on cell location, allowing more efficient frequency reuse and interference management.
- *Enhanced Mobility Management*: Handover optimization benefits from understanding the movement patterns and behavior of the UE. The network can anticipate handovers and prepare target cells in advance, reducing handover failures and maintaining service continuity. Context-aware services can provide UEs with relevant information and services based on their current geographical position.
- *Quality of Service Improvement*: Using beamforming in targeted areas, beamforming techniques can be optimized to enhance signal quality and data rates.
- *Network Planning*: Infrastructure deployment can leverage geographical data to guide the placement of new BSs, small cells, and other network infrastructure. Capacity planning can anticipate areas and times of high UE densities, helping to manage demand.

There are significant research gaps in location-based UE clustering within the 5G NR system. The authors of [10] explored the classification of UEs in dynamic millimeter-wave scenarios using conventional ML techniques on simulated CSI-RS measurements without directly considering physical positioning. The study in [11] proposes dynamic UE-group-based interference management by adjusting data transmission powers in small cell deployments. As a reference, our study utilizes a high-resolution uplink (UL) Sounding Reference Signal (SRS) dataset, recently showcased in a highly accurate positioning model [12]. This model outperforms previous studies by regressing UE positions using an attention-based approach. The major contributions of this paper are as follows:

- We propose an accurate outdoor positioning model that utilizes SRS channel estimates to infer the actual user position and the Course Over Ground (COG) or heading direction.
- To the best of our knowledge, this is the first study to introduce geographical-based user grouping through clustering methods in a commercial 5G system.

II. System Model and Data Collection

The fundamental concept behind positioning with UL SRS channel estimates is that a specific physical location under similar radio channel conditions corresponds to unique SRS-generated channel estimates. Each UE transmit antenna acts as a resource where channel measurements are gathered by the numerous antennas of the massive Multiple-Input Multiple-Output (MIMO) receiver in the UL. At the BS end, a MIMO system enables improved channel measurements across multiple frequency resource instances of the entire bandwidth, incorporating the UE location information via AoA, ToA and TDoA. We consider a commercial 5G NR Time Division Duplex (TDD) system in a single-user massive MIMO scenario, where the BS processes a time series of SRS measurements that capture the angular delay spectrum of the radio channel in the beam domain. At time t, the UE, equipped with M_{UE} antenna elements, transmits an UL pilot signal.

This pilot signal reaches the BS at an azimuth arrival angle ϕ and an elevation angle θ. The BS is equipped with M_{BS} antennas, half of which is vertically polarized and the other half horizontally polarized. We suppose that the number of multipath components is P, and denote $\tau_{p,t}$ as the time delay between UE and BS w.r.t. the p-th path at time t, and $\zeta_{p,m,t}$ denotes the complex amplitude of each multipath component. The BS utilizes all vertical-polarized antennas to form N_{V} beams, the response of the i-th beam w.r.t. the p-th path is $\psi_{\text{V},i}(\phi_p, \theta_p, f)$, where f denotes frequency, and ϕ_p and θ_p represent the azimuth and elevation arrival angles for the p-th multipath, respectively. Another set of N_{H} beams uses all horizontal polarized antennas, and the response of the i-th beam is $\psi_{\text{H},i}(\phi_p, \theta_p, f)$. For the m-th UE antenna, the propagation channel is modeled for each beam at time index t as

$$\begin{aligned} h_{\text{V},i,m,t}(f) &= \sum_{p=1}^{P} \psi_{\text{V},i}(\phi_p, \theta_p, f)\zeta_{p,m,t} \exp\{-j2\pi f \tau_{p,t}\} \\ h_{\text{H},i,m,t}(f) &= \sum_{p=1}^{P} \psi_{\text{H},i}(\phi_p, \theta_p, f)\zeta_{p,m,t} \exp\{-j2\pi f \tau_{p,t}\}. \end{aligned} \tag{1}$$

By collecting all $h_{\text{V},i,m,t}(f)$ and $h_{\text{H},i,m,t}(f)$ for the F subcarriers, we formulate two beam space matrices of the Channel Transfer Functions (CTFs), $\mathbf{H}_{\text{V},m,t} \in \mathbb{C}^{N_V \times F}$ and $\mathbf{H}_{\text{H},m,t} \in \mathbb{C}^{N_{\text{H}} \times F}$ at time t, which correspond to the vertical and horizontal polarized antenna groups, respectively. We further define the matrix $\mathbf{H}_t \in \mathbb{C}^{N \times F} = \left[\mathbf{H}_{\text{H},1,t}^T, \mathbf{H}_{\text{V},1,t}^T, ..., \mathbf{H}_{\text{H},M_{\text{UE}},t}^T, \mathbf{H}_{\text{V},M_{\text{UE}},t}^T\right]^T$ that combines the channel matrices of all UE antennas as $N = M_{\text{UE}}(N_{\text{H}} + N_{\text{V}})$ that depends on the UE position and therefore meets the criteria needed to perform ML-based localization.

A. Outdoor 5G NR measurement campaign

To assess our localization pipeline, we conducted an outdoor measurement campaign in a parking lot near the Ericsson office in Lund, Sweden. Fig. 1 shows photos of the BS antenna and measurement locations. Throughout the campaign, a commercial UE was placed on top of a test vehicle alongside a high-performance GNSS receiver, providing ground truth reference with centimeter-level positioning accuracy and COG parameter featuring GNSS multi-band and multi-constellation support. To ensure uninterrupted SRS transmission, the UE remained in a connected state while simultaneously downloading data at a rate of 750 Mbit/s. The UL SRS pilot signals were received and processed by a commercial Ericsson 5G BS in TDD mode, operating in the mid-band at a center frequency of 3.85 GHz, compliant with the 5G NR 3GPP standard 38.104 Rel15 [13]. The BS was equipped with an integrated radio with 64 transmitters/receivers (TX/RX) and 32 dual polarized antennas. For digital beam forming, the 64 TX/RX formulate 64 beams in DL/UL respectively. As illustrated in Fig. 1, our measurement campaign includes two distinct scenarios: Line-of-Sight (LoS) and non Line-of-Sight (NLoS). In both scenarios, the velocity of the test vehicle was approximately 5 m/s. The trajectory for each of the two measurement scenarios consists of 4 laps with 4 different UE mobility patterns: clockwise, clockwise random, anticlockwise, and anticlockwise random. This approach creates four distinct movement trajectories for each scenario, which makes them suitable for clustering. As UEs move, the clockwise and anticlockwise patterns cause them to dynamically move towards or away from each other, while the random trajectories introduce additional variation to the geographical distribution of the UEs. To increase the total number of UEs in each scenario, the data generated

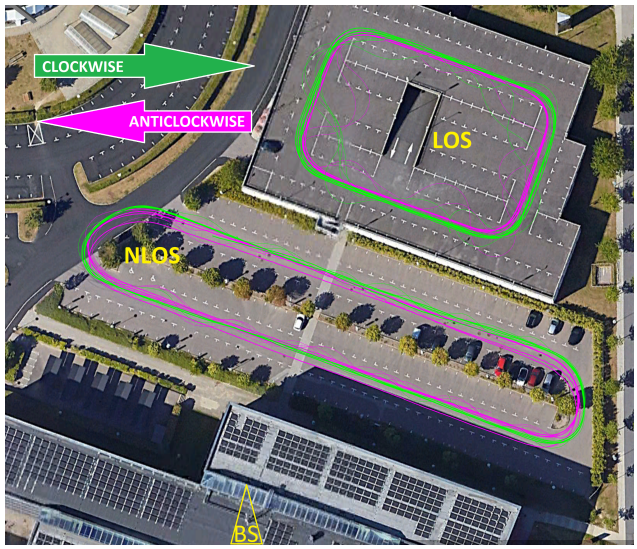


Fig. 1: The 5G NR base station was installed on top of a 20-meter-high building. During this measurement campaign, a test vehicle traversed two predefined routes: A 10-meter-high garage path for LoS measurements and a ground-level path for NLoS measurements below the base station building. Each route features four different movement patterns. Thinner lines depict random trajectories.

during the 4 laps were divided in half, resulting in an additional 4 *virtual* users, totaling 8 users with 2 laps in each scenario.

B. Signal Processing Pipeline

As illustrated in Fig. 2, the SRS channel estimates cover 273 Physical Resource Blocks (PRBs) over a 100 MHz bandwidth. Each channel snapshot comprises 273 PRBs for all 64 beams based on SRS reporting periodicity of 20 ms. The PRBs are grouped in adjacent pairs and averaged by downsampling, taking the mean value of the sampled data points resulting in 137 PRB Subgroups (PRSGs). Downsampling was performed on every third PRSG, resulting in a total of 46 PRSGs. The UE, equipped with 4 antennas (i.e., 4 UE layers), transmits the SRS pilots. We recorded the SRS pilots from 2 UE layers at a time, forming two channel transfer function matrices $\mathbf{H}_1, \mathbf{H}_2 \in \mathbb{C}^{N \times F}$. We define a matrix $\mathbf{H}' \in \mathbb{C}^{2N \times F}$ to collect those two matrices, specifically, $\mathbf{H}' = [\mathbf{H}_1, \mathbf{H}_2]$ (with $N = 64, F = 46$). After initial processing, the final CTF snapshot for 64 gNodeB antennas and two UE layers has a dimension of 1x128 amplitude instances, as depicted in Fig. 2, collected and averaged over 46 PRSGs for each gNodeB antenna. Gathering UL SRS channel measurements in a commercial 5G NR BS faces constraints when retrieving data-rich structures such as SRS channel measurement samples. The extensive SRS data, generated at millisecond intervals, typically reside within the BS's baseband entity, primarily for internal processing. However, accessing these data externally may be impeded by hardware and software constraints. As not all PRSG values are

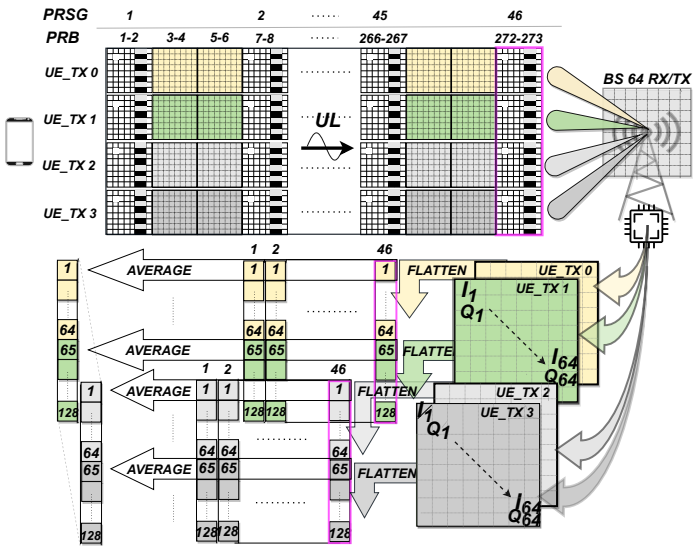


Fig. 2: SRS data stream collection and pre-processed CTF dataset.

updated during SRS transmission, it is necessary to represent the missing channel estimate values. To ensure the validity of the CTF for missing PRSGs, we employ the simplest method, such as forward-filling, using the latest known values.

III. Proposed ML-based Clustering Framework

ML-driven UE grouping framework is illustrated in Fig. 3 consisting of two sequential blocks:

1) **Positioning block:** This block is designed to achieve precise positioning and incorporates a Convolutional Neural Network (CNN) [14] in conjunction with a Feedforward Neural Network (FNN) [15]. Both networks utilize features extracted from the SRS dataset as input to regress the local P_{XY} position, along with COG_{DEG}.
2) **Clustering block:** Leveraging the highly accurate positioning block, we utilize nonparametric clustering methods that do not require a pre-setting of the number of clusters, namely Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [16] and Hierarchical clustering [17] for the UE grouping.

A. DBSCAN

DBSCAN uses tree techniques called *dendograms* [18], a tree-structured graph, and groups points into clusters based on their density, identifying areas with a high data point density separated by regions of low density. Since dendograms use features only indirectly as the basis for distance calculation, they partition the given data rather than entire instance space, and hence represent descriptive clustering rather than predictive one. This makes DBSCAN especially qualified for handling clusters of arbitrary shapes and sizes, even in noisy data. DBSCAN requires two

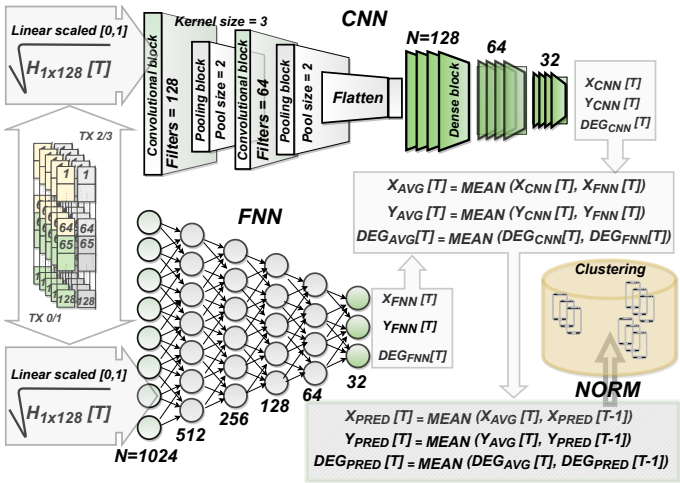


Fig. 3: The architecture of the ML-driven grouping framework along with the input pipeline and intended output. The output of the positioning model is averaged with the previous prediction as a post-processing step and normalized before forwarding it to the clustering algorithms for user grouping.

parameters: epsilon (*eps*) and the minimum number of samples *MinPts*. The *eps* represents the radius around a data point, and *MinPts* the minimum number of data points within *eps* to form a dense region.

B. Hierarchical clustering

Hierarchical clustering has the distinct advantage that any valid measure of distance can be used. In hierarchical clustering, deciding which clusters to combine or where to split requires a measure of dissimilarity between sets of observations. Most methods achieve this by using an appropriate distance *threshold*, such as the Euclidean distance, between individual observations in the dataset. As there is a need to measure how close two clusters are, a *linkage criterion* is employed, which is a general way to turn pairwise point distances into pairwise cluster distances. Our model used the WARD *linkage criterion* defined as:

$$\Delta(A,B) = \frac{\mu_A \mu_B}{\mu_A + \mu_B} \|\vec{m_A} - \vec{m_B}\|^2 \tag{2}$$

where A and B are two sets of observations with a centre of cluster *i* denoted as $\vec{m_i}$ and the number of points in it as μ_i. Tables I and II summarize the hyperparameter settings of the ML model illustrated in Fig. 3.

IV. Results And Discussion

The accuracy of the positioning block is summarized in Table III and Fig. 4 demonstrating a sub-1 m accuracy level of precision and sub-9 ° heading direction.

TABLE I: Hyperparameters employed by the positioning block

Layers	Activation	Batch s.	Epochs	Optimizer	Loss f.
		FNN			
8	ReLU	64	200	ADAM	MSE
		CNN			
2	ReLU	64	200	ADAM	MSE

TABLE II: Parameters employed by the clustering block

Hyperparameter	DBSCAN
Distance Metric	Euclidian
eps	0.5, 0.6
minPTS	1
Algorithm	Auto
Hyperparameter	**HIERARCHICAL**
Distance Metric	Euclidian
Distance Threshold	0.5, 1.0
Compute Full Tree	Auto
Linkage criterion	Ward

Based on the highly accurate performance of the proposed positioning block, the subsequent clustering block aims to partition all points in the dataset into groups of similar objects, where the notion of similarity is highly domain-dependent. As illustrated in Fig. 3, the two positioning features (X and Y coordinates) and the heading feature (°) were input into the clustering block. To address the dynamic nature of cellular networks, which incorporate UE mobility, non-parametric clustering algorithms using various hyperparameter values are deemed more suitable than parametric ones. Assessing the quality of user clustering in a 5G system requires understanding how well the clustering meets the specific criteria of various network functionalities. The goal of this study is not to define the best clustering hyperparameters, as these are highly dependent on specific network function domains. Therefore, we refrain from optimizing the clustering parameter settings or determining the optimality of one approach over another. Instead, we focus on demonstrating the potential of the proposed user grouping framework. We envision it as an internal capability of the 5G system, primarily for use cases requiring real-time user localization, such as beamforming, handover, or cell interference management when the UE remains connected. This also implies that users will be dynamically added to and removed from the clusters as they enter or exit the cell.

To underscore the heading feature's significance, we conduct user grouping based on estimated position alone and, in the subsequent step, include the heading feature. Figures 5, 6, 7, and 8 present some time aligned comparison results of UE grouping based on various parameter settings, where each UE is represented by an arrow indicating its regressed direction and location. Colors are used to distinguish between different UE clusters. Our approach, besides the novel

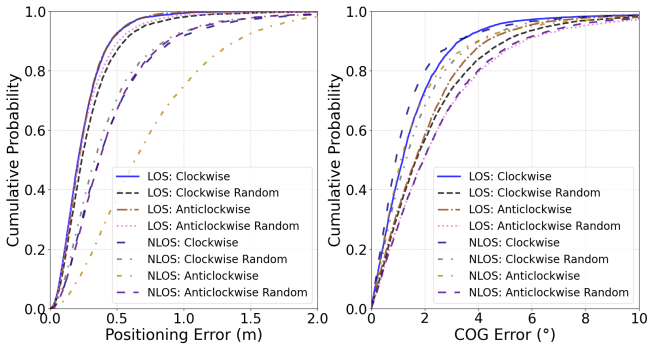


Fig. 4: Positioning and COG errors.

TABLE III: Performance of the positioning block

Scenario	Metric	X (m)	Y (m)	Heading (°)	Dist. (m)
		CLOCKWISE			
LoS	RMSE	0.2	0.23	7.1	0.3
	R2 Score	0.99988	0.99976	0.995	NA
NLoS	RMSE	0.53	0.35	6.33	0.64
	R2 Score	0.9998	0.9998	0.993	NA
		CLOCKWISE RANDOM			
LoS	RMSE	0.29	0.35	8.97	0.454
	R2 Score	0.9997	0.9995	0.99	NA
NLoS	RMSE	0.52	0.39	5.8	0.65
	R2 Score	0.9997	0.9994	0.996	NA
		ANTICLOCKWISE			
LoS	RMSE	0.23	0.21	8.36	0.31
	R2 Score	0.9998	0.9998	0.993	NA
NLoS	RMSE	0.53	0.75	8.76	0.92
	R2 Score	0.9997	0.998	0.991	NA
		ANTICLOCKWISE RANDOM			
LoS	RMSE	0.26	0.26	8.56	0.36
	R2 Score	0.9998	0.9997	0.993	NA
NLoS	RMSE	0.48	0.44	8.37	0.65
	R2 Score	0.9998	0.9991	0.991	NA

location-based user grouping itself, successfully incorporates the heading direction feature, which becomes particularly interesting for cases such as user movement predictions, dynamic adjustment of beamforming patterns, etc. For more detailed positioning and clustering results, we refer to the work conducted in [19] which used the same datasets and ML pipeline introduced in this paper.

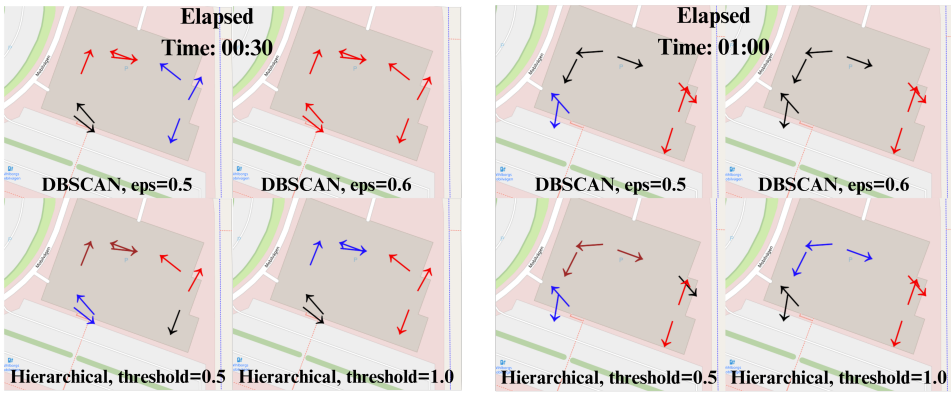


Fig. 5: LoS clustering results based on position

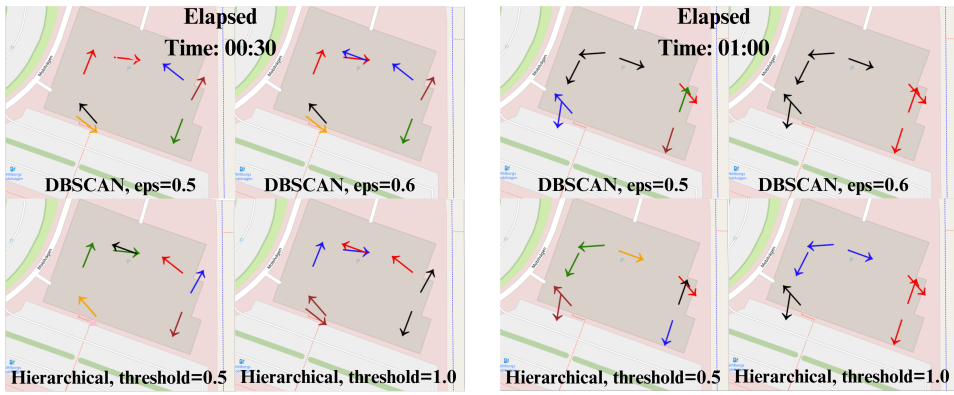


Fig. 6: LoS clustering results based on position and heading

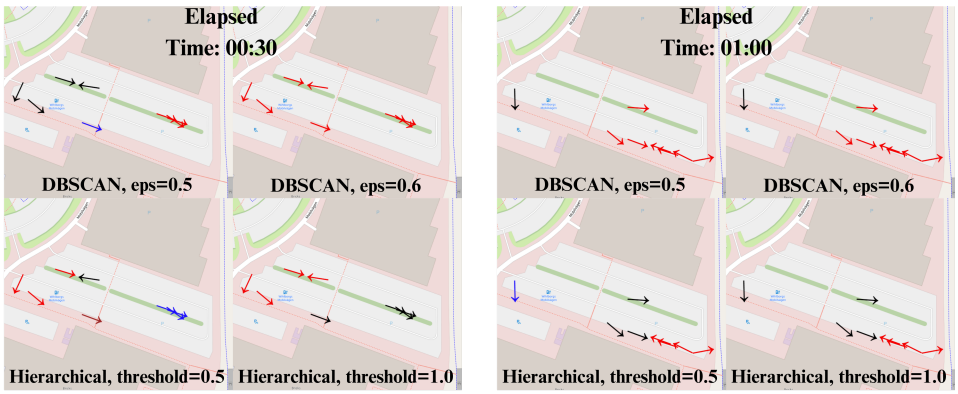


Fig. 7: NLoS clustering results based on position

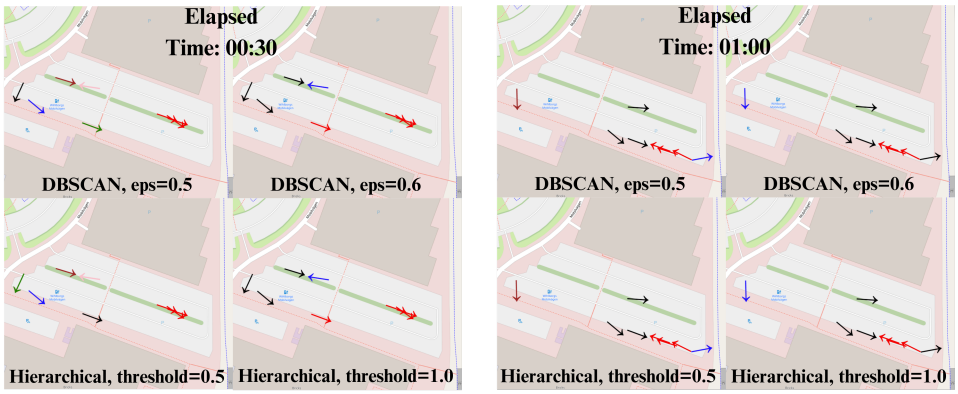


Fig. 8: NLoS clustering results based on position and heading

V. Conclusions And Future Work

By leveraging geographical positioning and heading direction, the proposed UE grouping framework can significantly enhance operational efficiency across various functional domains in 5G, some of which may not yet be fully realized. Future work could involve transitioning from 2-dimensional to 3-dimensional coordinate-based positioning, adding an extra dimension to the clustering method and making user grouping even more suitable for network functionalities such as beamforming. Integrating heading direction into user grouping and network management enables 5G systems to provide more intelligent location-based user grouping services.

REFERENCES

[1] A. Grenier, E. S. Lohan, A. Ometov, and J. Nurmi, "A survey on low-power GNSS," IEEE Commun. Surv. Tutor, vol. 25, no. 3, pp. 1482–1509, 2023.

[2] K. Ohno, T. Tsubouchi and S. Yuta, "Outdoor map building based on odometry and rtk-gps positioning fusion", IEEE International Conference on Robotics and Automation— 2004. Proceedings. ICRA '04. 2004, vol. 1, pp. 684-690, April 2004.

[3] H. Wymeersch, G. Seco-Granados, G. Destino, D. Dardari and F. Tufvesson, "5G mmWave Positioning for Vehicular Networks," IEEE Wireless Communications, vol. 24, no. 6, pp. 80-86, Dec. 2017.

[4] X. Cai, W. Fan, X. Yin, and G. F. Pedersen, "Trajectory-Aided Maximum-Likelihood Algorithm for Channel Parameter Estimation in Ultrawideband Large-Scale Arrays," IEEE Transactions on Antennas and Propagation, vol. 68, no. 10, pp. 7131-7143, May 2020.

[5] X. Li, E. Leitinger, M. Oskarsson, K. Åström, and F. Tufvesson, "Massive MIMO-based localization and mapping exploiting phase information of multipath components," IEEE Wireless Communications, vol. 19, no. 9, pp. 4254-4267, June 2019.

[6] M. M. Butt, A. Rao, and D. Yoon, "RF Fingerprinting and Deep Learning Assisted UE Positioning in 5G," IEEE 91st Vehicular Technology Conference (VTC2020-Spring), May 2020.

[7] D. Burghal, A. T. Ravi, V. Rao, A. A. Alghafis, and A. F. Molisch, "A Comprehensive Survey of Machine Learning Based Localization with Wireless Signals," [Online]. Available: https://arxiv.org/pdf/2012.11171.pdf, Dec. 2020.

[8] R. Whiton, J.Chen, and F. Tufvesson, "Wiometrics: Comparative Performance of Artificial Neural Networks for Wireless Navigation," IEEE Transactions on Vehicular Technology, Oct. 2024.

[9] S. Bartoletti, L. Chiaraviglio, S. Fortes, T. E. Kennouche, G. Solmaz, G. Bernini, D. Giustiniano, J. Widmer, R. Barco, G. Siracusano, A. Conti, N. B. Melazzi, "Location-Based Analytics in 5G and Beyond," IEEE Communications Magazine, vol. 59, no. 7, pp. 38-43, July 2021.

[10] D. Pjanić, A. Sopasakis, H. Tataria, F. Tufvesson, and A. Reial, "Learning-Based UE Classification in Millimeter-Wave Cellular Systems with Mobility," 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Oct. 2021.

[11] Ji-Hwan Choi, "Dynamic UE-Grouping Based Interference Management for Ultra-Dense Networks," 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Jan. 2019.

[12] G. Tian, D. Pjanić, X. Cai, B. Bernhardsson, and F. Tufvesson, "Attention-aided Outdoor Localization in Commercial 5G NR Systems," arXiv preprint arXiv:2405.09715, May 2024.

[13] 3GPP TR 38.104, "Base Station (BS) radio transmission and reception," Third Generation Partnership Project (3GPP), Dec. 2020.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, no. 521, pp. 436-444, May 2015.

[15] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks, vol. 2, no. 5, pp. 359-366, Mar. 1989.

[16] M. Ester, H-P. Kriegel, J. Sander, X. Xu, E. Simoudis, J. Han, U. M. Fayyad, "A density-based algorithm for discovering clusters in large spatial databases with noise," Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, pp. 226-231, Aug. 1996.

[17] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, pp. 236-244, Aug. 2012.

[18] B. Everitt, "Dictionary of Statistics," Cambridge, UK: Cambridge University Press. p. 96. ISBN 0-521-59346-8, 1998.

[19] K. E. Arslantürk, "User Equipment Grouping in 5G TDD System using Machine Learning," Master's thesis, Lund University, Sweden, June 2024.

Paper V

Paper V

Reproduced, with permission from IEEE

G. Tian, D. Pjanić, X. Cai, B. Bernhardsson, F. Tufvesson, "Attention-aided Outdoor Localization In Commercial 5G NR Systems," *IEEE Transactions on Machine Learning in Communications and Networking*, Vol. 2, Nov. 2024, doi: 10.1109/TMLCN.2024.3490496.

Attention-aided Outdoor Localization in Commercial 5G NR Systems

Guoda Tian[1,*], Member, IEEE, Dino Pjanić[1,2,*], Student Member, IEEE,
Xuesong Cai[1], Senior Member, IEEE,
Bo Bernhardsson[3], and Fredrik Tufvesson[1], Fellow, IEEE
[1] Dept. of Electrical and Information Technology, Lund University, Sweden
[2] Ericsson AB, Sweden
[3] Dept. of Automatic Control, Lund University, Sweden
[*] Equal contribution

Abstract

The integration of high-precision cellular localization and machine learning (ML) is considered a cornerstone technique in future cellular navigation systems, offering unparalleled accuracy and functionality. This study focuses on localization based on uplink channel measurements in a fifth-generation (5G) new radio (NR) system. An attention-aided ML-based single-snapshot localization pipeline is presented, which consists of several cascaded blocks, namely a signal processing block, an attention-aided block, and an uncertainty estimation block. Specifically, the signal processing block generates an impulse response beam matrix for all beams. The attention-aided block trains on the channel impulse responses using an attention-aided network, which captures the correlation between impulse responses for different beams. The uncertainty estimation block predicts the probability density function of the user equipment (UE) position, thereby also indicating the confidence level of the localization result. Two representative uncertainty estimation techniques, the negative log-likelihood and the regression-by-classification techniques, are applied and compared.

This work has been funded by Ericsson AB, the Swedish Foundation for Strategic Research, and partly by the Horizon Europe Framework Programme under the Marie Skłodowska-Curie grant agreement No. 101059091.

Furthermore, for dynamic measurements with multiple snapshots available, we combine the proposed pipeline with a Kalman filter to enhance localization accuracy. To evaluate our approach, we extract channel impulse responses for different beams from a commercial base station. The outdoor measurement campaign covers Line-of-Sight (LoS), Non Line-of-Sight (NLoS), and a mix of LoS and NLoS scenarios. The results show that sub-meter localization accuracy can be achieved.

Index Terms

5G New Radio, Sounding Reference Signal, self-attention, uncertainty estimation, radio-based positioning

I. INTRODUCTION

RADIO-based positioning is envisioned to pave the way for numerous sophisticated yet practical applications, including vehicle navigation, intelligent traffic management, and autonomous driving [1]–[7]. In contemporary 5th generation mobile network (5G) systems, there is a pronounced demand for precise localization capabilities. Currently, most localization-aware applications are facilitated by Global Navigation Satellite Systems (GNSS). However, the effectiveness of these systems is limited by many factors, such as shadowing, multipath propagation, and clock drifts between the GNSS transmitter and receiver [8]. Consequently, there is an increasing need to investigate cellular-based technologies and seamlessly integrate those techniques into existing localization systems.

Existing cellular-based localization methods can be broadly classified into two categories, namely conventional signal processing methods [7], [9]–[15], and machine learning (ML) based methods [16]–[24]. Conventional signal processing methods, such as Time of Arrival (ToA), Angle of Arrival (AoA), and Time Difference of Arrival (TDoA), require the estimation of essential channel parameters, such as signal propagation time between user equipment (UE) and base stations (BS). In the next step, the location of the UE can be estimated using these parameters. Although some of these methods have reached maturity, they can be constrained by calibration needs and algorithmic complexities [7]. On the other hand, ML methods present a promising solution but require access to data for training and a radio environment with enough unique features that can be learned. To implement an ML-based localization approach, the initial step involves obtaining various channel

fingerprints, such as the raw transfer function [21], [22], received signal strength [16], angle-delay spectrum [17], [20], [23] and/or covariance matrix [18], [19]. These fingerprints then serve as input for the ML algorithms. It should be noted that an effective method of combining several different fingerprints has the potential to significantly increase the localization accuracy, see [18], [19]. ML-based localization algorithms can also be divided into two categories, namely classical ML approaches such as K-nearest neighbors (KNN) [19], Gaussian process regression [16], adaptive boosting [18], and deep learning based approaches, such as fully connected neural networks (FCNN) [22], [23], convolutional neural network [16], [18], [24], [25], and attention-aided networks [21]. In particular, the attention-aided approach holds significant promise, as its embedded attention mechanism enables ML algorithms to recognize relationships between different input feature vectors, irrespective of their actual spatial or temporal separation among those vectors. This mechanism is also the core of widely used transformer techniques, producing fruitful results in various domains such as language translations, image recognition, and speech recognition [26]. Another crucial aspect for localization is uncertainty prediction, which is particularly important in life-critical tasks such as autonomous driving. This research problem has been initially tackled by previous works [19], [27], which provide not only the estimated location coordinates but also the corresponding variances using the negative log-likelihood (NLL) loss function.

However, to the best of our knowledge, there are still notable research gaps. Primarily, the application of attention-aided localization algorithms in 5G new radio (NR) systems represents a novel, yet unexplored, area. Secondly, the NLL uncertainty estimation technique assumes a Gaussian distribution for the estimation error of the UE position. However, such an assumption often diverges from reality. Consequently, it becomes crucial to explore further uncertainty estimation methods capable of estimating distributions other than Gaussian. To address the issues stated above, we propose a novel localization pipeline and evaluate it using data from a commercial 5G NR BS. Very few studies in the literature have been conducted on commercial grade 5G NR systems. Our research contributions are listed as follows:

- We apply attention-aided neural networks as the backbone to perform localization, we also demonstrate the advantages of this network in terms of localization accuracy.
- We apply a novel regression-by-classification method that can predict the uncer-

tainty of localization estimates. Compared with the NLL approach, this approach provides better uncertainty estimation since it is not bounded by the assumption of Gaussianity.
- We further enhance localization accuracy by applying a Kalman filter to exploit temporal correlation between multiple channel snapshots, which smoothes the estimated trajectory.
- Finally, we verify the novel ML-powered pipeline with real measurement data obtained using a commercial 5G NR test setup, covering both Line-of-Sight (LoS) and non-Line-of-Sight (NLoS) scenarios. The results show that our approach achieves submeter-level localization accuracy.

Our initial outdoor UE localization results have been presented in the conference paper [28]. Differ from [28], we utilize a higher subchannel resolution of the UL SRS channel estimates and a high-accuracy GNSS receiver. Furthermore, we apply more advanced ML approaches such as attention mechanisms and uncertainty estimation algorithms. Compared with [28], the localization accuracies have significantly improved.

The remainder of this paper is organized as follows. Section II introduces the signal model and discusses the selected fingerprints. In Section III, we elaborate on the localization algorithms. Section IV illustrates the measurement campaign and Section V presents the results. Finally, conclusive remarks are included in Section VI.

II. System model and dataset generation

We consider a commercial 5G NR system in a single-user massive Multiple-Input Multiple-Output (MIMO) scenario, where the BS processes uplink (UL) Sounding Reference Signal (SRS) data. The system utilizes orthogonal frequency division multiplexing (OFDM) with F subcarriers, and the SRS data is a time series of UL measurements in the beam domain. With this approach, we essentially capture the angular delay spectrum of the radio channel, an approach that has been shown to be advantageous for accurate localization based on ML [20], [29]. The BS is equipped with M_{BS} antenna ports, half of which is vertically polarized and the other half horizontally polarized, while the UE is equipped with M_{UE} antenna ports. We suppose that the number of multipath components is P, and denote $\tau_{p,t}$ as the time delay between UE and BS w.r.t. the p-th path at time t, and $\alpha_{p,m,t}$ indicates the complex coefficient of each multipath component. The BS utilizes all vertical-

polarized antennas to formulate N_{V} beams, the response of the i-th beam w.r.t. the p-th path is $\beta_{\mathrm{V},i}(\phi_p, \theta_p, f)$, where f denotes frequency, and ϕ_p and θ_p represent the azimuth and elevation arrival angles for the p-th multipath, respectively. Similarly, another N_{H} set of beams uses all horizontal polarized antennas, and the response of the i-th beam is $\beta_{\mathrm{H},i}(\phi_p, \theta_p, f)$. For the m-th UE port, the propagation channel model for each beam at time index t can be formulated as

$$\begin{aligned} h_{\mathrm{V},i,m,t}(f) &= \sum_{p=1}^{P} \beta_{\mathrm{V},i}(\phi_p, \theta_p, f)\, \alpha_{p,m,t}\, \exp\{-j2\pi f \tau_{p,t}\} \\ h_{\mathrm{H},i,m,t}(f) &= \sum_{p=1}^{P} \beta_{\mathrm{H},i}(\phi_p, \theta_p, f)\, \alpha_{p,m,t}\, \exp\{-j2\pi f \tau_{p,t}\}. \end{aligned} \tag{1}$$

By collecting all $h_{\mathrm{V},i,m,t}(f)$ and $h_{\mathrm{H},i,m,t}(f)$ for the F subcarriers, we can formulate two beam space matrices of the channel transfer function (CTF), $\mathbf{H}_{\mathrm{V},m,t} \in \mathbb{C}^{N_V \times F}$ and $\mathbf{H}_{\mathrm{H},m,t} \in \mathbb{C}^{N_{\mathrm{H}} \times F}$ at time t, which correspond to the vertical and horizontal polarized antenna groups, respectively. We further define matrix $\mathbf{H}_t \in \mathbb{C}^{N \times F} = \left[\mathbf{H}_{\mathrm{H},1,t}^T, \mathbf{H}_{\mathrm{V},1,t}^T, ..., \mathbf{H}_{\mathrm{H},M_{\mathrm{UE}},t}^T, \mathbf{H}_{\mathrm{V},M_{\mathrm{UE}},t}^T\right]^T$ that combines channel matrices of all UE antenna ports. Specifically, $N = M_{\mathrm{UE}}(N_{\mathrm{H}} + N_{\mathrm{V}})$. This matrix depends strongly on the UE position, therefore they can be selected as raw channel fingerprints to perform ML-based localization.

III. The ML-based localization approach

In this paper, our study focuses on car navigation applications, where two-dimensional (2-D) localization is adequate for most scenarios. However, a similar approach can be extended to three-dimensional (3-D) coordinate-based localization by altering the dimension of the output layer in our neural network. The ML-based localization pipeline, as described in Fig. 1, consists of five sequential blocks. First, the raw CTF $\mathbf{H}_t$ is fed into a data cleaning block to evaluate the validity of the input data. After this, valid CTFs are forwarded to a digital signal processing block to generate an impulse response beam matrix $\mathbf{G}_t \in \mathbb{C}^{N \times F}$. The amplitudes in this matrix then serve as input to a deep neural network, which incorporates a self-attention mechanism at its core. The network's final layer outputs an estimated probability density function (PDF) representing the location, thereby facilitating uncertainty estimation. To further enhance localization accuracy, a filter may be

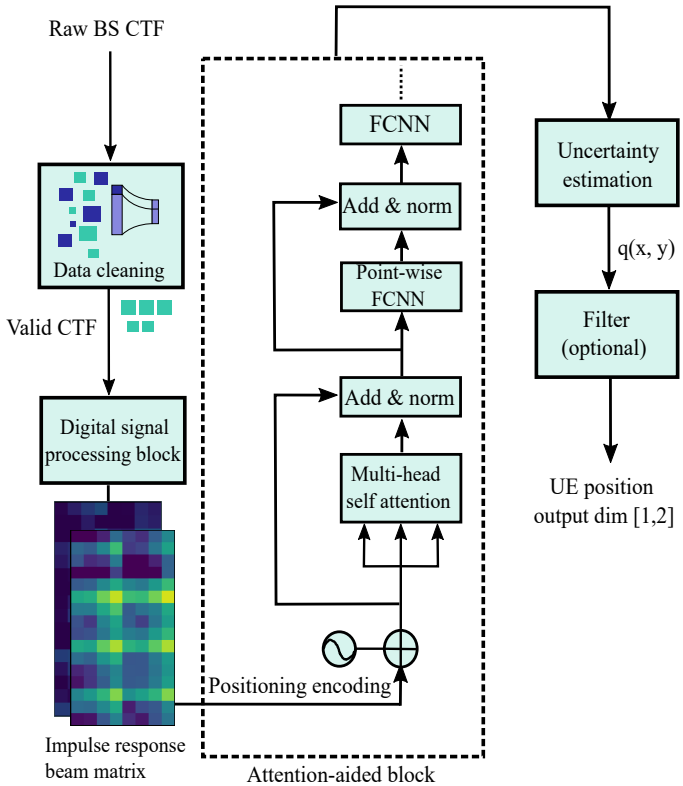


Fig. 1: The ML-based localization pipeline for the 5G NR system.

applied after the final layer of the pipeline, provided that information from multiple snapshots is available.

A. The attention mechanism

1) Fundamental basics of the attention operation

An example of the attention block is illustrated in Fig. 2, which takes a matrix $\mathbf{X} \in \mathbb{R}^{A_1 \times A_2}$ as the input, generating the output matrix $\mathbf{Z} \in \mathbb{R}^{A_1 \times A_3}$. Initially, $\mathbf{X}$ are multiplied by three matrices, namely, the query matrix $\mathbf{W}_q \in \mathbb{R}^{A_2 \times A_3}$, the key matrix $\mathbf{W}_k \in \mathbb{R}^{A_2 \times A_3}$ and the value matrix $\mathbf{W}_v \in \mathbb{R}^{A_2 \times A_3}$. The multiplication operations

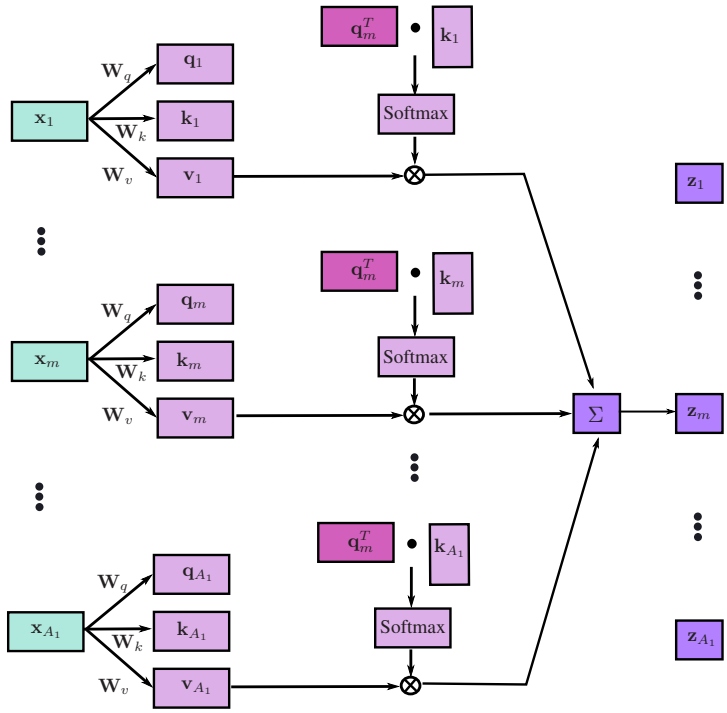


Fig. 2: An illustration of basic attention mechanism to generate $\mathbf{z}_j$ and same mechanism can be applied to generate $\mathbf{Z}$.

yield three matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{A_1 \times A_3}$, specifically,

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_q,\ \mathbf{K} = \mathbf{X}\mathbf{W}_k,\ \mathbf{V} = \mathbf{X}\mathbf{W}_v. \tag{2}$$

In the self-attention mechanism, the query ($\mathbf{Q}$) and key $\mathbf{K}$ matrices play a crucial role in determining the relevance of each row vector in the matrix $\mathbf{X}$ to the other row vectors. The elements of these three matrices act as hyperparameters that can be fine-tuned during the training process. The second step is to calculate the pairwise correlations between all columns of matrices $\mathbf{Q}$ and $\mathbf{K}$, resulting in a new matrix $\mathbf{A} \in \mathbb{R}^{A_3 \times A_3}$, specifically,

$$\mathbf{A} = \frac{1}{\sqrt{A_2}}\mathbf{Q}^T\mathbf{K}. \tag{3}$$

These correlations reflect the similarities between each pair of row vectors in $\mathbf{X}$. We then apply the *softmax* operation to normalize $\mathbf{A}$ and obtain another matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{A_3 \times A_3}$. Each element $\tilde{\mathbf{A}}_{i,j}$ is positive and the sum of all the elements in each column is equal to 1. Specifically, $\tilde{\mathbf{A}}_{i,j}$ is calculated as

$$\tilde{\mathbf{A}}_{i,j} = \frac{\exp \mathbf{A}_{i,j}}{\sum_k \exp \mathbf{A}_{i,k}}. \tag{4}$$

Finally, the output matrix $\mathbf{Z}$ is calculated as

$$\mathbf{Z} = \mathbf{V}\tilde{\mathbf{A}}, \tag{5}$$

where each column of $\mathbf{Z}$ represents a weighted sum, and the weights are determined by the corresponding column in $\tilde{\mathbf{A}}$. In addition to the fundamental attention operation, we further introduce the *multi-head attention* mechanism that can improve model capabilities. This mechanism employs a total of $\mathcal{P}$ attention heads, each associated with sets of query matrices $(\mathbf{W}_q^1, ..., \mathbf{W}_q^{\mathcal{P}})$, key matrices $(\mathbf{W}_k^1, ..., \mathbf{W}_k^{\mathcal{P}})$, and value matrices $(\mathbf{W}_v^1, ..., \mathbf{W}_v^{\mathcal{P}})$. The multi-head attention mechanism operates in $\mathcal{P}$ steps. In the initial step, the matrices $\mathbf{W}_q^1, \mathbf{W}_k^1, \mathbf{W}_v^1$ are applied to the input matrix $\mathbf{X}$ following equations (2)-(5), resulting in the output $\mathbf{Z}_1 \in \mathbb{R}^{A_1 \times A_3}$. This process is then repeated $\mathcal{P} - 1$ times, generating additional output matrices $\mathbf{Z}_2, ..., \mathbf{Z}_P \in \mathbb{R}^{A_1 \times A_3}$. Finally, we concatenate all output matrices obtained from each step, formulating a matrix $\mathbf{Z}_{tl} \in \mathbb{R}^{A_1 \times \mathcal{P}A_3}$. The final output matrix $\mathbf{Z}' \in \mathbb{R}^{A_1 \times A_2}$ can then be expressed as

$$\mathbf{Z}' = \mathbf{Z}_{tl}\mathbf{W}_O, \tag{6}$$

where $\mathbf{W}_O \in \mathbb{R}^{\mathcal{P}A_3 \times A_2}$ is another hyperparameter matrix.

2) Positioning encoding

It is important to note that the attention mechanism neglects the inherent sequence order of the input vectors in $\mathbf{X}$. Consequently, when employing such a mechanism, particularly for tasks dependent on the order of vector arrangement, it is imperative to apply a *positioning encoding* technique to incorporate and preserve this sequential information. The idea of positioning encoding is to add another fixed matrix $\mathbf{X}_k \in$

$\mathbb{R}^{A_1 \times A_2}$ to $\mathbf{X}$ [26], a standardized positioning encoding matrix $\mathbf{X}_k$ is

$$\begin{aligned} \mathbf{X}_k(x, y) &= \sin\left(\frac{x}{10000^{y/A_2}}\right), \text{for odd } y; \\ \mathbf{X}_k(x, y) &= \cos\left(\frac{x}{10000^{(y-1)/A_2}}\right), \text{for even } y. \end{aligned} \tag{7}$$

The matrix $\mathbf{X}_k$ is fixed and will not be fine-tuned during the training process. The advantages of using cosine and sine structures are as follows:

- The values of the sine and cosine functions are bounded between -1 and 1, providing stable input magnitudes for the model.
- The smooth variation of sine and cosine functions allows the model to capture gradual changes in positions.
- The use of sine and cosine functions, as given by Eq. (7), ensures that each position is uniquely encoded.

3) Residual mechanism, Layer normalization and position-wise FCNN

After collecting the matrix $\mathbf{Z}'$, we add the input matrix $\mathbf{X}$ to $\mathbf{Z}'$ to obtain the matrix $\tilde{\mathbf{Z}} \in \mathbb{R}^{A_1 \times A_2}$. We apply the *residual mechanism* since it preserves the original information of the input matrix. The matrix $\tilde{\mathbf{Z}}$ is then fed to a *layer normalization block*, which first vectorizes $\tilde{\mathbf{Z}}$ into a vector $\tilde{\mathbf{z}} \in \mathbb{R}^{A_1 A_2}$. Subsequently, each element $\tilde{z}_i$ in $\tilde{\mathbf{z}}$ is scaled to derive a new vector $\hat{\mathbf{z}} \in \mathbb{R}^{A_1 A_2}$ as in [26], specifically,

$$\hat{z}_i = \gamma \frac{\tilde{z}_i - \mu}{\sigma} + \beta, \tag{8}$$

where μ and σ^2 represent the mean and variance of vector $\tilde{\mathbf{z}}$. The parameters γ and β denote the amplitude scaling and the bias, respectively. By default, $\gamma = 1$ and $\beta = 0$, although these parameters can be adjusted as learning hyperparameters. We then reformulate $\hat{\mathbf{z}}$ into a matrix $\hat{\mathbf{Z}} \in \mathbb{R}^{A_1 \times A_2}$. To enhance the capacity to capture nonlinear relationships, we feed the output matrix $\hat{\mathbf{Z}}$ into a pointwise FCNN to get $\hat{\mathbf{Z}}' \in \mathbb{R}^{A_1 \times A_2}$ [26], specifically,

$$\hat{\mathbf{Z}}' = \mathbf{W}_2 f_{\text{Relu}}(\mathbf{W}_1 \hat{\mathbf{Z}} + \mathbf{B}_1) + \mathbf{B}_2, \tag{9}$$

where $f_{\text{Relu}}(.)$ represents the rectifier activation function, and $\mathbf{W}_1, \mathbf{W}_2, \mathbf{B}_1, \mathbf{B}_2$ are hyperparameter matrices, and the bias matrices $\mathbf{B}_1, \mathbf{B}_2$ are optional. After collecting $\hat{\mathbf{Z}}'$, we apply the same residual mechanism and layer normalization to derive $\breve{\mathbf{Z}} \in$

$\mathbb{R}^{A_1 \times A_2}$. Finally, $\breve{\mathbf{Z}}$ is vectorized and fed into another FCNN. Such an operation can also help to match the vector sizes for possible subsequent blocks.

B. Data cleaning and signal processing

The collection of UL SRS channel measurements in a commercial 5G NR BS builds limitations when retrieving data-intense structures such as SRS channel measurement samples. The vast amounts of SRS data generated at milliseconds level are normally enclosed within the baseband entity of a BS and primarily intended for internal processing, whereas external access to these data sets may be compromised by hardware and software restrictions. To mitigate these challenges, it is essential to equip our pipeline with the ability to discern the validity of the input data. As retrieving all the necessary data in a complete format has been challenging, we implemented a threshold that defines a cut-off point for discarding datasets when insufficient information has been retrieved from the BS and introduced a *data-cleaning* block to pre-process the measurement data. Its primary objective is to determine whether the raw transfer function is valid or invalid. A raw transfer function is labeled invalid under the following conditions:

- Insufficient CSI in the beam or frequency domain: the number of non-zero elements in $\mathbf{H}_t$ is lower than a given threshold.
- Update failure: the values of all subcarriers or all beams remain the same.

After filtering out all invalid data, the next step is to process the raw CTF to generate impulse response beam matrices. To suppress the side lobes, we apply Hann windowing across all rows of the matrix $\mathbf{H}_t$ to obtain matrix $\hat{\mathbf{H}}_t \in \mathbb{C}^{N \times F}$. The F-length Hann window in the frequency domain is given by

$$w[f] = \sin^2\left(\frac{\pi f}{F}\right), \quad f = 0, \ldots, F-1. \tag{10}$$

After the windowing operation, the impulse response beam matrix $\mathbf{G}_t$ is produced by performing the inverse discrete Fourier transform along each row of $\hat{\mathbf{H}}_t$. Given the potential difficulty in achieving a stable phase for $\mathbf{G}_t$, here we opt to use its amplitude $|\mathbf{G}_t|$ as the training fingerprint, although this means throwing away potentially useful information.

C. Single-snapshot localization

We hereby introduce our single-snapshot localization approach, which focuses on performing the localization task using only a single channel sample of the received transfer function, generated at an SRS reporting periodicity of 20 ms containing 64 symbols, transformed from antenna to beam space. The proposed positioning model analyzes the time series of these samples. As illustrated in Fig. 1, the architecture comprises multiple attention-aided blocks, followed by an output layer that has three alternatives corresponding to three loss functions, namely the Mean Square Error (MSE), NLL, and Regression-by-classification loss functions. We use $\mathbf{p}_i = [p_{x,i}, p_{y,i}]^T$ to represent the 2-D ground truth of the moving UE at the i-th position. Notably our approach can be readily adapted for 3-D localization.

1) Alternative 1: MSE loss function

This approach directly estimates the UE locations by setting a 2-D regression head at the output layer of the last attention block. Let $f_{\text{MSE}}(.)$ denote the overall function and vector $\boldsymbol{\theta}_2$ all hyperparameters, $\hat{\mathbf{p}}_i = [\hat{p}_{x,i}, \hat{p}_{y,i}]^T$ the estimated i-th UE locations generated by $f_{\text{MSE}}(\boldsymbol{\theta}_2, |\mathbf{G}_t|)$, the loss Ψ_1 can be expressed as

$$\Psi_1 = \frac{1}{N_{tr}} \sum_{i \in \Omega'_{tr}} ||\mathbf{p} - \hat{\mathbf{p}}||_F^2, \tag{11}$$

where Ω'_{tr} and N_{tr} denote the training set and the number of training samples, respectively, and $||.||_F$ denotes the Frobenius matrix norm.

2) Alternative 2: NLL loss function

Unlike the first approach, this method employs the NLL criterion, which models the estimated UE position as a multivariate Gaussian distribution defined by its mean $\breve{\mathbf{p}} = [\breve{p}_{x_i}, \breve{p}_{y_i}]^T$ and variance $\breve{\boldsymbol{\sigma}}_i^2 = [\breve{\sigma}_{x_i}^2, \breve{\sigma}_{y_i}^2]^T$. Consequently, a 4-dimensional regression head is required at the output layer. Similar to [19], the NLL loss Ψ_2 is expressed as

$$\Psi_2 = \frac{1}{2N_{tr}} \sum_{i \in \Omega'_{tr}} \Big(\frac{\log \breve{\sigma}_{x_i}^2 \breve{\sigma}_{y_i}^2}{2} + \frac{(p_{x_i} - \breve{p}_{x_i})^2}{2\breve{\sigma}_{x_i}^2} + \frac{(p_{y_i} - \breve{p}_{y_i})^2}{2\breve{\sigma}_{y_i}^2} \Big). \tag{12}$$

3) Alternative 3: Regression-by-Classification (RbC)

The core of this approach [30], [32] lies in converting a regression task to a classification task. This is achieved by first defining a feasible range for the target parameter and then dividing this range into discrete bins. For the localization task,

the lower and upper bounds of the UE x-coordinates are denoted as $B_{lw,x}$ and $B_{up,x}$, respectively. Similarly, $B_{lw,y}$ and $B_{up,y}$ represent the bounds for the y-coordinates. To accomplish this discretization, we divide the x-coordinate range into L_x equally sized bins. The y-coordinate range is divided into L_y bins in a similar fashion. For each bin, we denote $\bar{l}_{x,k}$ and $\bar{l}_{y,k}$ as the lower endpoint values of the k-th interval for the x- and y-coordinates, respectively.

Unlike the NLL method, RbC does not inherently model the output probability as a Gaussian distribution. Instead, it estimates the probability and bias values of each bin for both the x- and y-coordinates. The bias value can be used to reduce the quantization error. To this end, in total 4 vectors are generated: the probability vectors $\boldsymbol{\omega}_x \in \mathbb{R}^{L_x}$ and $\boldsymbol{\omega}_y \in \mathbb{R}^{L_y}$, as well as the deviation vectors $\mathbf{d}_x \in \mathbb{R}^{L_x}$ and $\mathbf{d}_y \in \mathbb{R}^{L_y}$. Note that $\boldsymbol{\omega}_x$ refers to the probability vector corresponding to a specific position, and there are N_{tr} instances of $\boldsymbol{\omega}_x$ when considering the entire training dataset; the same applies to $\boldsymbol{\omega}_y$. It is crucial to apply a *softmax* operation as shown in (4) when generating $\boldsymbol{\omega}_x$ and $\boldsymbol{\omega}_y$ to ensure that the elements within each vector sum to 1. One special case for deviation vectors is when all L_x elements in $\mathbf{d}_x$ have the same value, and the same for $\mathbf{d}_y$. In other words, a uniform shift is applied to the probability density function, which also aids in the reduction of the output vector dimensions. We denote $\omega_{x,k}$ and $d_{x,k}$ as the k-th elements of $\boldsymbol{\omega}_x$ and $\mathbf{d}_x$, similarly for $\omega_{y,k}$ and $d_{y,k}$. Inspired by [30], the η-norm loss Ψ_3^η is formulated as

$$\Psi_3^\eta = \frac{1}{2N_{tr}} \sum_{i \in \Omega'_{tr}} \Big(|| \sum_{j=1}^{L_x} \omega_{x,j,i} \bar{l}_{x,j,i} - p_{x,j,i} + d_{x,j,i} ||^\eta + || \sum_{j=1}^{L_y} \omega_{y,j,i} \bar{l}_{y,j,i} - p_{y,j,i} + d_{y,j,i} ||^\eta + \gamma_1 ||\mathbf{d}_x|| + \gamma_2 ||\mathbf{d}_y|| \Big). \tag{13}$$

Here, η is usually chosen as $\eta = 1$ or $\eta = 2$, which corresponds to the *Taxicab* and *Euclidean* norms, respectively. Two penalty terms, $\gamma_1 ||\mathbf{d}_x||$ and $\gamma_2 ||\mathbf{d}_y||$, are added to the cost function. The estimated coordinate $\hat{\mathbf{p}}_i^{\text{RbC}} = [\hat{p}_{x,i}^{\text{RbC}}, \hat{p}_{y,i}^{\text{RbC}}] \in \mathbb{R}^2$ is then

given by

$$\hat{p}_{x,i}^{\text{RbC}} = \sum_j \omega_{x,j,i} \bar{l}_{x,j,i} + d_{x,j,i},$$
$$\hat{p}_{y,i}^{\text{RbC}} = \sum_j \omega_{y,j,i} \bar{l}_{y,j,i} + d_{y,j,i}. \tag{14}$$

4) Comparison between different uncertainty estimates

Our previous work [19] used the NLL score in the test data set to assess the effectiveness of uncertainty estimation. However, applying the same criterion to evaluate the RbC method presents challenges because of the non-Gaussian nature of its output. To address this challenge, another criterion named Area Under the Sparsification Error (AUSE) [31], [33] is used. Sparsification is a way to assess the quality of uncertainty estimates. It works by progressively discarding fractions of the predictions that the model is most uncertain about and verifying whether this corresponds to a proportional decrease in the remaining average endpoint error. To calculate AUSE, the first step is to compute the discrete entropy u_H based on the predicted probability. In the following discussion, we illustrate this process using the predicted $\boldsymbol{\omega}_{x,i}$ vector for the x-coordinate as an example, noting that the result can be readily extended to the y-coordinate. The entropy $u_{H,x,i}$ for $\boldsymbol{\omega}_{x,i}$ is given by [32]

$$u_{H,x,i}(\boldsymbol{\omega}_{x,i}) = -\sum_{k=1}^{L_x} \omega_{x,k,i} \log \omega_{x,k,i}. \tag{15}$$

To enable a fair comparison between the NLL and RbC methods, we need to discretize the predicted Gaussian distributions determined by $\breve{\mathbf{p}}$ and $\breve{\sigma}_i^2$. To this end, the x-axis is segmented into L_x bins. As detailed in [34], the value for the k-th bin of the discretized function, denoted $\breve{p}_{x,k}$, is calculated as

$$\breve{p}_{x,k}^{ds} = \frac{\frac{1}{\breve{\sigma}_k} \exp(-\frac{(\breve{p}_{x,k} - \bar{l}_k)^2}{2\breve{\sigma}_k^2})}{\sum_j \frac{1}{\breve{\sigma}_j} \exp(-\frac{(\breve{p}_{x,j} - \bar{l}_j)^2}{2\breve{\sigma}_j^2})}. \tag{16}$$

We now organize the discrete entropies for the N_{ts} testing samples calculated from (15) in descending order to form the vector $\mathbf{u}_{H,x} \in \mathbb{R}^{L_x}$. Similarly, we calculate the absolute errors between the estimated values $\hat{p}_{x,i}^{\text{Ws}}$ and the ground truth p_x for all testing samples, arranging these errors in descending order to create the vector

$\boldsymbol{\xi}_x \in \mathbb{R}^{L_x}$. Let ξ_{max} be the maximum absolute error. We scale all elements in $\mathbf{u}_{H,x}$ by a factor such that the first element of the resulting vector $\hat{\mathbf{u}}_{H,x}$ equals ξ_{max}.

Next, we define a *sparsification* function $s(\varphi)$, which is calculated by removing the initial φ-fraction of samples from $\hat{\mathbf{u}}_{H,x}$ and averaging the remaining data, with φ ranging from 0 to 1. A similar process is applied to $\boldsymbol{\xi}_x$, which yields the oracle function $g(\varphi)$. Finally, AUSE is calculated as

$$\text{AUSE} = \int_0^1 |s(\varphi) - g(\varphi)| d\varphi, \tag{17}$$

which represents the area between the sparsification and the oracle curves. A smaller area indicating a better uncertainty estimator.

D. Kalman-Filter-based trajectory smoothing

To further improve the localization accuracy, we exploit the temporal correlation between successive positions by applying a Kalman filter as a straightforward method for trajectory smoothing. The BS can select the appropriate motion model based on several factors, including the sampling rate of channel snapshots, the vehicle's velocity, and the availability of velocity or acceleration parameters. In this paper, we introduce the constant velocity model as the simplest option. This model is effective for scenarios with low vehicle speed and a high sampling rate. However, the same concepts can be extended to more advanced models that account for changes in velocity or even acceleration. While these advanced models may deliver better performance, especially in high-speed scenarios, they also require more complex hardware. More detailed information see be refered to [35].

We define a vector $\boldsymbol{\xi}_t \in \mathbb{R}^4 = [p_{x,t}, v_{x,t}, p_{y,t}, v_{y,t}]^T$ to represent the UE position and velocity at time t, where $v_{x,t}$ and $v_{y,t}$ denote the speed in the x and y-directions, respectively. The state-space model for the UE is given by

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \boldsymbol{\lambda}_t, \tag{18}$$

where $\mathbf{F} \in \mathbb{R}^{4\times4}$ denotes the state-transition matrix, while $\boldsymbol{\lambda}_t \in \mathbb{R}^4$ the additive

noise. Specifically,

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & \Delta_t & 0 \\ 0 & 1 & 0 & \Delta_t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{19}$$

where Δ_t denotes the time differences between snapshots. We then define $\boldsymbol{\Xi}_t \in \mathbb{R}^{4\times4}$ as the covariance matrix of $\boldsymbol{\xi}_t$. The relationship between $\boldsymbol{\Xi}_t$ and $\boldsymbol{\Xi}_{t-1}$ can be written as

$$\boldsymbol{\Xi}_t = \mathbf{F}\boldsymbol{\Xi}_{t-1}\mathbf{F}^T + \boldsymbol{\Lambda}, \tag{20}$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{4\times4}$ is the covariance matrix of the noise vector $\boldsymbol{\lambda_t}$. We further denote $\breve{\mathbf{p}}_t \in \mathbb{R}^2 = [\breve{p}_{t,x}, \breve{p}_{t,y}]$ as the predicted UE position and express the observation model as

$$\breve{\mathbf{p}}_t = \boldsymbol{\Phi}_t\boldsymbol{\xi}_t + \boldsymbol{\zeta}, \tag{21}$$

where $\boldsymbol{\zeta} \in \mathbb{R}^2$ represents observation noise and $\boldsymbol{\Phi}_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$. Given the error signal $\mathbf{e}_t = \hat{\mathbf{p}}_t - \breve{\mathbf{p}}_t$, the state vector $\boldsymbol{\xi}_t^+$ is updated as

$$\boldsymbol{\xi}_t^+ = \boldsymbol{\xi}_t + \boldsymbol{\Gamma}_t\mathbf{e}_t. \tag{22}$$

In (22), $\boldsymbol{\Gamma}_t$ represents the Kalman gain matrix, which balances the predictions from the state-space model and the ML-based pipeline, specifically,

$$\boldsymbol{\Gamma}_t = \boldsymbol{\Xi}_t\boldsymbol{\Phi}_t^T\left[\boldsymbol{\Phi}_t\boldsymbol{\Xi}_t\boldsymbol{\Phi}_t^T + \mathbf{R}\right]^{-1}, \tag{23}$$

where $\mathbf{R}$ is the covariance matrix of $\boldsymbol{\zeta}$. After computing $\boldsymbol{\Gamma}_t$, the covariance matrix $\boldsymbol{\Xi}_t$ is updated using

$$\boldsymbol{\Xi}_t^+ = (\mathbf{I} - \boldsymbol{\Gamma}_t\boldsymbol{\Phi}_t)\boldsymbol{\Xi}_t, \tag{24}$$

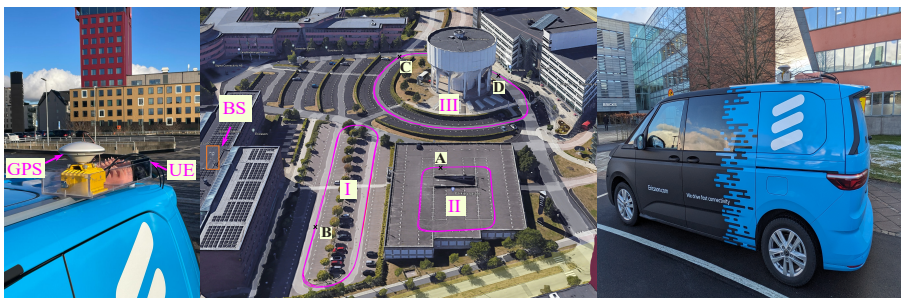
where $\mathbf{I}$ denotes the identity matrix. By applying the process outlined by (18)-(24), we can significantly mitigate the impact of prediction outliers, as will be further illustrated in Section V.

IV. Outdoor 5G NR measurement campaign

To evaluate our localization pipeline, an outdoor vehicular measurement campaign was conducted at a parking lot outside of the Ericsson office in Lund, Sweden.

Photos of the test vehicle, the BS antenna, the UE as well as the measurement areas are presented in Fig. 3.

A. Introduction to the measurement campaign



(a) (b) (c)

Fig. 3: The 5G NR BS was equipped with an antenna integrated radio with 64 transmitters and receivers, placed on top of a 20 m high building. In this measurement campaign, a vehicle moves along three pre-defined routes: **I** A route on a 10 meter-high garage for LOS measurements. **II**: A ground-level route for NLoS measurements below the building of the BS. **III**: A ground-level route for combined LoS and NLoS measurements. (a) GPS and UE, (b) Measurement Scenario, (c) Measurement van.

During the measurement campaign, the test vehicle carried a GNSS receiver, and a commercial UE, see Fig. 3(a). Centimeter-level ground truth positioning accuracy was achieved using a Swift Duro high-performance GNSS receiver with real time kinematics technology, GNSS multi-band and multi-constellation support. To ensure that the UE remained in connected state, it simultaneously downloaded data at a 750 Mbit/s rate enabling continuous SRS UL transmission. The UL SRS pilot signals were received and processed by a commercial Ericsson 5G BS operating in mid-band at 3.85 GHz center frequency. The BS was compliant to the 5G NR 3rd Generation Partnership Project (3GPP) standard 38.104 Rel15 [36] and equipped with a time division duplexing (TDD) antenna integrated radio with 64 transmitters/receivers (TX/RX) consisting of 32 dual-polarized antennas covering a 120 degree sector. As for digital beam forming, 64 TX/RX formulate 64 beams in downlink (DL) and UL respectively. As illustrated in Fig. 4, the SRS channel estimates are reported for 273 physical resource blocks (PRBs) over a 100 MHz bandwidth. Each channel snapshot

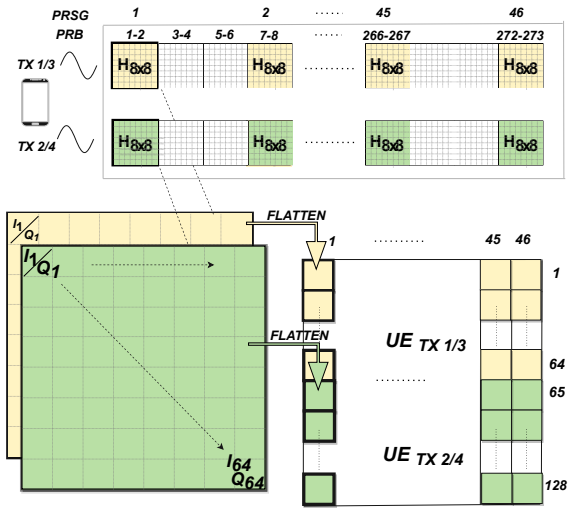


Fig. 4: SRS data collection and CTF generation. I/Q means in-phase/quadrature.

contains the 273 PRBs for all 64 beams. The PRBs are grouped and averaged in pairs, resulting in 137 Physical Resource Blocks Sub Groups (PRSG). Down sampling was done so that every third PRSG was further used generating 46 PRSGs in total. The UE was equipped with 4 antenna ports, i.e. 4 UE layers, sounding SRS pilots. Due to the capacity of our data-streaming system, the BS recorded the channel responses of 2 UE antenna ports which formulate two channel transfer function matrices $\mathbf{H}_1, \mathbf{H}_2 \in \mathbb{C}^{N \times F}$. We define a matrix $\mathbf{H}' \in \mathbb{C}^{2N \times F}$ to collect those two matrices, specifically, $\mathbf{H}' = [\mathbf{H}_1, \mathbf{H}_2]$ ($N = 64, F = 46$). As illustrated in Fig. 3, our measurement campaign comprises three distinct scenarios: LoS, NLoS, and a mixed scenario. In all scenarios, the velocity of the vehicle is approximately 15 km/h. The trajectory for each of the three measurement scenarios consists of 5 laps. In the LoS scenario, the test vehicle drove at an open parking lot, while in the NLoS scenario, the vehicle was driving next to a tall building that obstructed the LoS path. As for the mixed scenario, NLoS conditions occurred when the LoS was blocked by the water tower. For all three measurements, the BS station recorded channel snapshots with 20 ms periodicity, resulting in $\mathcal{T}_1 = 22000$, $\mathcal{T}_2 = 24603$

and $\mathcal{T}_3 = 27087$ channel snapshots. We formulate three tensors $\mathcal{A}_{\text{LoS}} \in \mathbb{C}^{\mathcal{T}_1 \times 2N \times F}$, $\mathcal{A}_{\text{NLoS}} \in \mathbb{C}^{\mathcal{T}_2 \times 2N \times F}$, $\mathcal{A}_{\text{mix}} \in \mathbb{C}^{\mathcal{T}_3 \times 2N \times F}$ to collect all snapshots. Those three tensors are normalized by multiplying each with a scalar so that their Euclidean norms equals $\mathcal{T}_i MN$, where $i = 1, 2, 3$.

B. Measured propagation channel characteristics

In Fig. 5, we illustrate the range of single-frequency point SNR across three typical scenarios. As shown in the figure, most SNR samples in the LoS (Line-of-Sight) scenario are concentrated between 14.2 and 21.6 dB, with a median value of 17.6 dB. Similarly, in the mixed scenario, most of the SNR values fall within the range of 14.3 dB to 19.6 dB, with a median of 17.2 dB. In contrast, the NLoS (Non-Line-of-Sight) scenario exhibits significantly lower SNR values, ranging from 9.4 dB to 16.9 dB, with a median of 12.3 dB. It is important to note that our processing pipeline can achieve approximately 15 dB gain through antenna beamforming. To futher display the measured channel property, we choose four UE positions (positions A-D, see Fig. 3 (b)) from the three measurement scenarios and show representative channel impulse responses (CIR) in Fig. 6 (a)-(d). To be specific, Fig. 6 (a) illustrates

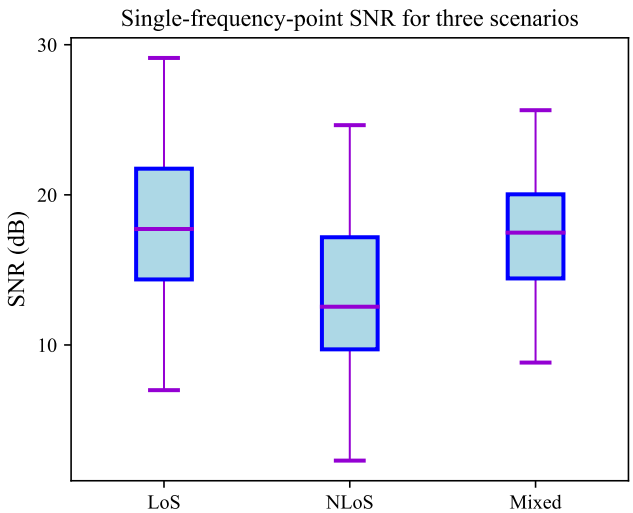


Fig. 5: Single-frequency-point SNR for three scenarios.

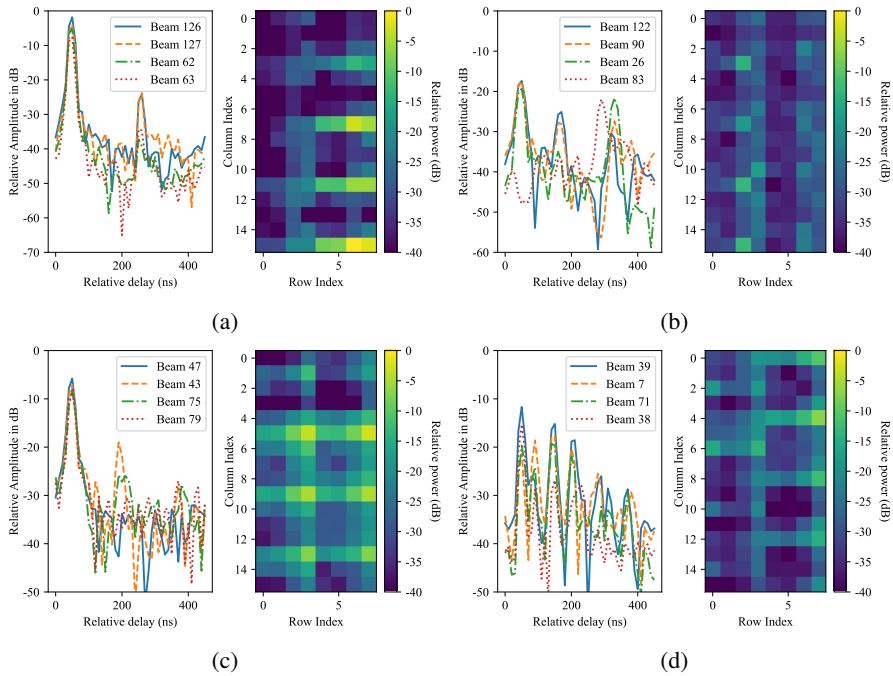


Fig. 6: CIR and relative power of all 128 beams of four locations (a) LoS at point A, (b) NLoS at point B, (c) LoS at point C, (d) NLoS at point D. Beam diagrams are arranged as follows: row $0-3$ and row $4-7$ represent the 32 vertical and 32 horizontal-polarized beams respectively for UE layer 1; row $8-11$ and row $12-15$ represent the co-polarized beams for UE layer 2. Beam index is $8*(i-1)+j$, where i and j denote the row and column index respectively. We select the first 4 strongest beam and plot the relative amplitude of CIR. The strongest beam among all figures (a)-(d) is normalized to 0-dB. The relative amplitude refers to the power difference of a specific beam to the strongest beam among all 4 figures.

a typical LoS scenario where a dominant LoS path can be seen from both the CIR and the beam patterns. Few beams exhibit dominant power levels, while others remain comparatively weaker. Although few NLoS-paths can still be observed, their strengths are much weaker compared to the direct path. This is because the UE is located in an open parking lot, where the reflected signals from other buildings are relatively weak. From the beam power pattern, one can observe the signal strength variations of different BS antenna polarizations and UE transmission layers as well. In contrast, Fig. 6 (b) displays NLoS channel characteristics where the BS captures

several reflected paths and there is no path with a dominant power. Thus, the signal strength in Fig. 6 (b) is lower compared to the case in Fig. 6 (a). Fig. 6 (c) and Fig. 6 (d) present the measured channels in a mixed scenario, where more local scatters surround the UE. The distance between UE position C and the BS is greater than that of UE position A, resulting in a decrease in the strength of the received LoS signal. Nevertheless, the BS is capable of detecting stronger reflective paths in addition to the LoS path, attributed to reflections from surrounding buildings. Similarly, in Fig. 6 (d), a rich number of multipath components can be observed in both the CIR and the beam pattern, despite the LoS path being obstructed.

We focus on simple deployments and mobility scenarios to showcase the novel approach, specifically targeting typical urban and rural environments, including LoS, NLoS, and mixed scenarios common in commercial networks. The dense, controlled test scenarios provide a robust evaluation of the proposed positioning algorithm. In contrast, larger network deployments would increase complexity and pose significant challenges in data generation, collection, and processing, which fall beyond the scope of this study.

V. Results and discussion

In this section, we evaluate our ML-based localization pipeline using the measurements. We initially compare the single-snapshot localization performance for different ML algorithms under different scenarios. Then, we demonstrate the performance gain achieved by smoothing multiple position estimates with a Kalman filter.

A. Single snapshot localization

Our approach starts with assessing the validity of the input channel snapshot, as outlined in Section. III. B. The first criterion, related to the CTF matrix $\boldsymbol{\Xi}$, employs a cut threshold set at 3500 out of 5888 (128×46) available physical resource elements, approximately 60%, so that the channel information is sufficient. With such threshold setting, signal paths can be clearly visualized from the channel impulse responses. After discarding snapshots with insufficient data, we generate the amplitude of impulse response beam matrix $|\mathbf{G}_t|$ and feed it to the attention-aided localization block. This block, with detailed parameters in Table. I, comprises three cascaded sub-blocks. Initially, positioning encoding is applied to $|\mathbf{G}_t|$ using (7). Subsequently, a layer normalization procedure follows according to (8). The

TABLE I: Overview of our ML-based single snapshot localization pipeline

Item	Network Structures or Parameters
Input Features	Amplitude of CIRs for all beams
Network Output	Estimated position labels or probabilities
Intermediate block 1	Residual 2-Heads Self-attention Network
Intermediate block 2	Residual Position-wise FCNNs
Intermediate block 3	3 cascaded ordinary FCNNs
Time Complexity	NF^2

normalized matrix is then input into a simple 2-head self-attention block with a single self-attention layer, generating matrix $\mathbf{Z}'$ via (2-6). The pairwise correlation values in matrix $\mathbf{A}$ reflect the similarities between each pair of row vectors in $\mathbf{H}_t$ in the beam domain, which provides valuable information for UE localization. Considering the simplicity of future hardware implementation work, the exact parameter settings are displayed as follows: $A_1 = 128, A_2 = 46, A_3 = 64$. After the Add & Norm operation, the output is transferred to the second sub-block, consisting of two FCNNs with sizes $\mathbf{W}_1 \in 46 \times 128$ and $\mathbf{W}_2 \in 128 \times 46$. Following this, the output matrix of the second sub-block is vectorized to yield a vector of length 5888. This vector is fed into the last FCNN sub-block, with sizes as given in Table II. Network sizes are shown in Table II, and they vary based on the selected cost functions. As seen in Table II, the neural network requires more resources when RbC is used as the cost function. This is because RbC needs to calculate the probability of each bin in the final layer, rather than simply estimating the 2-D position of the UE. Despite this, the size of all three networks remains under 10 MB, classifying them as lightweight neural networks. As illustrated in We compare the localization performance when using three different loss functions and in three typical scenarios. As illustrated, the output matrix of the second intermediate block is first vectorized and fed to the input layer of the third sub-block, which consists of 2-3 FCNNs depending on the choice of loss functions. When the loss function RbC is used, its corresponding network delivers the probability of all L_x and L_y bins. In scenario I, $L_x = L_y = 200$ while in the other two scenarios $L_x = L_y = 100$. The deviation vectors $\mathbf{d}_x$ and $\mathbf{d}_y$ are set as: $\mathbf{d}_x = \delta_x \mathbf{1}, \mathbf{d}_y = \delta_y \mathbf{1}$, where $\mathbf{1}$ denotes the all-ones vector, δ_x and δ_y denote the deviation value of the x- and y-axis, respectively. Accordingly, the output dimension

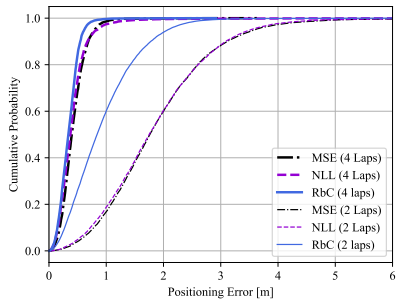
TABLE II: Structures and parameter settings of the third FCNN sub-block using three different loss functions. Lr: Learning rate.

Items \ Loss F.	MSE	NLL	RbC
Input layer size	5888×1	5888×1	5888×1
Hidden layer 1	5888×32	5888×32	5888×128
Hidden layer 2	32×2	32×2	$128 \times \tilde{L}$
Batch size	64	64	64
Lr: LoS (4 laps)	0.0006	0.0006	0.0006
Lr: NLoS (4 laps)	0.0006	0.0006	0.0006
Lr: Mixed (4 laps)	0.0006	0.0006	0.0006
Lr: LoS (2 laps)	0.0002	0.0002	0.0002
Lr: NLoS (2 laps)	0.0001	0.0001	0.0001
Lr: Mixed (2 laps)	0.0002	0.0002	0.0002
Learning Epoch	500	500	500
Dropout Rate	0.05	0.05	0.05
Cost function	(11)	(12)	(13)
Network Size	1.137 MB	1.138 MB	7.44 MB

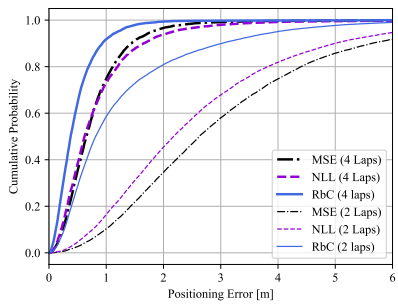
$\tilde{L}$ equals $L_x + L_y + 2$. The penalty term γ_1 and γ_2 are set as: $\gamma_1 = \gamma_2 = 1$. In addition, the Euclidean norm loss function is utilized, i.e. $\eta = 2$.

1) Comparisons of different uncertainty estimations

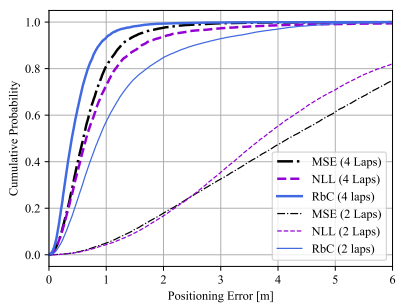
Fig. 7 compares the positioning accuracy of our single-snapshot localization pipeline using three loss functions in three scenarios under different training densities. As shown, the RbC method outperforms the other two methods in all three scenarios and under both high and low training densities. Compared to the other two methods, RbC learns better the non-Gaussian probability distribution of the UE position, while the performance of the NLL method is constrained by its underlying Gaussian assumption, and the MSE method does not estimate uncertainty. The performance of these three methods differ less in the LoS scenario and high training density, because the estimated UE position has less uncertainty in this situation. However, in other scenarios or lower training density, the uncertainty of the estimated UE position increases due to reduced SNR or training samples. Consequently, an accurate uncertainty estimation is more essential, and thus the RbC method performs much better. At both high and low training densities, our pipeline performs best in



(a)



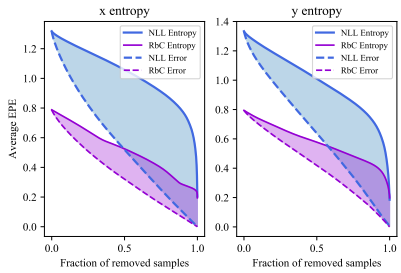
(b)



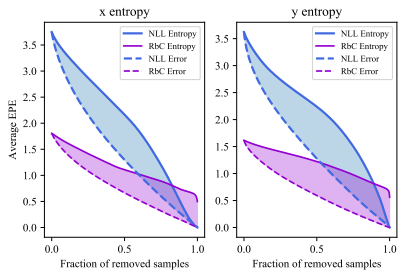
(c)

Fig. 7: Positioning errors of different training densities in the three scenarios: (a) LoS, (b) NLoS, (c) Mixed.

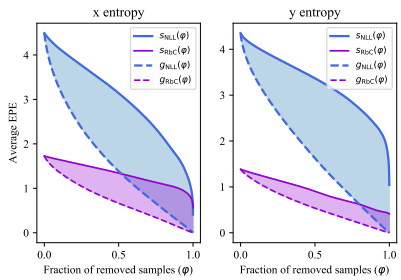
LoS scenarios, the mixed scenario ranks 2nd, while the localization performance in the NLoS scenario is the worst. We postulate that in the LoS scenario, the much



(a)



(b)



(c)

Fig. 8: Sparsification curves of NLL and RbC methods under high training density (4 laps as training data) and across three scenarios: (a) LoS, (b) NLoS, (c) Mixed. EPE: Endpoint error.

higher SNR contributes to very good positioning accuracy. To further compare the uncertainty estimation quality of the NLL and RbC methods, we demonstrate the sparsification and oracle curves of the probability density functions of the estimated

UE-x and y coordinates under high training density in Fig. 8. Specifically for the NLL method, we discretize the predicted Gaussian functions to achieve the same number of discrete bins as the RbC method, according to (16). The AUSE values for all training densities are calculated according to (17) and are displayed in Table III. To reduce the effect of outliers, the starting point of the sparsification and oracle curves equals 99% of the positioning error. As depicted in Fig. 8, the discrepancies between the sparsification (entropy) and oracle curves are significantly reduced in all three scenarios when the RbC method is used. This improvement is reflected in the improved AUSE values presented in Table III. These findings underscore the quality of the uncertainty estimation achieved with our approach.

We finally compare the NLL method to another popular approach, the Monte Carlo (MC) dropout method [37]. This technique estimates uncertainty by applying dropout to a trained neural network. During testing, the network is evaluated multiple times, with a percentage of neurons randomly deactivated on each run. This randomness results in slightly different predictions on each evaluation. The mean of these predictions provides the final estimate, while the variance among them represents

TABLE III: AUSE values of two uncertainty estimation algorithms under different training densities across three channel scenarios.

	NLL-x	RbC-x	NLL-y	RbC-y
LoS (4 laps)	0.480	0.179	0.351	0.163
NLoS (4 laps)	0.579	0.427	0.704	0.548
Mixed (4 laps)	1.428	0.616	1.543	0.325
LoS (2 laps)	1.951	0.968	2.023	1.181
NLoS (2 laps)	3.816	1.868	3.407	2.475
Mixed (2 laps)	4.682	0.809	3.540	1.138

TABLE IV: Negative-log-likelihood values of MC Dropout methods compared with NLL.

	LoS (4 laps)	NLoS (4 laps)	Mixed (4 laps)
NLL	-0.144	0.069	0.066
MC dropout	2.142	0.014	0.138

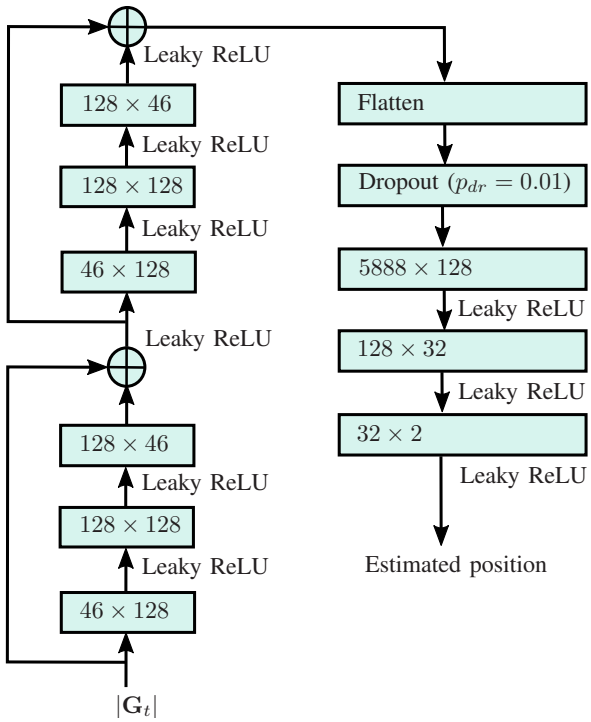
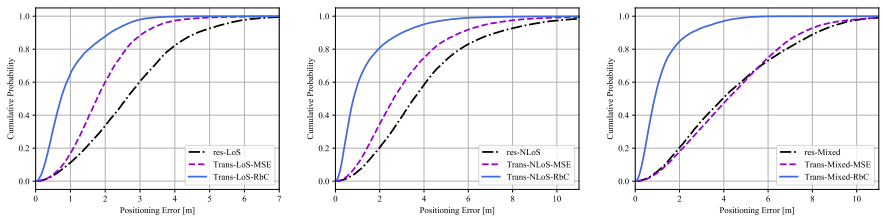


Fig. 9: Structure of the residue network.

the uncertainty. Using this approach, we evaluate our pipeline with MSE as the cost function. During testing, the dropout rate is set to 0.05, and for each input $\mathbf{G}_t$, the network is evaluated 50 times, after which we compute the mean predictions $\tilde{\mathbf{p}} = [\tilde{p}_{x_i}, \tilde{p}_{y_i}]^T$ and variances $\breve{\boldsymbol{\sigma}} i^2 = [\breve{\sigma}^2_{x_i}, \breve{\sigma}^2_{y_i}]^T$. Four laps are used for training, with the remainder allocated for testing. To assess performance, we calculate the negative log-likelihood (NLL) score of the MC dropout and NLL methods on the testing dataset, according to (12). The results are presented in Table IV. As shown, the MC dropout method performs similarly to the NLL method in both NLoS and Mixed scenarios. However, under the LoS scenario, the MC dropout method shows overconfidence, with a significantly lower estimated variance compared to the NLL method. We attribute this to the fact that MC dropout primarily captures uncertainty related to the network's weights, but it does not fully account for other sources of



(a) (b) (c)

Fig. 10: Comparison between positioning error of our pipeline and a residue neural network under three different scenarios. (a) LoS, (b) NLoS, (c) Mixed. For all three scenarios, two laps are used for training and the rest for testing.

uncertainty, such as model misspecification or uncertainty in the underlying data distribution.

2) Compare with the start-of-the-art

We compare performances of our approach with a residue neural network (ResNet), which is widely used in solving regression and classification tasks. The structure and parameter settings of the neural network are illustrated in Fig. 9. As seen in Fig. 9, the residue network consists of two residue blocks, followed by three fully connected layers. The time complexity of this model is $O(N \max(F^2, FN, N^2))$. Dropout is also applied to avoid overfitting and Leaky ReLU (with negative slope -0.3) is selected as the activation function. The learning rate is set as follows: LoS 0.00001, NLoS 0.00005 and Mixed 0.00005. Fig. 10 illustrates the localization errors of attention-aided and ResNet based pipeline under all three scenarios. Channel data of two laps are selected as training and the rest three laps are used for testing purposes. Compared with the residue network, our transformer-based approach performs better under both LoS and NLoS scenarios, if MSE is selected as the loss function. Localization accuracy can be further significantly improved, if we use RbC approach to estimate uncertainty. We postulate that compared with the state-of-the-art, our processing pipeline benefits both from the attention mechanism and the advanced uncertainty estimation algorithms.

B. Smoothing the trajectory by Kalman filtering

Next, we investigate the performance when using a Kalman filter for smoothing within our pipeline. To clearly visualize the effect of the Kalman filter, we apply a low training density, using two laps for training and one lap for testing. First, the

TABLE V: Parameter settings and rooted mean square errors (RMSE) when applying the Kalman Filtering

	ϵ_1	ϵ_2	RMSE (m), before filter	RMSE (m)
LoS	0.05	1.2	0.99	0.93
NLoS	0.05	1.2	2.00	1.76
Mixed	0.05	1.2	1.01	0.82

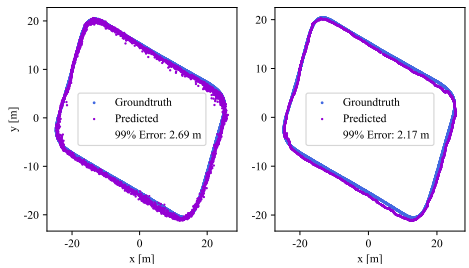
validity of each channel CSI is assessed by the data cleaning block. All test channel samples classified as valid are then utilized for evaluation. Similarly to Section V.B, we apply an attention-aided block as the backbone and the output layer utilizes the RbC uncertainty estimation. For simplicity, the matrix $\mathbf{\Lambda}$ in (20) and the matrix $\mathbf{R}$ in (23) are set as

$$\mathbf{\Lambda} = \epsilon_1^2 \mathbf{I}, \ \mathbf{R} = \epsilon_2^2 \mathbf{I}, \tag{25}$$

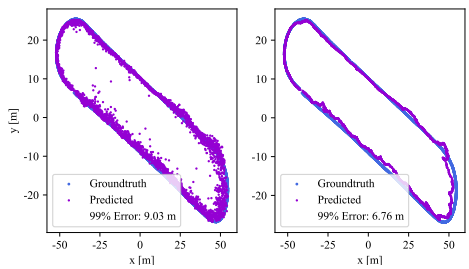
where ϵ_1 and ϵ_2 denote the standard deviation, which indicates the state and observation noise levels, respectively. Their exact values for the three scenarios are listed in Table V. Fig. 8 shows the predicted UE trajectories both with (right) and without (left) the Kalman filter for the three scenarios. The MSE between the predicted trajectories and their ground truths is shown in Table. V. As expected, the results demonstrate a significant improvement with the inclusion of the Kalman filter: the trajectories become considerably smoother, and outliers are mitigated to a large extent. Consequently, there is a substantial enhancement in localization accuracy, particularly evident in NLoS and mixed propagation scenarios. This improvement can be attributed to the ability of the Kalman filter to utilize relationships between different snapshots, which effectively balances the newly predicted UE position with previous positional states, leading to more accurate localization.

VI. Conclusions

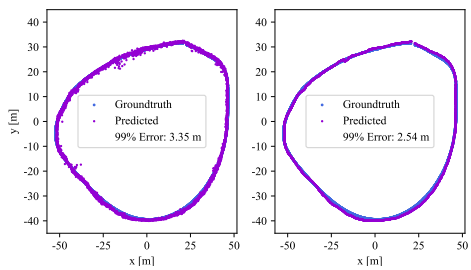
In this paper, machine learning is applied to a 5G NR cellular system for UE localization. A novel ML-based localization pipeline is presented, which utilizes attention-aided techniques to estimate UE positions by employing impulse response beam matrices as channel fingerprints. In addition, we implement two uncertainty estimation techniques, namely the NLL and RbC methods, to estimate the probability density function of the UE position error and compare their performances. Finally,



(a)



(b)



(c)

Fig. 11: Comparison between the raw (the left) and Kalman-filtered trajectory (a) LoS, (b) NLoS, (c) Mixed.

a Kalman filter is applied to smooth consecutive position estimates. To evaluate our pipeline, an outdoor cellular 5G measurement campaign was conducted at 3.85 GHz with a 100 MHz bandwidth, covering both LoS and NLoS scenarios, achieving submeter-level localization accuracy. The measurement results indicate several key

findings: 1) The attention-aided block shows promising potential to deliver high-precision localization accuracy. 2) The RbC uncertainty method outperforms the traditional NLL method, particularly with low training density or in more complex channel propagation scenarios. This advantage likely stems from the fact that the RbC method is not constrained by a Gaussian assumption on position errors. 3) Applying a Kalman filter to smooth consecutive position estimates significantly reduces position outliers, thereby enhancing localization accuracy. In future work, we plan to increase the diversity of our training data and expand the evaluation scenarios by testing our approach in various urban environments. Additionally, we will explore combining model-based and data-driven methods to further enhance the generalizability and robustness of our approach.

APPENDIX: LIST OF ABBREVATIONS

Abbreviation	Definition
2-D	Two-Dimensional
3-D	Three-Dimensional
3GPP	3rd Generation Partnership Project
5G	Fifth Generation
AoA	Angle of Arrival
AUSE	Area Under the Sparsification Error
BS	Base Station
CIR	Channel Impulse Response
CSI	Channel State Information
CTF	Channel Transfer Function
DL	Downlink
EPE	Endpoint Error
FCNN	Fully Connected Neural Network
GNSS	Global Navigation Satellite Systems
KNN	K-Nearest Neighbors
Lr	Learning Rate
LoS	Line-of-Sight
MC	Monte Carlo
ML	Machine Learning
MSE	Mean Square Error
NLL	Negative Log Likelihood
NLoS	None Line-of-Sight
NR	New Radio
OFDM	Orthogonal Frequency Division Multiplexing
PRB	Physical Resource Block
PRSG	Physical Resource Blocks Sub Groups
RbC	Regression-by-Classification
Rel	Release
RMSE	Rooted Mean Square Error
RX	Receivers
SNR	Signal-to-noise Ratio
SRS	Sounding Reference Signal
ToA	Time of Arrival
TDD	Time-division Duplexing
TDoA	Time Difference of Arrival
Tx	Transmitters
UE	User Equipment
UL	Uplink

ACKNOWLEDGMENT

The authors thank PhD candidate Ziliang Xiong for valuable suggestions regarding the uncertainty prediction.

REFERENCES

[1] X. Cai, X. Cheng, and F. Tufvesson, "Toward 6G with Terahertz Communications: Understanding the Propagation Channels," *IEEE Communications Magazine.*, vol. 62, no. 2, pp. 32–38, Feb. 2024, doi: 10.1109/MCOM.001.2200386.

[2] R. Whiton, "Cellular Localization for Autonomous Driving: A Function Pull Approach to Safety-Critical Wireless Localization," *IEEE Veh. Technol. Mag.*, vol. 17, no. 4, pp. 28–37, Dec. 2022, doi: 10.1109/mvt.2022.3208392.

[3] A. A. Abdallah, C. S. Jao, Z. M. Kassas, and A. M. Shkel, "A Pedestrian Indoor Navigation System Using Deep-Learning-Aided Cellular Signals and ZUPT-Aided Foot-Mounted IMUs," *IEEE Sensors Journal.*, vol. 22, no. 6, pp. 5188–5198, Mar. 2022, doi: 10.1109/jsen.2021.3118695.

[4] M. Maaref and Z. M. Kassas, "Ground Vehicle Navigation in GNSS-Challenged Environments Using Signals of Opportunity and a Closed-Loop Map-Matching Approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 2723–2738, Jul. 2020, doi: 10.1109/tits.2019.2907851.

[5] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Intelligent reflecting surface enhanced indoor robot path planning: A Radio Map-based approach,"*IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4732–4747, Jul. 2021, doi:10.1109/twc.2021.3062089.

[6] R. Whiton, J. Chen, T. Johansson and F. Tufvesson, "Urban Navigation with LTE using a Large Antenna Array and Machine Learning," 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), Helsinki, Finland, 2022, pp. 1-5.

[7] J. A.del P Rosado, R. Raulefs, J. A. López-Salcedo, and G. Seco-Granados, "Survey of Cellular Mobile Radio Localization Methods: From 1G to 5G," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 2, pp. 1124–1148, 2018, doi: 10.1109/COMST.2017.2785181.

[8] A. Grenier, E. S. Lohan, A. Ometov, and J. Nurmi, "A Survey on Low-Power GNSS," *IEEE Commun. Surv. Tutor.*, vol. 25, no. 3, pp. 1482–1509, Jan. 2023, doi: 10.1109/comst.2023.3265841.

[9] A. Hu, T. Lv, H. Gao, Z. Zhang, and S. Yang, "An ESPRIT-Based Approach for 2-D Localization of Incoherently Distributed Sources in Massive MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 996–1011, Oct. 2014, doi: 10.1109/JSTSP.2014.2313409.

[10] X. Zeng, F. Zhang, B. Wang, and K. J. R. Liu, "Massive MIMO for High-Accuracy Target Localization and Tracking," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10131–10145, Jun. 2021, doi: 10.1109/jiot.2021.3050720.

[11] X. Li, E. Leitinger, M. Oskarsson, K. Åström, and F. Tufvesson, "Massive MIMO-Based Localization and Mapping Exploiting Phase Information of Multipath Components," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4254–4267, Jun. 2019, doi: 10.1109/twc.2019.2922264.

[12] L. Lian, A. Liu, and V. K. N. Lau, "User Location Tracking in Massive MIMO Systems via Dynamic Variational Bayesian Inference," *IEEE Trans. Signal Process.*, vol. 67, no. 21, pp. 5628–5642, Nov. 2019, doi: 10.1109/tsp.2019.2943226.

[13] N. Garcia, H. Wymeersch, E. G. Larsson, A. M. Haimovich, and M. Coulon, "Direct Localization for Massive MIMO," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2475–2487, May 2017, doi: 10.1109/tsp.2017.2666779.

[14] X. Cai, W. Fan, X. Yin, and G. F. Pedersen, "Trajectory-Aided Maximum-Likelihood Algorithm for Channel Parameter Estimation in Ultrawideband Large-Scale Arrays," *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 10, pp. 7131–7143, Oct. 2020, doi: 10.1109/tap.2020.2996774.

[15] X. Cai and W. Fan, "A Complexity-Efficient High Resolution Propagation Parameter Estimation Algorithm for Ultra-Wideband Large-Scale Uniform Circular Array," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5862–5874, Aug. 2019, doi: 10.1109/tcomm.2019.2916700.

[16] R. Surya, E. Hossain, and V. K. Bhargava, "Machine Learning Methods for RSS-Based User Positioning in Distributed Massive MIMO," *IEEE*

Trans. Wireless Commun., vol. 17, no. 12, pp. 8402–8417, Dec. 2018, doi: 10.1109/twc.2018.2876832.

[17] X. Sun, C. Wu, X. Gao, and G. Y. Li, "Fingerprint-Based Localization for Massive MIMO-OFDM System With Deep Convolutional Neural Networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 10846–10857, Nov. 2019, doi: 10.1109/tvt.2019.2939209.

[18] X. Guo and N. Ansari, "Localization by Fusing a Group of Fingerprints via Multiple Antennas in Indoor Environment," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 9904–9915, Nov. 2017, doi: 10.1109/tvt.2017.2731874.

[19] G. Tian, I. Yaman, M. Sandra, X. Cai, L. Liu, and Fredrik Tufvesson, "Deep-learning based high-precision localization with massive MIMO," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 2, pp. 19–33, Jan. 2024, doi: 10.1109/tmlcn.2023.3334712.

[20] J. Vieira, E. Leitinger, M. Sarajlic, X. Li and F. Tufvesson, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, 2017.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)., Red Hook, NY, USA.

[22] E. Gonultas, E. Lei, J. Langerman, H. Huang, and C. Studer, "CSI-Based Multi-Antenna and Multi-Point Indoor Positioning Using Probability Fusion," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2162–2176, Apr. 2022, doi: 10.1109/twc.2021.3109789.

[23] Sibren De Bast, E. Vinogradov, and S. Pollin, "Expert-Knowledge-Based Data-Driven Approach for Distributed Localization in Cell-Free Massive MIMO Networks," *IEEE Access*, vol. 10, pp. 56427–56439, Jan. 2022, doi: 10.1109/access.2022.3177837.

[24] R. Bharadwaj, A. Alomainy, and S. K. Koul, “Experimental Investigation of Body-Centric Indoor Localization Using Compact Wearable Antennas and Machine Learning Algorithms,” *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 2, pp. 1344–1354, Feb. 2022, doi: 10.1109/tap.2021.3111308.

[25] R. Surya, E. Hossain, and V. K. Bhargava, “Low-Dimensionality of Noise-Free RSS and Its Application in Distributed Massive MIMO,” *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 486–489, Aug. 2018, doi: 10.1109/lwc.2017.2787764.

[26] A. Salihu, S. Schwarz and M. Rupp, ”Attention Aided CSI Wireless Localization,” 2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC), Oulu, Finland, 2022.

[27] M. Stahlke, T. Feigl, S. Kram, B. M. Eskofier and C. Mutschler, ”Uncertainty-based Fingerprinting Model Selection for Radio Localization,” 2023 13th International Conference on Indoor Positioning and Indoor Navigation (IPIN), Nuremberg, Germany, 2023.

[28] A. Ráth, D. Pjanić, B. Bernhardsson and F. Tufvesson, ”ML-Enabled Outdoor User Positioning in 5G NR Systems via Uplink SRS Channel Estimates,” IEEE International Conference on Communications (ICC), Rome, Italy, 2023.

[29] R. Whiton, J. Chen, and F. Tufvesson, “Wiometrics: Comparative Performance of Artificial Neural Networks for Wireless Navigation,” *IEEE Trans. Veh. Technol.*, pp. 1–16, Jan. 2024, doi: 10.1109/tvt.2024.3396286.

[30] S. F. Bhat, I. Alhashim and P. Wonka, ”AdaBins: Depth Estimation Using Adaptive Bins,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021.

[31] Z. Xiong, A. Jonnarth, A. Eldesokey, J. Johnander, B. Wandt, and P.-E. Forssen, “Hinge-Wasserstein: Estimating multimodal aleatoric uncertainty in regression tasks,” in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop, to appear, 2024.

[32] S. K. Lind, Z. Xiong, P. Forssén, V. Krüger, Uncertainty quantification metrics for deep regression, *Pattern Recognition Letters*, 2024, doi:10.1016/j.patrec.2024.09.011.

[33] E. Ilg, et al., “Uncertainty estimates and multi-hypotheses networks for optical flow,” in 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 2018.

[34] C. L. Canonne, G. Kamath, and T. Steinke, “The discrete Gaussian for differential privacy,” in Proceedings of the 34th International Conference on Neural Information Processing Systems (Neurips), Vancouver, Canada, 2020.

[35] S. M. Kay, Fundamentals of statistical signal processing: estimation theory. USA: Prentice-Hall, Inc., 1993.

[36] 3GPP, “Base station (BS) radio transmission and reception,” 3rd Generation Partnership Project, Technical Specification (TS) 38.104, 2023, version 15.19.0.

[37] Y. Gal and Z. Ghahramani. ”Dropout as a Bayesian approximation: representing model uncertainty in deep learning,” in Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, USA, 2016.

Paper VI

Paper VI

Reproduced, with permission from IEEE

D. Pjanić, G. Tian, A. Reial, X. Cai, B. Bernhardsson, F. Tufvesson, "Illuminating the Path: Attention-Assisted Beamforming and Predictive Insights in 5G NR Systems," *IEEE Transactions on Vehicular Technology*, submitted May 2025.

Illuminating the Path: Attention-Assisted Beamforming and Predictive Insights in 5G NR Systems

Dino Pjanić, *Member, IEEE*, Guoda Tian, *Student, IEEE*,
Andres Reial, *Senior Member, IEEE*, Xuesong Cai, *Senior Member, IEEE*,
Bo Bernhardsson, and Fredrik Tufvesson, *Fellow, IEEE*

Abstract

Artificial intelligence advances have recently influenced wireless communications, including beam management in fifth-generation (5G) new radio systems. AI-driven models and algorithms are being applied to enhance tasks such as beam selection, prediction, and refinement by leveraging real-time and historical data. These approaches address challenges such as mobility under complex channel conditions, showing promising results compared to traditional methods. Beam management in 5G refers to processes that ensure optimal alignment between the base station and user equipment for effective signal transmission and reception based on real-time channel state information and user positioning. This study leverages accurate beam prediction to identify a smaller subset of beams, resulting in a more efficient, streamlined, and link-adaptive communication system. The innovative

Dino Pjanić, is with Ericsson AB, Lund, Sweden (e-mail: dino.pjanic@ericsson.com) and the Department of Electrical and Information Technology, Lund University, Lund, Sweden. (e-mail: dino.pjanic@eit.lth.se) (*Corresponding author: Dino Pjanić*)

Guoda Tian and Andres Reial are with Ericsson AB, Lund, Sweden. (e-mail: guoda.tian, andres.reial@ericsson.com)

X. Cai is with the School of Electronics, Peking University, Beijing, 100871, China (email: xuesong.cai@pku.edu.cn) and the Department of Electrical and Information Technology, Lund University, Lund, Sweden (email: xuesong.cai@eit.lth.se).

Bo Bernhardsson is with the Department of Automatic Control, Lund University, Lund, Sweden. (e-mail: bo.bernhardsson@control.lth.se)

Fredrik Tufvesson is with the Department of Electrical and Information Technology, Lund University, Lund, Sweden. (e-mail: fredrik.tufvesson@eit.lth.se)

Dino Pjanić and Guoda Tian contributed equally to this work.

The work is partially sponsored by the Swedish Foundation for Strategic Research and Ericsson AB, Sweden.

approach presented introduces a precise, attention-based prediction model that derives the entire downlink transmission chain in a commercial grade 5G system. The predicted downlink beams are specifically tailored to handle the complexities of none line-of-sight environments known for high-dimensional channel dynamics and scatterer-induced signal variations. This novel method introduces a paradigm shift in utilizing environmental and channel dynamics in contrast to conventional procedures of beam management, which entails complex methods involving exhaustive techniques to predict the best beams. The presented beam prediction results demonstrate robustness in addressing the challenges posed by signal-dispersive environments, showcasing great potential in mobility scenarios.

Index Terms

Beam Management, Beam Prediction, Beamforming Weights, 5G New Radio, Self-Attention, Sounding Reference Signal.

I. Introduction

Beamforming is a signal processing method that directs radio energy through the channel toward a targeted receiver. Massive multiple-input multiple-output (MIMO) is an advanced antenna technology that provides high flexibility in beamforming due to the many radio frequency chains it employs [1]. By adjusting the phases and amplitudes, the system can create constructive interference in the desired area and destructive interference in others. This approach enables focused beams toward specific receivers, enhancing signal strength and providing greater spatial diversity and multiple data streams. To sustain effective beamforming, especially with moving users, the system requires precise channel state information (CSI), which reflects the current characteristics of the communication channel between the transmitter and receiver. Beamforming performs best in line-of-sight (LoS) scenarios, where there is an unobstructed path between the transmitter and receiver. In contrast, None LoS (NLoS) scenarios, such as urban environments with numerous obstructions, present significant challenges. In these cases, as users or obstacles move, signals often reflect off surfaces, resulting in multipath propagation, which demands advanced algorithms and precise CSI processing to dynamically match the instantaneous multipath propagation. This makes it particularly difficult to maintain accurate real-time CSI estimates for multiple beams, especially in scenarios involving high-mobility users such as vehicles. In NLoS environments with rapid changes or where CSI is noisy or incomplete, a reduced beam set allows the system to focus on the most reliable beams or those contributing the most signal energy, rather than trying to support numerous weak or scattered beams. Prediction of the strongest beams and beam reduction are closely interrelated, as both help to improve 5G beam management (BM) [2]. They are recognized as resource optimization strategies in MIMO systems, aiming to minimize the number of active beams or spatial streams employed during transmission or reception.

Recently, advances in machine learning (ML) and artificial intelligence (AI), particularly deep neural networks (DNNs) such as transformer models introduced [3], have emerged as powerful tools to tackle a wide range of tasks. Originating in natural language processing (NLP), transformers utilize a unique mechanism known as self-attention, enabling them to capture long-term dependencies more effectively than traditional recurrent neural networks (RNNs). This makes transformers especially suitable for analyzing long sequences in time series data, such as CSI measurements influenced by UE mobility patterns and surrounding scattering characteristics in wireless environments. In this paper we study beam prediction in 5G NR systems based on attention models and channel fingerprints.

The rest of the article is organized as follows: the next section outlines the motivation for this study and provides an overview of key ML/AI applications in BM within 5G communication systems. Subsequently, we propose a transformer architecture with an attention-based model tailored for BM, with a focus on downlink beam predictions. Finally, we evaluate the performance of the proposed model with a particular emphasis on long-term beam prediction in demanding NLoS environments.

In wireless systems, coherence time refers to the period during which the channel impulse response or the transfer function, both with respect to phase and amplitude, remain relatively stable. Hence, CSI acquisition must be processed at millisecond level to track channel dynamics during mobility under varying environmental conditions. In legacy BM techniques, primarily employed in millimeter-wave (mmWave) communications, the base station (BS) transmits reference signals (RSs) and configures UEs to measure and report them. These measurements, along with associated reporting, impose significant overhead, a challenge that becomes particularly pronounced in dynamic and dispersive NLoS scenarios. Predictive methods, such as AI-assisted beam predictions, present a promising solution to reduce the reliance on continuous RS transmissions and measurements, addressing the limitations of traditional parametric models and solutions to meet the required capacity and performance improvements [4]. Furthermore, conventional mathematical approaches to beam management often rely on idealized assumptions, such as pure additive white Gaussian noise, which may not accurately reflect real-world conditions [5], creating opportunities for AI/ML models to capture and model complex nonlinear factors effectively. Advances in AI and ML have introduced a transformative perspective to 5G NR standardization [6] [7], particularly through Release-18 [8], which explores AI/ML-driven approaches to address scalability issues of MIMO systems such as increased antenna array sizes. The authors of [9] and [11] provide an overview of current standards in relation to AI / related to AI / ML techniques. Many of the suggested methods replace traditional sequential beam sweep with predictive algorithms operating in the temporal and spatial domains; detailed insights are provided in [12] - [14]. Key use cases, such as CSI feedback, beam management, and positioning, capitalize on the channel's temporal stability within the coherence time to reduce complexity. Recent research demonstrates the feasibility of short-term beam predictions using variations in the angles of arrival (AoA) and departure (AoD) in mobile environments. Hybrid approaches [15] - [19] leverage prior low-frequency

channel information to predict optimal mmWave beams, reducing training overhead. In general, these approaches illustrate how the use of spatial channel characteristics in the sub-6 GHz band can simplify the complexity of the mmWave beam prediction encountered at mmWave frequencies. Different cross-domain approaches are also proposed in [20], using LidarDAR sensors to improve time-domain beam prediction, while the authors of [21] employ multipoint radar sensing to enhance beam tracking.

However, most recent BM studies rely on simplified LoS-dominant scenarios, often using simulated data as input of AI/ML models. In LoS environments, the best beam generally remains stable, suggesting that beam predictions could feasibly extend beyond coherence time if the UE follows a predictable movement trajectory. Another significant aspect is the fact that short-term prediction faces limitations: coherence regions typically span only a few decimeters or centimeters, depending on the frequency band, restricting prediction to brief intervals. We foresee advantages in longer-term beam predictions, where fingerprinting approaches can be explored by leveraging spatial and historical data patterns that indirectly incorporate AoA and AoD information through high-dimensional features. While fingerprinting is well-suited for stable and predictable radio channel environments, it requires robust augmentation with adaptive algorithms to handle the complexities of NLoS and dynamic scenarios effectively. However, in dispersive, NLoS environments even when the UE's trajectory is approximately known, the optimal beam prediction becomes highly sensitive to precise UE locations, often down to fractions of a wavelength.

Since the UE position inherently correlates with the best beam, this spatial information becomes a key factor in our prediction. The proposed attention-based solution adaptively learns the features of the measured UL channel and captures the features of the propagation characteristics, such as UE locations and surrounding environmental structures. In subsequent sections, we compare the energy efficiency based on the beam predictions of the proposed attention-based model. This involves evaluating the total energy of the full beam array against the energy represented by the predicted subset, providing insights into how efficiently the subset captures the beam energy relative to the entire array.

A. *Contributions*

- We present an accurate attention-based model for beam prediction that utilizes UL channel estimates from a commercial 5G system to derive the entire DL transmission chain, specifically tailored to handle the complexities of NLoS environments.
- The novel approach introduced here enables accurate beam prediction far beyond the coherence time by utilizing high-dimensional fingerprinted features. These features predict temporal changes in the AoA and AoD, offering a more robust solution for dynamic and complex wireless environments.

II. System Model and Problem Formulation

In our single-user massive MIMO setup, the BS uses orthogonal frequency division multiplexing (OFDM) with F subcarriers. At time t, the UE transmits an SRS pilot

signal using M_{UE} antennas. The BS has M_{BS} antennas, evenly split between vertical and horizontal polarization. Furthermore, let P denote the number of multipath components, $\tau_{p,t}$ represent the time delay for the p-th path between the UE and BS, and $\alpha_{p,m,t}$ indicate the complex coefficient of the p-th multipath component at time t. UE transmits a pilot signal that reaches the BS antenna array at an azimuth arrival angle ϕ_p and an elevation angle θ_p for the p-th multipath component, respectively. All vertically polarized antennas are used to form M_{bm}^{Vt} beams, with the response of the i-th beam given by $\beta_{V,i}(\phi_p, \theta_p, f)$, where f represents the pilot signal frequency. Similarly, horizontally polarized antennas generate M_{bm}^{Hz} beams, with the response of the i-th beam defined as $\beta_{H,i}(\phi_p, \theta_p, f)$. The total number of beams is $M_{bm} = M_{bm}^{Vt} + M_{bm}^{Hz}$. Consequently, for the m-th UE antenna the propagation channel model at time t for each beam is:

$$
\begin{aligned}
h_{\mathbf{V},i,m,t}(f) &= \sum_{p=1}^{P} \beta_{V,i}(\phi_p, \theta_p, f)\, \alpha_{p,m,t}\, e^{-j(2\pi f \tau_{p,t})} \\
h_{\mathbf{H},i,m,t}(f) &= \sum_{p=1}^{P} \beta_{H,i}(\phi_p, \theta_p, f)\, \alpha_{p,m,t}\, e^{-j(2\pi f \tau_{p,t})}.
\end{aligned}
\tag{1}
$$

By aggregating $h_{\mathbf{V},i,m,t}(f)$ and $h_{\mathbf{H},i,m,t}(f)$ from the F subcarriers, we construct two channel transfer functions (CTF), $\mathbf{H}_{\mathbf{V},m,t} \in \mathbb{C}^{M_{bm}^{Vt} \times F}$ and $\mathbf{H}_{\mathbf{H},m,t} \in \mathbb{C}^{M_{bm}^{Hz} \times F}$ representing the vertical and horizontal polarized antennas, respectively, at time t. These matrices are strongly influenced by the UE's position, making them effective raw channel fingerprints for predicting DL beamforming weights, as the beam direction explicitly depends on the UE's location and implicitly on the CSI. Finally, for all UE antennas, the combined channel matrice is $\mathbf{H}_{UL,t} \in \mathbb{C}^{N \times F} = \left[\mathbf{H}_{\mathrm{H},1,t}^T, \mathbf{H}_{\mathrm{V},1,t}^T, ..., \mathbf{H}_{\mathrm{H},M_{\mathrm{UE}},t}^T, \mathbf{H}_{\mathrm{V},M_{\mathrm{UE}},t}^T\right]^T$, where $N = M_{UE} M_{bm}$. In the subsequent stage of the signal processing chain in a time division duplex (TDD) system, the beamforming weights are determined using minimum mean square error (MMSE) channel estimator algorithm (2). These algorithms are designed to optimize the signal strength in the desired direction while minimizing interference. In a system with M_{BS} antennas, the DL CTF can be estimated using the complex conjugate of the UL channel matrix, exploiting the reciprocity principle inherent in TDD systems, $\hat{\mathbf{H}}_{DL} = \mathbf{H}_{UL}^*$.

$$
\hat{\mathbf{H}}_{DL} = (\mathbf{H}_{UL}^H \mathbf{H}_{UL} \ \ + \sigma^2 \mathbf{I})^{-1} \mathbf{H}_{UL}^H. \tag{2}
$$

where $\mathbf{H}_{UL}^H$ denotes conjugate transpose of the channel matrix, σ^2 noise variance, and $\mathbf{I}$ identity matrix.

The MMSE calculation detailed in (2) is computationally intensive. In practice, only a small subset of beamforming weights in the matrix $\mathbf{H}_{DL}$ constitute beams that account for the majority of the beamforming energy. Predicting the strongest beams for future time instances can help reduce both energy consumption and computational load. However, it is challenging to accurately predict $\mathbf{H}_{UL}$ and eventually $\mathbf{H}_{DL}$ for future time instances,

especially when the time gap exceeds the channel coherence time. This is because the further the prediction is, the more difficult it becomes to model the channel dynamics with high precision. This drives the need to explore non-traditional approaches, such as AI-based algorithms, to predict the strongest beams in advance. The proposed attention-based model learns a functional relationship $\boldsymbol{\psi} = f(\mathbf{H}_{UL,t}, \Delta_t)$, where $\boldsymbol{\psi} \in \mathbb{R}^n$ represents the n strongest beams in the predicted matrix $\hat{\mathbf{H}}_{DL}$, and Δ_t denotes the time difference between current and future snapshots.

III. Methodology

CSI acquisition in time TDD systems can benefit from channel reciprocity, meaning that the uplink (UL) and downlink (DL) channels are related since both use the same frequency band. This allows for estimating the CSI on the UL and applying it to the DL beamforming. For example, a BS equipped with 64 transceivers can leverage a single uplink pilot to estimate the full 64-dimensional channel across the entire bandwidth. This is facilitated by the transmission of sounding reference signals (SRS) from the multiantenna UE, enabling the BS to estimate the DL channel and compute DL beamforming weights (BFWs). These weights, represented as complex coefficients, are applied to the BS's MIMO antenna elements to control signal direction and shape by adjusting the amplitude and phase of the DL signal.

A. Dataset collection

To enable TDD-based downlink beam prediction using UL SRS channel estimates, a commercial grade 5G BS was used, compliant with 3GPP standards [22] - [29] for radio resource management, physical layer considerations and beam measurement procedures. The BS operated at a center frequency of 3.85 GHz with a bandwidth of 100 MHz and was equipped with a rooftop-mounted phased array antenna module (PAAM) comprising 64 cross-polarized antenna elements. A 5G-capable UE was kept connected while simultaneously downloading data at a sustained rate of 750 Mbit/s to maintain continuous UL SRS transmission throughout the measurements. The baseband unit of the BS processed the time-varying SRS reports to extract channel estimates, operating initially in the antenna element domain. The SRS data received, represented as complex samples, underwent additional unpacking from 16-bit floating-point format to 2xSQ15. The pre-processing steps included undoing the normalization of the SRS channel estimates averaged over the physical resource blocks (PRB) pairs to derive estimates for each PRB and beam. Finally, the processed SRS samples were transformed into the beamspace domain using a fast Fourier transform (FFT), producing a time series of beam measurements.

The measurement campaign encompassed two distinct propagation scenarios, LoS and NLoS, as illustrated in Fig. 1, as these provided a range of challenging environments to evaluate the proposed approach. Data collection was performed on these two approximately rectangular routes. Each route was repeated over five laps, serving as baselines for analysis. All measurements were performed using a test vehicle moving at a steady velocity of

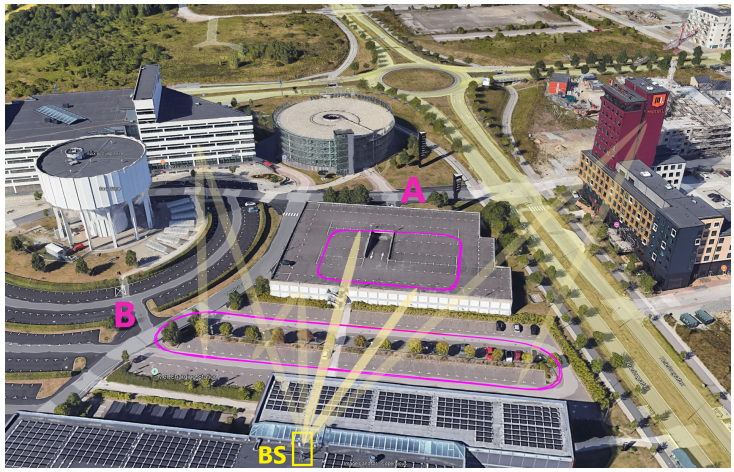


Fig. 1. A BS placed at a 20 m high rooftop. SU-MIMO scenario was tested in the two pre-defined measurement routes: **A**: The roof of a 10 m high garage building was used for LoS propagation measurements with a strong dominant path along the entire route. **B**: A ground-level route where the signals reflect off multiple surfaces and surrounding buildings block the signal causing NLoS propagation.

15 km/h (4.2 m/s). The approach of comparing two fundamentally opposite propagation scenarios serves to establish a simplified baseline with the LoS scenario while contrasting it with the more complex NLoS scenario. The NLoS scenario introduces a non-trivial relationship between the UE trajectory and the scatterers, collectively influencing the optimal beam direction. The prediction method relies on repetitive UE movement paths, emulating a car traveling along a road or a robot/machine following a specific route in a factory setting. These scenarios provide consistent movement patterns that can be effectively captured during model training and utilized during inference, allowing the model to represent the dominant candidate beam paths. However, it is important to note that the optimal beam may still vary between different movement realizations during inference. However, in practice, most commercial site deployments exhibit predictable UE movement patterns due to the static nature of the surrounding environment, which further supports the feasibility of this approach.

B. Signal Processing Framework

This section outlines the BS processing of the UL SRS channel estimates, which form the primary training data set. The BS handles a time series of SRS measurements, representing the angular delay spectrum of the radio channel in the beam domain. Simultaneously, a parallel computation determines the corresponding DL BFWs using MMSE-

based algorithms, as described in Section II. In this study, the UL SRS data serve as the input for the Attention-aided model, while the prediction task focuses on the generated DL beams, as detailed in the following chapters. As illustrated in Fig. 2, the UL SRS channel estimates span 273 PRBs within a 100 MHz bandwidth. Each channel snapshot includes data for all 64 beams across these PRBs, based on an SRS reporting interval of 20 ms. To reduce complexity, the PRBs are grouped into adjacent pairs, with the average value of each pair calculated by downsampling. This process results in 137 PRB subgroups (PRSGs). Further refinement is achieved by grouping every three consecutive PRSGs, where the first PRSG is downsampled, and the second and third are interleaved. This approach ultimately yields 46 PRSGs. The UE, equipped with four antennas (corresponding to four UE layers), is responsible for transmitting the SRS pilot signals. The SRS pilots recorded from all four UE layers form CTF matrices $\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3, \mathbf{H}_4 \in \mathbb{C}^{N \times F}$. The matrix $\mathbf{H}' \in \mathbb{C}^{4N \times F}$ contains all four matrices, specifically $\mathbf{H}' = [\mathbf{H}_1^T, \mathbf{H}_2^T, \mathbf{H}_3^T, \mathbf{H}_4^T]^T$. After further processing, the final 20-ms-based CTF snapshot is structured with 4 x [4 x 46] dimensions, representing amplitude instances. This data is collected for the 64 BS antennas and four UE layers in 46 PRSGs, as illustrated in Fig. 2.

Obtaining uplink UL SRS channel measurements in a commercial 5G BS introduces significant challenges, particularly when working with large, complex data structures like SRS measurement samples. These measurements, generated at millisecond intervals, are typically confined to the baseband unit of the BS for internal operations, with external access often limited by hardware and software restrictions. Moreover, since not all PRSG values are updated during UL SRS transmissions, it becomes essential to account for and address missing channel estimate values. To ensure the reliability of the collected UL CTF, the processing pipeline must include mechanisms to verify the validity of input data and handle incomplete PRSGs effectively. SRS channel estimates are classified as valid

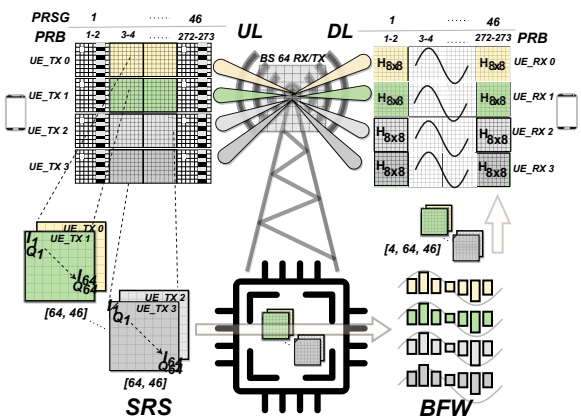


Fig. 2. Beamforming weights generation based on SRS channel estimates.

only if all PRSGs and UE transmit antennas have successfully transmitted SRS during a given discontinuous reception (DRX) cycle. Each PRSG comprises two PRBs, so channel estimates are derived by averaging the PRB pairs for every PRSG and beam. Validation of the raw UL CTF matrix involves a two-step process. The matrix is deemed invalid if it meets any of the following criteria based on the number of missing subcarrier samples (represented as zero elements):

- Insufficient CSI snapshots in the beam and frequency domain: the number of non-zero elements in $\mathbf{H}_t$ is lower than a given threshold of 60%.
- Updating procedure stalling: The values at all sub-carriers or all beams remain the same compared to the previous reporting interval.

Note: In cases where the UL CSI was determined to be insufficient, the calculation of DL BFWs was halted.
After discarding all invalid data, the next step is to process the raw UL CTF to generate impulse response beam matrices. To suppress the side lobes, we apply Hann windowing in all rows of the matrix $\mathbf{H}_t$ to obtain matrix $\hat{\mathbf{H}}_t \in \mathbb{C}^{N \times F}$. The F-length Hann window in the frequency domain is given by:

$$w[f] = \sin^2\left(\frac{\pi f}{F}\right), \quad f = 0, \ldots, F-1. \tag{3}$$

After the windowing operation, the impulse response beam matrix $\mathbf{G}_t$ is produced by performing the inverse discrete Fourier transform along each row of $\hat{\mathbf{H}}_t$. Given the potential difficulty in achieving a stable phase for $\mathbf{G}_t$, here we opt to use its amplitude $|\mathbf{G}_t|$ as the training feature, although this means discarding potentially useful information.

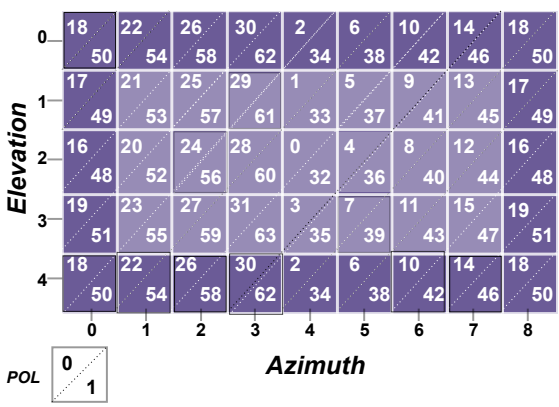


Fig. 3. Grid of Beams.

Digital beamforming involves the adjustment of antenna weights during the digital baseband processing. The independently calculated DL BFWs are applied to the downlink signal after transforming them from the beam domain to the antenna domain. In this process, each antenna in the array is assigned a specific phase and amplitude as determined by the computed weights. The PAAM explained above generates a grid-of-beams (GoB), forming a structured set of beams that span the coverage area. Each beam represents a spatially focused transmission or reception pattern, enabling efficient signal delivery to or from specific UE locations. As illustrated in Fig. 3, the GoB structure yields 64 distinct beams, providing precise control over the signal direction and maximizing spatial selectivity.

IV. Transformer Architecture

The transformer architecture, introduced in [3], has served as the foundation for numerous state-of-the-art models in natural language processing, thanks to its ability to effectively process input sequences and generate accurate output sequences. Unlike traditional methods that analyze tokens sequentially, transformers relate each token to all others within a sequence. Their self-attention mechanism replaces sequential processing with parallel computation, distinguishing them from recurrent neural networks (RNNs) [30] and convolutional neural networks (CNNs) [31]. Using parallelism, transformers efficiently capture long-range contexts and dependencies across distant positions in input or output sequences.

The transformer is a deep neural network (DNN) model [32] composed of multiple layers with a uniform architecture. These layers are organized into stacks that differ from those in classical DNN models. Each stack, which can function as an encoder or a decoder, operates from bottom to top. The input and output sequences are transformed into vectors

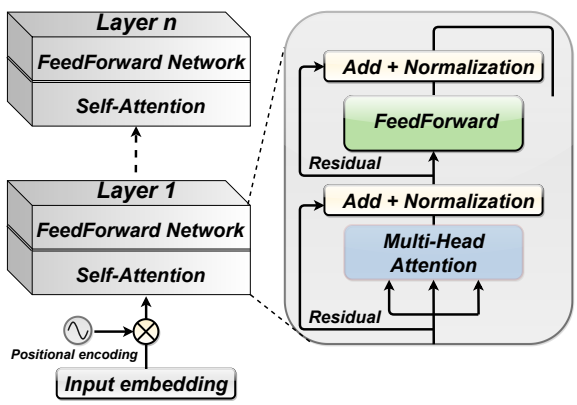


Fig. 4. A stacked architecture of computational encoder layers. The proposed model deploys 3-layer architecture.

of dimension *d* through embedding and positional encoding layers. Each layer sequentially passes its learned representations to the next layer until the final prediction is achieved. In particular, each layer comprises sublayers, all of which share an identical structure across different layers, enhancing hardware optimization. In its original design, the transformer includes two key sublayers: a self-attention sublayer and a feedforward network, as depicted in Fig. 4. The self-attention sublayer is further divided into *n* independent and identical components, called heads. The transformer architecture was originally designed for sequence-to-sequence tasks such as machine translation, and both encoder and decoder blocks were soon adapted as standalone models. Although there are hundreds of different transformer models, most of them belong to one of three types; encoder-only, decoder-only or encoder-decoder. In this study, we chose the encoder-only architecture to predict the best beams.

A. Input Embedding

The input embedding sublayer converts the input tokens to vectors of dimension d_{model}. Many embedding methods can be applied to the tokenized input. The later proposed model applies a simple lookup table that stores embeddings of a fixed dictionary and size. This module often stores word embeddings and retrieves them using indices. The input of the module is a list of indices, and the output is the corresponding word embeddings.

B. Positional Encoding

The idea behind positional encoding (PE) is to preserve sequential information in the input data sequence. This is achieved by adding value to the input embedding instead of having additional vectors to describe the position of a token in a sequence. Positional embedding provides sine and cosine functions that generate different frequencies for the PE for each entry i of the d_{model} entries in the PE vector:

$$\begin{aligned} \mathbf{PE}(pos\ 2i) &= \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \\ \mathbf{PE}(pos\ 2i+1) &= \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right). \end{aligned} \tag{4}$$

The sine function is applied to the even numbers and the cosine function to the odd numbers. These vectors follow a specific pattern that the model learns, which helps it determine the position of each token or the distance between different tokens in the sequence. In addition, positioning ensures meaningful distances between the embedding vectors once they are projected via dot-product operations in the attention mechanism.

C. Encoder Stack

As shown in Fig. 4, the encoder consists of a stack of n layers, each comprising two primary sublayers: a multihead self-attention block and a position-wise fully connected feed-forward network. To facilitate deeper models, a residual connection is applied around

each of these sublayers, followed by layer normalization. Specifically, each sublayer, denoted as `sublayer(x)`, includes a residual connection that transports the raw input x of the sublayer directly to the normalization function of the layer. This ensures that critical information, such as positional encoding, is preserved throughout processing. The output dimensions of all sublayers, as well as embeddings, are of size d_{model} which has a significant consequence, for example, all key operations are dot products. As a result, the dimensions remain stable, which reduces the number of operations.

D. Self-Attention

At the core of Transformers lies the self-attention mechanism. The input of a Transformer consists of a sequence of contiguous tokens, each represented as a vector in an embedding matrix. As part of the self-attention process, three projection matrices $\mathrm{W_q}$, $\mathrm{W_k}$, and $\mathrm{W_v}$ transform each input embedding vector into three distinct vectors: the Query, Key, and Value. For each token, its corresponding Key vector is compared to the Query vectors of all other tokens by computing dot products. This calculation provides a measure of similarity between Queries and Keys, forming the foundation of the attention mechanism. A Softmax function is then applied to normalize these similarity scores, amplifying the most relevant relationships. The softmax function is defined as below (5)

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{\mathbf{d_k}}}\right)\mathbf{V}, \tag{5}$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are representation matrices. In addition, $\boldsymbol{d}_{\mathrm{k}}$ represents the dimensionality of the Key vectors and $\sqrt{\boldsymbol{d_k}}$ is a scaling factor to prevent large values in the dot product. The softmax expression on the right-hand side of equation (5) normalizes the similarity scores into probabilities. The resulting matrix, known as the self-attention matrix, captures the contextual relationships between the tokens. This process is performed multiple times in parallel, using several independent sets of projections, resulting in a multi-head attention layer that enhances the model's ability to capture complex patterns across the input sequence.

V. Proposed Transformer Framework

The implemented Transformer model deploys a 3-layer, encoder-only architecture to predict the time series of DL beams based on UL SRS channel estimates. The paired data sets consist of the input UL channel impulse response (CIR) data, fed into the encoder, and the corresponding DL beam TF serving as the target dataset, described in detail in Section IV. The model is trained by minimizing the error between the encoder's output and the target DL beams. The input data, represented as the UL CIR beam matrix $\mathbf{G}_t \in \mathbb{C}^{N\times F}$, provides amplitude values used as input of a DNN equipped with an attention mechanism, while $\mathbf{B}_t \in \mathbb{C}^{N\times F}$ represents the DL TF. The attention-based model processes the input sequence to produce a rich numerical representation optimized for translating UL SRS channel estimation data into DL beam predictions. The architecture leverages

bidirectional attention, where the representation of a given token considers both preceding and succeeding tokens in the sequence. This property is particularly well-suited for time-series data, as it captures the temporal dependencies in channel measurement sequences effectively. The attention-based beam prediction pipeline, as described in Fig. 5, consists of an encoder-only deployment that comprises multiple attention-aided blocks, followed by an output layer that employs a loss functions, namely the Mean Square Error (MSE). We use $\boldsymbol{\eta}_i = [\eta_{bm_1,i}, ..., \eta_{bm_{64},i}]^T$ to represent the CTF of the DL beam as the ground truth of the moving UE at position i. This approach directly estimates the 64 DL beams by setting a regression head in the output layer of the last attention block. Let $f_{\text{MSE}}(.)$ denote the overall function and vector $\boldsymbol{\theta}_2$ all hyperparameters, $\boldsymbol{\eta}_i = [\eta_{bm_1,i}, ..., \eta_{bm_{64},i}]^T$ the estimated i-th 64-sized beam set generated by $f_{\text{MSE}}(\boldsymbol{\theta}_2, |\mathbf{G}_t|)$, the loss $\boldsymbol{\ell}$ can be expressed as

$$\boldsymbol{\ell} = \frac{1}{N_{train}} \sum_{i \in \Omega'_{train}} ||\boldsymbol{\eta}_i - \hat{\boldsymbol{\eta}}_i||_F^2, \tag{6}$$

where Ω'_{train} and N_{train} denote the training set and the number of training samples, respectively, and $||.||_F$ denotes the Frobenius matrix norm.

As illustrated in Fig. 5, an attention head operates in two key steps. First, as detailed in Section IV, the attention mechanism computes the keys **K** and queries **Q** from the input data. This process evaluates the relevance of each query vector **Q** with respect to all key vectors **K**, generating an energy score that reflects their importance. Or, simply explained

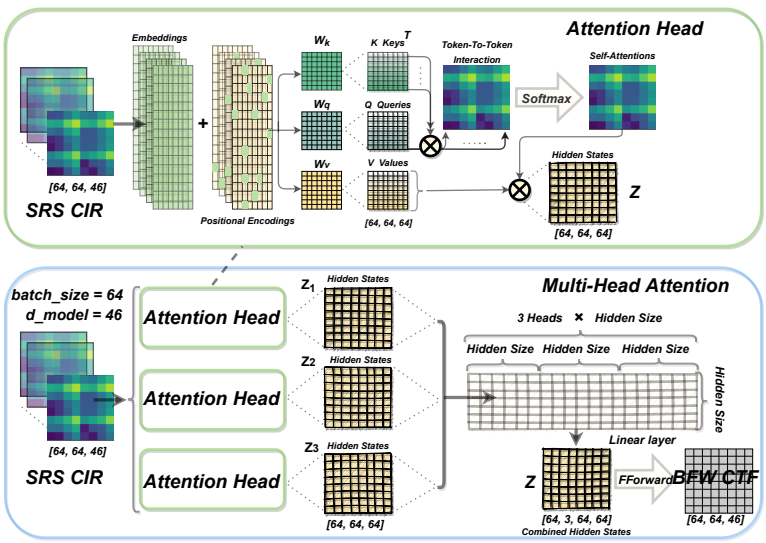


Fig. 5. Attention model implementation and tensor dimensions.

how much attention should a token pay to another token in the input sequence. A Softmax transform normalizes and further accentuates the high similarities, and the resulting matrix is called self-attention.
Next, the mechanism introduces a separate feature representation vector called values **V** that is combined with the attention weights (calculated from the dot products of **Q** and keys **K**) to produce the so-called hidden states. These hidden states represent a weighted sum of the values, highlighting the most relevant information. Notably, **Q**, **K** and **V** are all derived from the same input sequence, ensuring that the attention mechanism captures a rich and comprehensive representation of the data. The process is repeated multiple times with multiple attention layers, resulting in a multi-head attention layer. The final hidden states are combined into final hidden states by using a linear layer.

VI. Results and Discussion

This section evaluates the AI-based prediction pipeline. We start by assessing the data cleaning process to ensure the retention of sufficient channel information. Next, we assess the beam energy by quantifying the predicted DL beamforming performance. This involves analyzing the energy distribution within the selected beam subsets across different prediction time horizons, providing information on the effectiveness of the prediction model in maintaining energy efficiency and accuracy over time.

A. Single snapshot channel represenatation

As detailed in Section III-A, the BS recorded channel snapshots for two propagation scenarios, LoS and NLoS, resulting in $\mathcal{T}1 = 22000$ and $\mathcal{T}2 = 24603$ snapshots, representing time instances, collected on a 20 ms time resolution basis. These snapshots are structured into two tensors, $\mathcal{A}\text{LoS} \in \mathbb{C}^{\mathcal{T}1\times N\times F}$ and $\mathcal{A}\text{NLoS} \in \mathbb{C}^{\mathcal{T}2\times N\times F}$, where each subset of four adjacent snapshots corresponds to the signals transmitted by four UE antennas (layers). Each tensor is normalized by multiplying with a scalar such that its Euclidean norm is equal

TABLE I. Architectural overview of the proposed encoder-only model.

Model entity	Network Structures or Parameters
Input Features	Amplitude of CIRs for all beams
Network Output	Estimated DL TF for all beams
Intermediate block 1	Residual 3-head Self-Attention Layer
Intermediate block 2	Residual Position-wise FCNNs
Intermediate block 3	3 cascaded ordinary FCNNs
Encoder layers	3
$\mathrm{d}_{\text{model}}$	46
Batch size	64
Optimizer	Adam
Learning Epochs	2500
Time Complexity	NF^2

TABLE II. Overview of the last FCNN sub-block.

Model entity	Dimensions
Input layer size	2944×1
Hidden layer 1	2944×64
Hidden layer 2	64×32
Hidden layer 3	64×64
Cost function	(6)

to $\mathcal{T}_i MN$, where $i = 1, 2$. Following the validation of the input channel snapshot described in Section III-B, a single UL CTF matrix instance $\boldsymbol{\Xi}$ instance $\boldsymbol{\nu}$ applies a cut threshold of approximately 60%, yielding 1766 of the 2944 (64×46) components, corresponding to 64 BS antennas in 46 PRSG. This ensures that adequate channel information is retained. The amplitude of the UL CIR beam matrix, $|\mathbf{G}_t|$, is then derived and passed to the attention-aided prediction block. The prediction block architecture, described in Table I, comprises three cascaded sub-blocks, providing the foundation for downstream processing. Initially, position encoding is applied to $|\mathbf{G}_t|$ as described in (4), followed by layer normalization. The normalized matrix is then fed into three parallel self-attention blocks, each comprising a single self-attention layer, as illustrated in Fig. 5, to produce the output matrix $\mathbf{Z}$ via (5). These multi-head attention layers process a sequence of size $\mathcal{T}$ (single snapshot), with each head projecting the feature dimensions 1766 into smaller subspaces to compute the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ representations. After the Add & Normalization process, as depicted in Fig. 4, the output is transferred to the second sub-block, consisting of two position-wise fully connected neural networks (FCNNs) with sizes $\mathbf{W}_1 \in 46 \times 64$ and $\mathbf{W}_2 \in 64 \times 46$. Following this, the output matrix of the second sub-block is vectorized to produce a vector of original length 2944. This vector is fed into the last FCNN sub-block, with sizes defined in Table II. This entire computational process is repeated across all three layers of the encoder block, adhering to the architecture depicted in Fig. 4.

As detailed in Section V, the model uses a paired dataset, where the input UL CIR data is processed by the encoder and the corresponding DL beam TF, $|\mathbf{B}_t|$, serves as the ground truth. The model training optimizes, via Adam optimizer, the encoder's output by minimizing the error relative to the target DL beam TF. The model's final predicted output represents the beam energy levels across all 46 PRSG subcarriers and 64 antennas of the BS. It encapsulates the total energy distribution across the frequency and spatial domains and reflects the combined radiated energy output of the beamforming system, integrating contributions from each antenna and subcarrier.

B. Beam Energy Evaluation Methodology

The cumulative power across the predicted beam subset is the main subject of investigation to predict the total power expected in a series of predicted beams. In this study, instead of ranking the beams by their indices, we focus on selecting the strongest beams within

each subset, meaning the indices of the predicted strongest beams may not always align with those in the ground truth dataset. Ranked by their expected power (e.g., strongest to weakest), the cumulative power is the sum of the power contributions of individual beams in a beam subset is defined as

$$P_{\text{cumulative}}(n) = \sum_{i=1}^{n} P_i, \tag{7}$$

where $P_{cumulative}(n)$ is the cumulative power of the first n predicted beams, and P_i is the

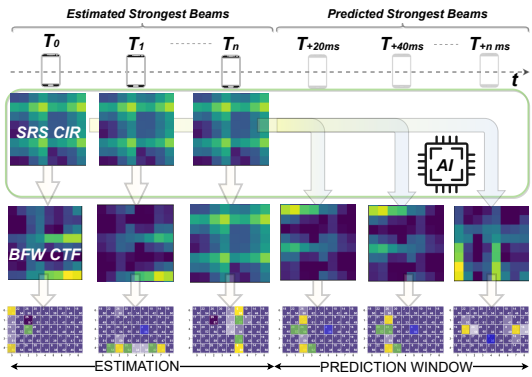


Fig. 6. Subsets of n strongest beams are selected for future transmission from the predicted beam-domain set.

power of the i-th beam. This cumulative metric can help quantify how well the predictions capture the total power available in the full set of beams, providing insights into how quickly the cumulative power saturates with increasing beams and helping determine the optimal subset size for energy-efficient operation. When comparing the energy of the complete set with a subset, $\mathcal{N}$, to a subset n, where $n \in \mathcal{N}$, discrete subsets of beams are selected with sizes n = [4,8,16,32]. Here, $\mathcal{N}$ represents the total energy from all beams within the antenna system, serving as a benchmark for evaluating the energy output. This approach enables a systematic assessment of the trade-offs between beam subset size and energy efficiency while maintaining the predictive accuracy of the system. The energy of the best beams predicted by the suggested attention-based model, with its concept shown in Fig 6, is associated with the location of the UE. Determining the significance of the subset relative to the whole is valuable for beam prediction, which is highly related to the mobility of UE and scatterers.

In commercial deployments, threshold levels can be used to define the minimum beam subset size corresponding to the optimal energy levels, allowing the subset size to adapt dynamically rather than being restricted to the predefined discrete sizes considered in this

study.

Statistical measures reduce extensive data sets to a single value, offering only one perspective on model errors by emphasizing specific aspects of model performance. To evaluate the performance of the proposed attention-assisted model, which uses time series data as input, we selected percentage-based error metrics such as the mean absolute percentage error (MAPE) (8) and the weighted mean absolute percentage error (WMAPE) (9). MAPE quantifies the average magnitude of the error, while WMAPE, a variant of MAPE, adjusts the error calculations by incorporating real values or weights.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\ y_i - \hat{y}\ |}{|y_i|} \times 100\% \tag{8}$$

$$\text{WMAPE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i|} \times 100\% \tag{9}$$

C. Evaluation of Estimation Results

While the primary focus of this study is to enhance BM by predicting sets of best beams, we also underscore a significant capability of the proposed model: the accuracy of DL transfer functions estimation via BFWs generated by the MMSE algorithm. As described earlier, BFWs are applied to the base station's MIMO antenna elements to shape and steer the signal by adjusting its amplitude and phase in the DL. Fig. 7 illustrates the high accuracy of the attention-based model in the LoS (right) and NLoS (left) scenarios. This result establishes a solid foundation for precise beam prediction in the DL. Furthermore, it highlights an important capability to generate BFWs, traditionally computed within the base station's baseband hardware, a process known for its intensive computational and hardware resource demands. These requirements are particularly challenging in multi-user MIMO scenarios, where the computational complexity grows linearly with the number of antennas at both the base station and user equipment. The NLoS results underscore the importance of focusing on a beam set, given the significant reduction in total energy caused by the

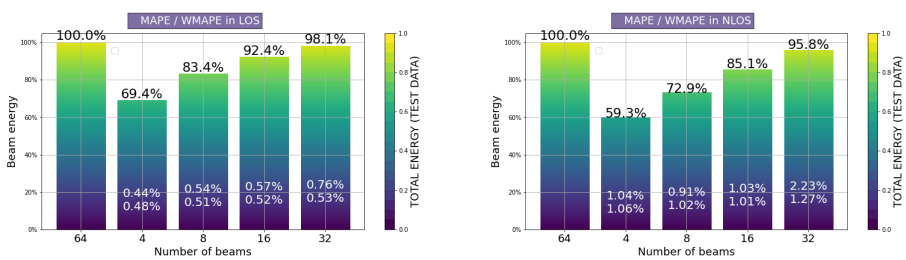
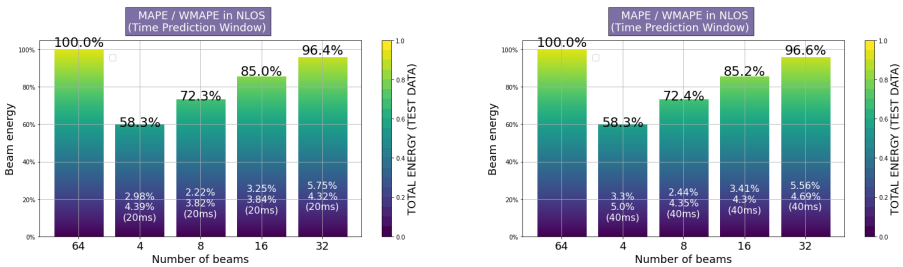


Fig. 7. LoS/NLoS comparison: Beam energy contributions of the estimated subsets relative to the total test data.

dispersive characteristics of NLoS propagation. Moreover, it provides valuable information to establish suitable energy threshold levels to guide beam prediction strategies effectively.

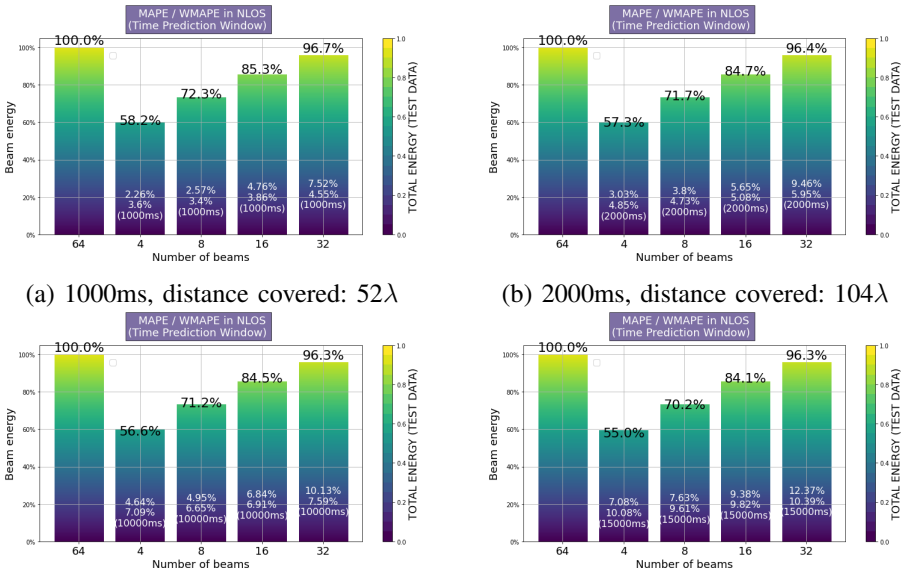
D. Evaluation of Prediction Results

The attention-based model enables long-term prediction of best beam candidate subsets for UEs following predictable movement paths, leveraging historical SRS data transmitted by the UE in the UL. These predictions extend over timelines spanning multiple seconds, equivalent to several hundred wavelengths. This allows for simplifying the BM procedure for future time instances by allocating BM CSI-RS resources and configuring UE measurements solely for the beam subset identified by the prediction algorithm. As described in Section III-A, the NLoS scenario involves numerous scatterers with varying locations, sizes, and shapes, creating a high-dimensional feature space. These scatterers collectively influence the optimal beam direction, making it challenging to accurately model the complex scattering characteristics. However, the proposed model exhibits strong prediction accuracy, particularly for larger beam subsets, such as n = 16, 32, ranked according to their expected power (e.g., strongest to weakest), the power in the weakest beams is expected to be very low compared to the cases of n = 4, 8, as Fig. 8 and Fig. 9 illustrate. The latter observation is particularly emphasized in Fig. 9 (d), where the attention model exhibits reduced prediction accuracy for larger beam subsets, such as 16 and 32 compared to Fig. 9 (a) for instance. This outcome is expected, given the long-term perspective that extends well beyond the coherence time window, corresponding to hundreds of multiples of λ-wavelengths. However, since these findings enable beam prediction well in advance, they have the potential to fundamentally transform the way 5G networks manage BM resources.



(a) 20ms, distance covered: 1.04λ (b) 40ms, distance covered: 2.08λ

Fig. 8. NLoS comparison: Beam energy contributions of the predicted subsets within short prediction windows



(a) 1000ms, distance covered: 52λ

(b) 2000ms, distance covered: 104λ

(c) 10000ms, distance covered: 521λ

(d) 15000ms, distance covered: 781λ

Fig. 9. NLoS comparison: Beam energy contributions of the predicted subsets within long prediction windows

VII. Conclusions

We explored the application of transformers in massive MIMO beam prediction and demonstrated their competitive performance using datasets derived from commercial 5G systems. The proposed approach incorporates spatial and environmental factors, such as multipath scattering and predictable movement patterns, enabling the model to maintain high prediction accuracy even when temporal correlations diminish. This represents a significant advancement in 5G beam management, showcasing accurate beam prediction beyond coherence time, thereby overcoming the limitations of traditional methods constrained by coherence-time boundaries. The number of CSI-RS resources allocated for each BM measurement and the CSI-RS reporting rate can be significantly reduced compared to legacy BM operations. This reduction may enhance the DL spectral efficiency by preserving resources for data transmission and/or mitigating interference in the DL. In addition, it helps to meet the processing demands of the UE and improves the energy efficiency of the UE. We acknowledge the repetitive nature of the selected UE trajectory and the UE antenna orientation toward the BS, as the UE was firmly mounted on the roof of the test vehicle. Although these factors reduced radio channel variations, the feasibility of attention-aided models for long-term beam prediction is still demonstrated. Another important considera-

tion is the computational complexity of transformer models, which may pose a limitation. Thorough evaluation of their real-time performance is essential, and using fast GPUs or specialized hardware can help minimize delays, ensuring efficient operation in commercial deployments.

REFERENCES

[1] E. G. Larsson, O. Edfors, F. Tufvesson, T. L. Marzetta, "Massive MIMO for next generation wireless systems," IEEE Communications Magazine, Vol. 52, No. 2, pp. 186-195, *https://arxiv.org/abs/1304.6690*, Feb. 2014.

[2] 3GPP TR 38.802, "Study on new radio access technology Physical layer aspects," *Third Generation Partnership Project (3GPP)*, 2017.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need," *31st Conference on Neural Information Processing System*, June 2017.

[4] Y. Wang, Z. Gao, D. Zheng, S. Chen, D. Gunduz, V. Poor, "Transformer-Empowered 6G Intelligent Networks: From Massive MIMO Processing to Semantic Communication," *IEEE Wireless Communications*, vol. 30, No. 6, pp. 98-104, Dec. 2023.

[5] K. Ma, Z. Wang, W. Tian, S. Chen, L. Hanzo, "Deep Learning for mmWave Beam-Management: State-of-the-Art, Opportunities and Challenges," *IEEE Wireless Communications*, vol. 30, no. 4, Aug. 2022.

[6] 3GPP TR 38.908, "Study on Artificial Intelligence/Machine Learning (AI/ ML) management," *Third Generation Partnership Project (3GPP)*, 2024.

[7] 3GPP TR 22.874, "5G System (5GS); Study on traffic characteristics and performance requirements for AI/ML model transfer," *Third Generation Partnership Project (3GPP)*, 2021.

[8] 3GPP TR 38.843, "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface," *Third Generation Partnership Project (3GPP)*, 2024.

[9] X. Lin, "An Overview of 5G Advanced Evolution in 3GPP Release 18," *IEEE Communications Standards Magazine*, vol. 6, No. 3, pp. 77-83, Oct. 2022.

[10] Q. Xue, J. Guo, B. Zhou, Y. Xu, Z. Li, S. Ma, "AI/ML for Beam Management in 5G-Advanced: A Standardization Perspective," *IEEE Vehicular Technology Magazine*, vol. 19, No. 4, pp. 64-72, Aug. 2024.

[11] W. Chen, J. Montojo, J. Lee, M. Shafi, Y. Kim, "The Standardization of 5G-Advanced in 3GPP," *IEEE Communications Magazine*, vol. 60, No. 11, pp. 98-104, Jun. 2022.

[12] J. Xu, I. Nakamura, R. Feng, L. Liu, and L. Chen, "Performance Evaluation of AI/ML Model to Enhance Beam Management in 5G-Advanced System," *IEEE Fourteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, pp. 1–6., Dec. 2023.

[13] Q. Xue, J. Guo, B. Zhou, Y. Xu, Z. Li, S. Ma, "AI-Based Beam Management in 3GPP: Optimizing Data Collection Time Window for Temporal Beam Prediction," *IEEE Open Journal of Vehicular Technology*, vol. 5, No. 4, pp. 48-55, Nov. 2023.

[14] J. Zuo, J. Zhang, Y. Cao, X. Chen, F. Wang, N. Hu, and X. Xu, "Artificial Intelligence-Based Spatial Domain Beam Prediction for 5G Beyond," *IEEE Globecom Workshops*, pp. 1460–1465, Nov. 2023.

[15] M. Alrabeiah, A. Alkhateeb, "Deep Learning for mmWave Beam and Blockage Prediction Using Sub-6 GHz Channels," *IEEE Transactions on Communications*, vol. 68, No. 9, pp. 5504-5518, Sept. 2020.

[16] K. Ma, D. He, H. Sun, Z. Wang, "Deep Learning Assisted mmWave Beam Prediction with Prior Low-frequency Information," *IEEE International Conference on Communications*, June 2021.

[17] M. Hashemi, C. E. Koksal and N. B. Shroff, "Out-of-band millimeter wave beamforming and communications to achieve low latency and high energy efficiency in 5G systems," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 875-888, Feb. 2018.

[18] Z. Ali, A. Duel-Hallen, H. Hallen, “Early Warning of mmWave Signal Blockage and AoA Transition Using sub-6 GHz Observations,” *IEEE Communications Letters*, vol. 8, pp. 207 - 211, Nov. 2019.

[19] M. S. Sim, Y. Lim, S. H. Park, L. Dai and C. Chae, “Deep learning-based mmWave beam selection for 5G NR/6G with Sub-6 GHz channel information: algorithms and prototype validation,” *IEEE Accesss*, vol. 8, pp. 51634-51646, 2020.

[20] D. Marasinghe, N. Jayaweera, N. Rajatheva, S. Hakola, T. Koskela, O. Tervo, J. Karjalainen, E. Tiirola, and J. Hulkkonen, “LiDAR aided Wireless Networks - Beam Prediction for 5G,” *IEEE 96th Vehicular Technology Conference*, Oct. 2022.

[21] K. Chen, D. Liu, Z. Zhang, “Beam Tracking Based on Multi-Point Radar Sensing and DNN Prediction,” *8th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*, Jan. 2024.

[22] M. Enescu, K. Jayasinghe, K. Ranta-Aho, K. Schober, and A. Toskala, “5G Physical Layer,” John Wiley & Sons, 2020, ch. 6, pp. 87–148.

[23] 3GPP TR 38.104, “Base Station (BS) radio transmission and reception,” Third Generation Partnership Project (3GPP), Dec. 2020.

[24] 3GPP TR 38.300 Rel. 16, “5G NR – Overall description Stage-2, 9.2.4 Measurements,” *Third Generation Partnership Project (3GPP)*, Dec. 2021.

[25] 3GPP TR 38.211, “NR; Physical channels and modulation,” *Third Generation Partnership Project (3GPP)*, 2022.

[26] 3GPP TR 38.215, “NR; Physical layer measurements,” *Third Generation Partnership Project (3GPP)*, 2023.

[27] 3GPP TR 38.321, “NR; Medium Access Control (MAC) protocol specification,” *Third Generation Partnership Project (3GPP)*, 2023.

[28] 3GPP TR 38.331, “5G NR – Radio Resource Control; Protocol specification,” *Third Generation Partnership Project (3GPP)*, 2021.

[29] 3GPP TR 38.901, “Study on Channel Model for Frequencies from 0.5 to 100 GHz,” *Third Generation Partnership Project (3GPP)*, 2022.

[30] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” Phys. D, Nonlinear Phenomena, vol. 404, Art. no. 132306., Mar. 2020.

[31] I. Goodfellow, Y. Bengio, and Aaron Courville, “Deep Learning,” Phys. D, Nonlinear Phenomena, John Wiley & Sons, Nov. 2016.

[32] Y. LeCun, Y. Bengio, G. Hinton, “Deep Learning,” *Nature*, vol. 521, pp. 436 - 444, May 2015

Printed by Tryckeriet i E-huset, Lund 2025

LUND
UNIVERSITY

Doctoral Dissertation
No. 184
ISBN 978-91-8104-485-0 (printed)
ISBN 978-91-8104-486-7-0 (digital)
ISSN 1654-790X184