IDENTIFICATION OF INDOOR RADIO ENVIRONMENT PROPERTIES BASED ON CHANNEL STATE INFORMATION USING MACHINE LEARNING APPROACHES

Teodora Kocevska

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Asst. Prof. Dr. Andrej Hrovat, Jožef Stefan Institute, Ljubljana, Slovenia
Co-Supervisor: Dr. Aleksandra Rashkovska Koceva, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Dr. Aleš Švigelj, Chair, Jožef Stefan Institute, Ljubljana, Slovenia
Asst. Prof. Dr. Bernard Ženko, Member, Jožef Stefan Institute, Ljubljana, Slovenia, and Faculty of Information Studies, Novo mesto, Slovenia
Prof. Dr. Enis Kocan, Member, Faculty of Electrical Engineering, University of Montenegro, Podgorica, Montenegro

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Teodora Kocevska

IDENTIFICATION OF INDOOR RADIO ENVIRONMENT PROPERTIES BASED ON CHANNEL STATE INFORMATION USING MACHINE LEARNING APPROACHES

Doctoral Dissertation

IDENTIFIKACIJA LASTNOSTI NOTRANJEGA RADIJSKEGA OKOLJA NA PODLAGI INFORMACIJ O STANJU KANALA Z UPORABO PRISTOPOV STROJNEGA UČENJA

Doktorska disertacija

Supervisor: Asst. Prof. Dr. Andrej Hrovat

Co-Supervisor: Dr. Aleksandra Rashkovska Koceva

Ljubljana, Slovenia, April 2023

To my parents

Acknowledgments

First and foremost, I would like to sincerely thank my supervisor, Asst. Prof. Dr. Andrej Hrovat, for his exceptional guidance, invaluable support, and encouragement throughout my studies, and especially for his confidence in me. He accepted me as his doctoral student without any hesitation. During my studies, he offered me so much advice, insightful feedback, and constructive criticism, patiently supervising me and always guiding me in the right direction. I appreciate all his contributions in time, expertise, and ideas to my research. I have learned a lot from him and I am honored to have worked with such an exceptional supervisor.

Especially, I would like to express my utmost appreciation and gratitude to Prof. Dr. Tomaž Javornik for giving me the invaluable opportunity to join his research group, for his guidance during my doctoral studies, and for his contributions to my research. He has been more than generous with his expertise and time. I am very grateful to him for the numerous discussions, insightful ideas, and constructive feedback that were crucial to my research, as well as for carefully reading the text of the dissertation. His comments and suggestions helped me to improve and finalize the dissertation.

I wish to extend my deep gratitude to my co-supervisor, Dr. Aleksandra Rashkovska Koceva, for her irreplaceable support and guidance, both academically and personally. Her encouragement was invaluable to me when I made the decision to pursue doctoral studies. I am grateful for her willingness to share her knowledge with me and her determination to help me throughout the process. Her expertise and feedback helped me to develop a deeper understanding of the field of machine learning and improve my research.

I would also like to thank Prof. Dr. Aleš Švigelj, Asst. Prof. Dr. Bernard Ženko and Prof. Dr. Enis Kocan for agreeing to be members of the dissertation evaluation committee.

I gratefully acknowledge the scholarship from the Slovene Human Resources Development and Scholarship Fund, which funded my doctoral studies.

I also express my gratitude to the Department of Communication Systems for the wonderful collaboration during my studies and to my colleagues from the Communication Technology Laboratory for creating a productive and supportive work environment.

A special thanks goes to Tomislav Shuminovski and Robert Adamovski, who are not only my colleagues in Macedonia, but also my friends. Thank you for being there to lift my spirit and bring joy to my day. Also, many thanks to Stojche and Irena in Ljubljana who helped me in many situations and made my life abroad easier.

My deepest gratitude goes to my loving family, my parents Ljubica and Cene and my sisters Zhaklina and Natasha, for their unconditional love and support. Last but not least, endless thanks to my loving, supportive, and patient Slavcho. Without the encouragement from you, none of this would have been possible. Thank you for believing in me and living every moment of my doctoral studies. Nothing else is able to motivate me more than you, and nothing can cheer me up more than a smile on your face.

Abstract

Characterization of the indoor radio environment (RE) is a prerequisite for advances in the design and optimization of next-generation indoor wireless networks and for the construction of a digital twin of the building. The need for comprehensive and accurate indoor characterization will be evident in the future hyper-connected mixed real-virtual world, where emerging applications in various areas such as wireless communications, spatial understanding, localization, automation, mediated reality, etc., will be based on environmental awareness. Novel, parsimonious, and intelligent environment awareness methodologies that do not rely on specialized infrastructure and manual intervention are needed.

The dissertation addresses the problem of indoor environment characterization using state-of-the-art wireless technologies and machine learning (ML) approaches. A novel methodology for identifying indoor RE properties based on channel state information (CSI) using ML approaches is proposed, formalized, and evaluated. The methodology is based on two assumptions: the received signal conveys a RE signature, and the RE signature can be estimated by analysing the wireless link.

The procedure for constructing the RE identification model from RE signatures is streamlined into a framework. The framework specifies the RE signature, RE signature acquisition, feature selection from CSI, CSI processing and storage, ML task, and ML workflow. A large data set of CSI data acquired using ultra wideband (UWB) technology in the microwave frequency band and annotated with environmental properties is built. The data set contains channel impulse response (CIR) from a large number of rooms with different sizes and materials for the surfaces and different positions of radio nodes.

The experiments presented in the dissertation provide an evaluation of the proposed methodology for identifying surface materials from CIR, acquired with ray tracing method in plain indoor environments and a comparative analysis along three main aspects: room size, CIR acquisition strategy, and learning method. The ability of the models to generalize to CIR acquisition strategies and room sizes not considered in the training process is evaluated. The results show that the methodology can be applied to identify the material of a single wall as well as the material of all surfaces in plain indoor environments. The impact of radio nodes' position and room size on the model performance is also confirmed.

The methodology is one of the main contributions of the dissertation. The material identification presented in the dissertation is an initial example of indoor RE using the proposed methodology. The methodology can be considered separately or as part of a larger methodology to create an accurate and detailed digital twin of the building. One of its main values lies in its extensibility to different CSI properties, CSI estimation methods, indoor environments, and indoor characterization tasks. The methodology provides the foundation (or environmental context) needed to develop state-of-the-art methods for environmentally aware indoor wireless communications. Its importance is emphasized in the era of next-generation communications, where a detailed description of the propagation environment is a prerequisite for improving communications performance to meet the requirements of emerging applications.

Povzetek

Opis in označba notranjega radijskega okolja (RE) je predpogoj za napredek pri načrtovanju in optimizaciji brezžičnih omrežij naslednje generacije v notranjih okoljih ter za izgradnjo digitalnega dvojčka stavbe. Potreba po celoviti in natančni oceni in označbi notranjega RE bo še posebej očitna v prihodnjem hiper-povezanem mešanem realno-virtualnem svetu, kjer bodo nastajajoče aplikacije na različnih področjih, kot so brezžične komunikacije, prostorsko razumevanje, lokalizacija, avtomatizacija, posredovana resničnost, itd., temeljile na zavedanju okolice. Zato potrebujemo nove, varčne in inteligentne metodologije zaznavanja okolja, ki niso odvisne od specializirane infrastrukture in ročnega posredovanja.

Disertacija obravnava problem ocene in označbe notranjega okolja z uporabo najsodobnejših brezžičnih tehnologij in pristopov strojnega učenja (ML). Predlagana, formalizirana in ovrednotena je nova metodologija za ugotavljanje lastnosti notranjega RE na podlagi informacij o stanju kanala (CSI) z uporabo pristopov ML. Metodologija temelji na dveh predpostavkah: sprejeti signal vsebuje podpis RE in podpis RE je mogoče oceniti z analizo brezžične povezave.

Postopek izgradnje ML identifikacijskega modela RE iz podpisov RE je strnjen v razvojno ogrodje, ki določa podpis RE, pridobivanje podpisa RE, izbiro značilnosti iz CSI, obdelavo in shranjevanje CSI, naloge in delovni tok ML. Ustvarili smo tudi velik nabor podatkov CSI, pridobljenih z upoštevanjem ultra širokopasovne tehnologije (UWB) v mikrovalovnem frekvenčnem področju. CSI smo opremili tudi z oznakami lastnosti okolja. Nabor podatkov vključuje podatke CIR velikega števila sob različnih velikosti, katerih stene so iz različnih materialov. Podatke smo pridobili z različnimi položaji radijskih vozlišč.

Testni scenariji, predstavljeni v disertaciji, zagotavljajo oceno ustreznosti predlagane metodologije za prepoznavanje površinskih materialov iz CIR, ki smo jih pridobili z metodo sledenja žarkom v preprostih notranjih okoljih. Opravili smo primerjalno analizo glede na tri glavne vidike, in sicer velikost sobe, strategijo pridobivanja podatkov CIR in metodo učenja. Ocenjena je sposobnost posplošitve modelov na strategije pridobivanja CIR in velikosti prostorov, ki niso bile upoštevane v postopku učenja. Rezultati kažejo, da se metodologija lahko uporablja za identifikacijo materiala ene stene in materialov vseh sten v preprostih notranjih okoljih. Rezultati tudi potrjujejo, da položaj radijskih vozlišč in velikost prostora vplivata na lastnosti modelov.

Metodologija je eden glavnih prispevkov disertacije. Identifikacija materiala stene v notranjih okoljih, predstavljena v disertaciji, je začetni primer uporabe predlagane metodologije. Mogoče jo je obravnavati ločeno ali kot del večje metodologije za izdelavo natančnega in podrobnega digitalnega dvojčka stavbe. Ena od njenih glavnih vrednosti je razširljivost na različne lastnosti CSI, metode ocenjevanja CSI, notranja okolja in naloge opisovanja notranjih prostorov. Metodologija zagotavlja podlago (ali okoljski kontekst), ki je potrebna za razvoj najsodobnejših metod za okoljsko ozaveščene brezžične komunikacije v notranjih okoljih. Njen pomen je poudarjen v luči komunikacije naslednje generacije, kjer bo podroben opis okolja razširjanja radijskih signalov predpogoj za izboljšanje komunikacijske zmogljivosti za potrebe nastajajočih aplikacij.

Contents

List of Figures

List of Tables

Abbreviations

1T	...	single target
2-D	...	two-Dimensional
3-D	...	three-Dimensional
5G	...	fifth generation wireless systems
6G	...	sixth generation wireless systems
BIM	...	Building Information Modeling
CAD	...	Computer-aided Design
CART	...	Classification and Regression Trees
CIR	...	Channel Impulse Response
CNN	...	Convolutional Neural Networks
CSI	...	Channel State Information
CSV	...	Comma Separated Values
CTF	...	Channel Transfer Function
DT	...	Decision Tree
EM	...	electromagnetic
EPC	...	Exact Path Calculator
FCF	...	Frequency Coherence Function
FDTD	...	Finite Difference Time Domain
FEM	...	Finite Element Method
FMCW	...	Frequency Modulated Carrier Wave
FN	...	False Negative
FP	...	False Positive
GO	...	Geometrical Optics
GPU	...	Graphics Processing Unit
GTD	...	Geometrical Theory of Diffraction
HP	...	horizontal polarization
ID3	...	Iterative Dichotomiser 3
IoE	...	Internet of Everything
IoT	...	Internet of Things

k-NN ... k-Nearest Neighbor
LCR ... inductance-capacitance-resistance
LiDAR ... Light Detection and Ranging
L ... large
LOS ... Line of Sight
MIMO ... Multiple-Input Multiple-Output
ML ... Medium+Large
MLP ... Multilayer Perceptron
M ... medium
MoM ... Method of Moments
MPC ... multipath component
MT ... multiple targets
OFDM ... Orthogonal Frequency-Division Multiplexing
OVA ... One-vs-All
OVO ... One-vs-One
OVR ... One-vs-Rest
ReLU ... Reflected Linear Function
RE ... Radio Environment
RF ... Random Forest
RGB ... Read-Green-Blue
RSSI ... Received Signal Strength Indicator
SBR ... Shooting and Bouncing Rays
SLAM ... Simultaneous Localization and Mapping
SL ... small+large
SM ... small+medium
SMO ... Sequential Minimal Optimization
S ... small
SVM ... Support Vector Machine
Tanh ... Hyperbolic Tangent Function
TDIDT ... Top-Down Induction of Decision Trees
TN ... True Positive
TP ... True Negative
UTD ... Uniform Theory of Diffraction
UWB ... ultra wideband
VNA ... Vector Network Analyser
VP ... vertical polarization
WLAN ... Wireless Local Area Network

Symbols

X	...	description space
X_i	...	i-th instance
D	...	number of descriptive variables
y	...	target vector
K	...	number of classes
I	...	instance space
N	...	number of instances
q	...	quality criterion
h	...	hypothesis
Λ	...	label space
Q	...	number of possible labels
k	...	number of propagation characteristics
R	...	number of propagation paths
P	...	received power
Φ	...	phase shift
τ	...	excess delay
M	...	material label
ϵ	...	electric permittivity
ϵ_0	...	permittivity in free space
ϵ_r	...	relative permittivity
$\epsilon_r^{'}$	...	real part of relative permittivity
$\epsilon_r^{''}$	...	imaginary part of relative permittivity
σ	...	conductivity
ω_c	...	angular frequency
f_c	...	frequency
μ	...	magnetic permeability
μ_0	...	permeability in free space
μ_r	...	relative permeability
Γ	...	reflection coefficient

Γ_{HP} ... reflection coefficient for horizontal polarization

Γ_{VP} ... reflection coefficient for vertical polarization

α ... angle of incidence

E^i ... incident electric field

E^r ... reflected electric field

$E^r_{\parallel}$ ... component of the electric field parallel to the reflection plane

$E^r_{\perp}$ ... component of the electric field perpendicular to the reflection plane

s_i ... i-th CIR component

P_i ... power carried by the i-th multipath component

Ψ_i ... phase of the i-th multipath component

V_i ... complex voltage

θ ... theta coordinate in spherical coordinate system

ϕ ... phi coordinate in spherical coordinate system

$E_{\theta,i}$ ... theta component of the electric field of i-th multipath component

$E_{\phi,i}$ ... phi component of the electric field of i-th multipath component

g_θ ... theta component of direction of arrival

g_ϕ ... phi component of direction of arrival

G_θ ... theta component of antenna gain

Ψ_θ ... relative phase of the theta component of the far zone electric field

$F1$ ... F1 score

Chapter 1

Introduction

1.1 Motivation

Wireless communication networks will play a crucial role in smooth interaction between people, data, processes, and things. To meet the requirements of a diverse range of Internet of Everything (IoE) applications beyond just human use, wireless communications must be intelligent, environmentally aware, and capable of adapting dynamically and rapidly to the propagation environment. Therefore, a detailed radio environment (RE) description is vital to support the development of wireless communication systems.

Current methods to characterize indoor environments are limited in the environmental information they provide, only providing details such as object distance, surface and object geometry, and isolated sample dielectric properties [1]. These methods are also costly and time-consuming, making them impractical for larger spaces. Additionally, specialized equipment and trained personnel are required, making them inaccessible for widespread use and presenting logistical challenges. Furthermore, their inflexibility due to limitations in adaptability, cost-effectiveness, and maintenance requirements makes them unsuitable for the future intelligent world. To accurately and seamlessly characterize buildings in future scenarios, it is more beneficial to use novel methods that reduce unnecessary complexity, eliminate the need for manual intervention, and incorporate automatic environment reconstruction techniques. These methods should also be efficient, flexible, and agile.

The motivation for the research presented in this doctoral dissertation is the need for a comprehensive characterization of the radio propagation environments within a building, with a specific focus on the electromagnetic (EM) characteristics of the building's structure. Characterizing building properties at the radio signal propagation level is crucial for designing and optimizing future wireless systems and enhancing their performance in indoor environments. This information can also facilitate spatial understanding, automation, and mediated reality in the future mixed real-virtual world. With the enormous number of connected devices in future wireless networks and their widespread deployment in indoor spaces with different geometries and EM properties, sensing the propagation environment with radio waves has become an unprecedented opportunity. The large amount of data that will be available paves the way for the joint use of wireless technologies and machine learning (ML) techniques for the identification of indoor environment properties.

The use of the received signal properties to achieve ML-based characterization of indoor radio propagation environment still needs to be investigated. A learning-based approach for propagation environment characterization holds great promise in tackling complex issues like material characterization, for which conventional methods are costly, time-consuming, and dependent on specialized equipment. The theory of EM radiation provides the foundation for understanding the impact of the indoor environment on radio propagation which is

crucial for the tailored construction of ML models. Nonetheless, how to utilize the domain knowledge to inform data-driven approaches, enhance the learning process, and ensure the reliability of the models is not reported.

Having accurate information about the propagation environment is crucial for developing solutions that aid wireless communications by sensing and utilizing the environment. This field of research holds immense potential for the future of environmentally aware communications and is vital for meeting performance requirements while reducing environmental pollution caused by EM waves. Furthermore, it has a wide range of benefits for various applications in other domains that depend on a detailed description of the indoor environment, ultimately facilitating human activities in future buildings. With the increasing urban population, the prevalence of indoor structures, and people engaging in indoor activities such as work, study, sport, and leisure, the importance of propagation environment characterization for high-performance indoor communication is further highlighted.

The possibility of estimating the propagation environment using the RE signature included in the received radio signals motivated our research toward radio-based scanning of buildings. The interaction between the propagating waves and the propagation environment surrounding the wireless link leaves information about the multipath characteristics of the indoor structure in the waves reaching the receiver. Recent advancements in wideband radio technologies, specifically in estimating CSI, allowing the received signal to be utilized for characterizing indoor environments. The idea of using RE signatures and ML models for RE characterization is illustrated in Figure 1.1. The large number of radio devices operating in high-frequency bands, with sensing capabilities, deployed indoors opens the possibility of sharing existing indoor wireless infrastructure for communication and sensing. Therefore, the wireless sensing infrastructure could be used for RE acquisition. The properties of the RE result in a unique distortion of the received signal, i.e., the RE signature. It is foreseen that the ML algorithms can extract knowledge about the relationship between the RE signature and RE properties based on a corpus of RE signatures from spatially distributed wireless links and learn a RE identification model. The RE identification model is expected to accurately predict the RE properties (EM properties of materials, size, shape, and roughness) from the RE signature for a new indoor environment. The detailed description of the RE properties could be included in a multi-layered description of the building, which holds information about various aspects of the building obtained from heterogeneous sensing methods, including cameras, infrared lasers, MIMO-based sensing methods, ML-based methods, and new methods that will be developed in the future. The enhanced RE awareness can serve as a basis for developing novel methods toward environmentally aware wireless communications.

Indoor scanning methodologies that do not depend on a cumbersome and expensive infrastructure for a specific environment scanning function are desirable for cost-effective and ubiquitous sensing. The correlation between the received signal and the propagation environment properties makes the propagation environment characterization problem adequate for ML analysis. Over the last decade, ML techniques gained prominence in different domains concerned with data-intensive problems and have proven the ability to extract knowledge from labeled data. To train models that can identify the propagation environment from CSI, a large amount of data from different propagation environments and communication system configurations is necessary.

In the hyper-connected world, the growth in data available across the infrastructure of wireless communication systems will persist and even intensify with the ubiquitous indoor deployment of emerging broadband wireless communication technologies. The expected availability of a vast amount of data, with embedded information about the multipath

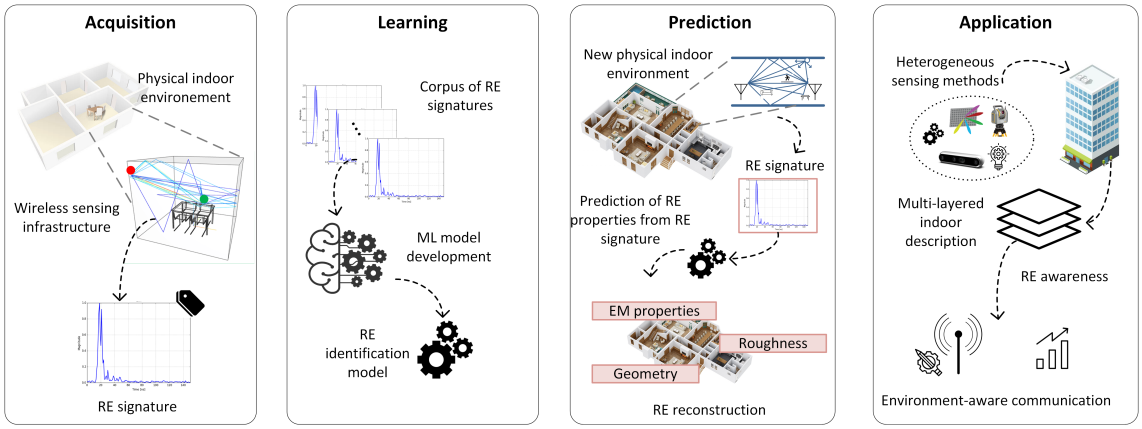

Figure 1.1: Indoor RE characterization using RE signature and ML. Illustration of the idea.

characteristics of the propagation environment, across the ubiquitous wireless infrastructure motivated our research in data-driven identification of the propagation environment properties. The wireless network infrastructure will become an essential enabler for using ML techniques to tackle the problem of propagation environment characterization and other complex problems from various verticals in smart buildings. The emerging wireless technologies and infrastructure will enable the ML-based identification of the properties of the propagation environment that affect indoor propagation by providing data that conveys information about the multipath characteristics of the environment and will provide potential for communication performance optimization and environmental-awareness enhancement.

1.2 Aims and Objectives

The main aim of this dissertation is to enhance environmental awareness about the physical inner building structure with information about the characteristics of the propagation environment that affect indoor wireless communication by exploiting the indoor radio networks, the RE signature included in the received signal, and the capability of broadband radio technologies to estimate the CSI. Environmental awareness is required to develop wireless communication systems that cope with the increasing demand for connectivity and elevate the user experience to a higher level. The wireless network infrastructure and ML methods are essential enablers toward this aim. In particular, the future radio networks deployed inside buildings will be the primary source of data containing the environmental signature. Consequently, they will foster the use of ML for emerging problems in wireless communications. ML is an essential tool to extract knowledge from the data to identify the propagation environment based on CSI.

Hence, the main scientific objective is to formalize and evaluate a methodology for the characterization of the indoor propagation environment in terms of the properties of the bounding surfaces that affect the indoor propagation from RE signature (or CSI) by taking advantage of the ubiquitous infrastructure of cutting-edge wireless technologies and state-of-the-art ML approaches capable of learning models from data.

To achieve the main objective, the following specific objectives shall be accomplished:

- O1: **Build relevant theoretical foundations from indoor radio propagation and ML domains for domain knowledge-driven use of ML.** The theoretical

foundations of both wireless communication and ML are crucial for the development of an optimal methodology that exploits the complementary strengths of both domains. The communication perspective is imperative for the informative use of powerful ML methods and accurate identification of the propagation environment.

- O2: **Develop a method for radio-based acquisition of the propagation environment properties.** Radio-propagation-based environment acquisition method is essential for seamless environment scanning and constructing an open sensing database.

- O3: **Collect data sets with propagation characteristics from various physical environments and communication system configurations and make them available to support research in intelligent wireless sensing.** The sensing database is imperative for constructing machine-learning models capable of predicting the propagation environment.

- O4: **Propose a ML workflow for propagation environment prediction based on CSI.**

- O5: **Evaluate the proposed methodology for material characterization of the inner building structure.**

1.3 Hypothesis

The main hypotheses of the dissertation are as follows.

- H1: **The channel impulse response (CIR) contains a signature of the indoor RE.** Due to the interaction of the radio waves with the surrounding, the CIR includes information about the geometry and EM properties of the indoor environment.

- H2: **Materials in the propagation environment can be predicted from CIR with ML models.** ML methods can build models capable of predicting the materials in the propagation environment for input CIRs estimated for traced rays.

- H3: **Due to the different EM properties, some materials can be identified more accurately than others.** Some materials have a greater effect on radio propagation than others; thus, the environment information loaded in the CIR is richer, and some materials can be identified more accurately than others.

- H4: **The wireless link position relative to the room affects the prediction performance.** Based on the position of the transmitter/receiver, some wireless links convey richer environmental information compared to others. Consequently, the propagation environment can be predicted more accurately based on some links than on others.

- H5: **Including data from various rooms and radio node positions in the training process results in more general models with wider practical applicability.** Including CIRs from a large number of rooms with different sizes and surface materials estimated with different radio node positions in the training process results in models able to make accurate predictions for larger set of rooms.

1.4 Scientific Contribution

This dissertation contributes to state-of-the-art interdisciplinary research on wireless technologies for seamless sensing of indoor radio propagation environments, using ML to analyze the abundance of propagation data provided by emerging communication systems and extract knowledge about the correlation between CSI and propagation environment characteristics.

In particular, the scientific contributions of the dissertation are summarized as follows:

- C1: **Use of RE signature for identification of the materials inside buildings.** We exploit the information the materials from the reflective surfaces leave on the received signal.

- C2: **Adoption of ML techniques as a tool for RE characterization.** Adopting ML methods for learning models capable of predicting the RE's propagation-level properties from RE signatures is an original unpublished/unattended approach.

- C3: **Novel interdisciplinary method for intelligent (i.e., ML-based) wireless sensing of radio propagation environment inside buildings using CSI.** The breakthrough provided by the proposed method is the ability to predict the indoor propagation environment properties by taking advantage of the environmental information embedded in the radio signals. Hence, the solution does not require dedicated sensing equipment and human operator, which makes it seamless, cost-effective, and less time-consuming.

  This contribution is composed of several important contributions:

  - C3.1: **Indoor environments and communication system configurations for training models.** We select several indoor environment setups which differ in the shape and dimension of the indoor structure and material of the surfaces. We chose several communication system setups described by wireless technology, wireless node location and setup, antenna configuration, frequency and bandwidth, and channel state estimation methods for collecting propagation data needed for building relevant models.
  - C3.2: **CIR parameters as input for the learning task.** We use domain expertise to select the CIR parameters as channel state characteristic that conveys rich environment description that can be used as a data point in the sensing data sets.
  - C3.3: **Suitable ML workflow.** We adopt ML techniques as a tool to implement the RE characterization with the CIR. The problem is formalized as a ML task, and we selected an appropriate state-of-the-art ML workflow for the learning task. The workflow includes (i) designing the data sets and generating features, (ii) selecting ML techniques suitable for the task and data, and (iii) selecting evaluation metrics and evaluation scenarios.
  - C3.4: **Inside building material characterization.** We evaluate the proposed method for material characterization of plain rooms with two levels of complexity in terms of the number of materials affecting wireless propagation. We investigated several scenarios in terms of room sizes and link layouts included in the learning process.

- C4: **Ultra-wideband indoor CSI data sets available as part of the open indoor sensing database.** We collect a large and diverse set of CSI from wireless

links located in various locations in numerous rooms with different sizes and EM properties related to the materials used for the surfaces. We label the datasets with environmental properties, i.e., the shape, dimensions, and material of each surface in the room. Thus, we provide representative annotated datasets for training, validating, and testing ML models and make them available to the research community for further characterization and modeling.

- C5: **The proposed method provides environment context, which can enhance the buildings' digital twin and contributes to developing pioneering methods for improved radio communication in the next-generation communication era.** The comprehensive description of the environment enables solutions for many vertical applications in future smart buildings and methods for predicting the radio channel beyond the current state of the art to improve indoor wireless communications.

1.5 Dissertation Structure

In the introductory chapter, we give the motivation for identifying indoor RE properties based on CSI using ML approaches. Also, we state the aims and hypotheses of the dissertation, followed by the original scientific contributions of the dissertation.

In Chapter 2, we provide an overview of the background knowledge from the radio propagation and ML areas used for the dissertation. The theoretical foundations related to indoor radio propagation, the impact of the indoor environment on radio propagation, the use of ray tracing techniques in indoor environments, and ML with emphasis on approaches for tackling classification learning tasks are covered.

In Chapter 3, we present a review of the state-of-the-art literature in the key research areas related to our work: convergence of sensing and communication, characterization of indoor environments, identification of indoor environment, and EM characterization of materials, and we identify the gaps in the literature on environmental awareness.

In Chapter 4, we propose a novel methodology for intelligent CSI-based characterization of the indoor RE. First, we describe the identification of RE properties based on RE signature using ML models. Next, we formalize the framework that streamlines the procedure for the characterization of indoor RE with a RE identification model. Finally, we expound upon the practical application of the RE identification model in characterizing a specific indoor environment.

In Chapter 5, we elaborate on the procedure for experimental evaluation of the proposed methodology for material characterization in indoor RE. Firstly, we delineate the experimental questions that we consider. Secondly, we specify the relevant characteristics of the indoor environments and communication system setup that are considered in the experiments. Thirdly, we define the experimental scenarios, and we outline the procedure for building the data sets, followed by a description of the data sets. Finally, we define the evaluation measures for assessing the ML model's performance.

In Chapter 6, we report the results obtained from the experiments and discuss them in the context of the research objectives and questions.

Finally, we conclude the dissertation with Chapter 7, where we summarize the dissertation, emphasize the scientific contributions and give directions for future work.

Chapter 2

Background

This dissertation addresses the problem of indoor environment characterization through the use of wireless technology and machine learning. Wireless technology is used to acquire indoor radio propagation data, and machine learning methods are utilized to learn models capable of predicting the indoor RE properties from the RE signature. This chapter provides the necessary background for this dissertation, covering indoor radio propagation and machine learning tasks and methods. It begins with an explanation of how surface materials affect radio wave propagation, followed by a description of the RE signature and the ray-tracing method. The chapter continues with a formal definition of the multi-class and the multi-label classification tasks and an overview of the learning methods used in this dissertation.

2.1 Radio Wave Propagation in Indoor Environments

The transmission medium between the transmitter and the receiver in indoor wireless communications is usually referred to as indoor radio channel [2], [3]. The indoor radio channel causes impairments to the radio signals, which can be mitigated by the appropriate design of transmitters and receivers. Another approach that can be used to mitigate the challenges encountered indoors is to proactively reconfigure the wireless propagation environment using reconfigurable intelligent surfaces [4]. The information that has to be transmitted from the transmitter to the receiver is carried by radio waves. Several wavelengths from the transmitting antenna, the radio waves are plane waves [3]. A plane wave's electric and magnetic fields are perpendicular to each other and the direction of the propagation of the wave. These two fields are in phase at any point in time or space and the wavefront is the plane normal to the direction of the propagation. Plane waves can be represented by radio rays which are lines indicating the direction of propagation [5]. When the ray concept is valid, the radio wave propagation inside buildings can be described with the radio propagation mechanisms such as reflection, refraction, transmission, diffraction, and scattering [3].

A radio link between a transmitter and a receiver with omnidirectional antennas in an indoor environment is depicted in Figure 2.1. The indoor radio propagation depends on the indoor environment geometry (size and shape), the EM properties of the construction materials, the transmitter/receiver location, the number of floors and walls in-between the transmitter and the receiver (in the case when the transmitter and receiver are located in different floor or rooms), the height where the transmitter and receiver are mounted, the distance between the transmitter and the receiver, the antenna parameters, the operating frequency, the polarization, the layout of rooms, and the obstacles (furniture, furnishing, plants, objects, and people) [3], [5]. The obstacles in the indoor propagation environment

interact with the propagating radio waves. Due to the various indoor propagation mechanisms, the transmitted radio wave often reaches the receiver through numerous propagation paths [2]. These additional copies of the transmitted wave, known as multipath components (MPCs), can be attenuated in power, delayed in time, and shifted in phase compared to the line of sight (LOS) path. The MPCs arriving from indirect paths and the direct path (if LOS exists) combine at the receiver, which often results in distortion of the received wave relative to the transmitted wave. Depending on the phases of the components, they can sum up as constructive or destructive.

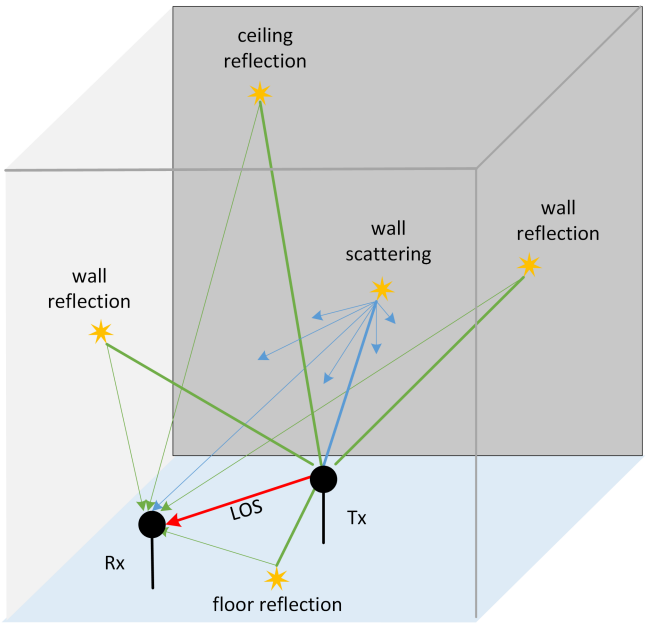

Figure 2.1: Schematic representation of multipath propagation in an indoor environment. The room is represented as a cube. The yellow stars show the rays' interactions with the surfaces. The direct path is denoted by a red line, single bounce reflections are illustrated by green lines, and the scattering is represented with blue lines. Tx: transmitter, Rx: Receiver, LOS: LOS.

When the obstacles have very large dimensions compared to the wavelength, the propagating radio wave is reflected and part of the energy penetrates the interacting object. Typically, inside a building, a reflection occurs from the large and smooth surfaces of the walls. The propagation direction of the reflected and the transmitted rays is determined by the law of reflection [3]. The magnitude of the reflected and transmitted fields is given relative to the magnitude of the incident field by Fresnel's reflection and transmission coefficients [3]. The coefficients depend on the EM properties of the materials, frequency, and polarization of the propagating radio wave, and angle of incidence [6]. If the surfaces of the interacting object are rough, the reflected waves are scattered. The degree of scattering depends on the angle of incidence and the roughness of the surface in comparison to the wavelength. The apparent roughness of the surfaces is reduced as the incident angle comes closer to the grazing incidence. Scattering becomes much more significant for waves with high frequencies. Diffraction is not a dominant indoor propagation mechanism, but it emerges on the border between two different materials, e.g., at the edges of windows and doors, where glass or wood meets with walls built of concrete, brick, or plaster, or on the sharp change in the shape of the building structure, e.g., at the corners and edges between walls or walls and ceiling.

The geometric and dielectric characteristics of the indoor propagation environment significantly vary for the various buildings in practice. For instance, the mechanisms behind the radio propagation are different in open spaces with minimal obstructions, typically factories in industrial environments, and enclosed spaces with many obstacles, typically cubical rooms in office and residential environments [7]–[12]. The age of the building and the area where the building is located may additionally determine the construction practices applied.

Due to the motion of the obstacles and the multipath propagation, the transmission medium varies strongly with time. In crowded and dynamic indoor environments, such as shopping malls, airports, etc., the channel statistics change even when the transmitter and the receiver are fixed. In residential and office buildings with low dynamics, it can be assumed that the channel is quasi-static over a short period.

2.1.1 Effect of building materials on radio wave propagation

Inside the building, the propagating radio wave traverses from one media to another and the boundaries between two media with different EM properties result in changes in the amplitude, phase, and direction of the propagating radio waves [13]. Usually, one of the media is air and the other is the construction material of the indoor structure. The EM properties and the material structure affect the radio wave, which results in signal attenuation. At high-frequencies, the propagation is affected also by the granularity on the surfaces, which comes from the type of construction materials used, their texture, and variation in density, and porosity.

The EM properties of materials have a dominant effect on the reflection from, and transmission of, radio waves through building materials and on the absorption of radio wave energy in these materials [3]. On the other hand, the EM properties of the building materials are not of primary importance when determining the effect of diffraction and scattering on the radio waves. A frequency-dependent assumption can be made that the reflection is the dominant mechanism in buildings where the space is bounded by walls, doors, and windows, which can be approximated as planar, quasi-two-dimensional structures with smooth surfaces compared to the wavelength. When the wave reaches the plane interface between air and another material with different electrical properties, due to changes in the wave impedance, part of the energy reflects while the remainder penetrates through. The magnitudes of the reflected and transmitted components can be determined from the incident field using the Fresnel reflection and transmission coefficients [3]. The coefficients depend on the EM properties of the two media, properties of the incident wave (frequency and polarisation), and the angle of incidence. The transmitted field exponentially decreases with the distance traveled through lossy material. For multi-layer slabs, multiple internal reflections occur in the layers. When the ends of the radio link are not separated by a wall or partitions, through the wall penetration can be just considered as transmission loss. Conventional building materials are non-ionised and non-magnetic, therefore the electrical properties of the material (permittivity and conductivity) affect the reflection. The evidence from practice shows that for many materials the real part of the permittivity is constant in the range of 5-10 GHz and after that starts to fall with frequency, while the conductivity strongly increases with frequency [13]. The frequency dependence makes it hard to obtain representative data with material properties close to the frequency of interest.

Some of the construction practices, mainly related to the construction method for the walls and layering method for the windows, may vary in different regions because of common practices specific to an area or government regulations. In some buildings, materials such as brick and concrete are usually used for the walls, regardless of whether they are

load-bearing or not, inner or outer. In other buildings, reinforced concrete is commonly used for load-bearing walls, plaster for interior partitions, and brick with insulation for the outer walls. The need to improve thermal performance and energy efficiency, driven by zero-carbon requirements in future smart buildings, means that different approaches to layering surfaces are currently or will be used in the future. Wood is commonly used as a building material, especially in residential buildings, but also in offices. Most often it is used for flooring, but it is not uncommon for it to be used for wall surfaces for thermal or interior design reasons. For windows, construction trends range from single-glazed windows with a single pane of glass to energy-efficient windows with multiple layers of glass and/or thin metal foils separated by a vacuum.

2.1.2 RE signature

The proposed methodology is based on the assumption that the received signal conveys RE signature as a result of an environment-specific distortion of the propagating waves [14]. The RE signatures can be further analysed and used to understand the characteristics of the space and the wireless propagation channel. We define the RE signature as distinctively unique information about the propagation environment around a wireless link that is specific to the layout and configuration of the end-nodes. The signature of the environment is closely related to the rich multipath channel in indoor environments. The indoor channel is characterized by the presence of MPCs with different properties (amplitude, delay, phase shift, and angle of arrival) that are created due to interactions between the waves and structures composing the space, such as walls, floors, ceilings, and other objects. The signature includes information about how the waves are reflected, absorbed, or scattered by the materials and the structures within the space. The signature that the environment leaves on the propagation of EM waves can vary significantly between different wireless links due to a variety of factors that influence the behavior of these waves within a space. One key factor is the frequency of the EM waves being used. Different wireless technologies use different ranges of wavelengths and frequencies, and these can be affected differently by the materials and structures within a space. For example, radio waves with longer wavelengths may be more likely to penetrate walls and other structures, while waves with shorter wavelengths may be more easily absorbed or scattered by these materials [3]. The use of multiple antennas and placement of antennas at specific heights or orientations can also affect the signature of the environment, as the waves may be reflected or absorbed differently by the materials and structures within a space depending on the angle of incidence. The layout and configuration of the space itself can also influence the signature. For example, large flat surfaces like walls will reflect the waves. Depending on the materials used and their texture compared to the wavelength, the walls may also absorb and scatter the waves. The presence of people and other objects within a space can additionally affect the signature. Therefore, each link depicts the surrounding in a unique way determined by the position of the radio nodes, geometry and EM properties of the environment, and configuration of the communication system.

2.1.3 Ray-tracing method

The ultimate details of the EM propagation characteristics could be exactly computed by solving Maxwell's equations with boundary conditions that express the physical properties of the walls and other structures in the building that interact with the radio waves [3]. In a real environment consisting of a large number of elements and made of materials with various EM properties, precise consideration of Maxwell equations requires complex mathematical operations and considerable computing power. Therefore, approximate numerical

methods are of interest [15].

Ray-tracing is a deterministic propagation prediction method based on Geometrical Optics (GO) that can be used for approximate estimation of the high-frequency EM fields [16]–[19]. Ray-tracing is less demanding in terms of computation than the numerical methods for solving the Maxwell's equation, such as Finite Difference Time Domain method (FDTD) [20], Method of Moments (MoM) [21], and Finite Element Method (FEM) [22]. The ray-tracing method is used to determine the MPCs including their amplitudes, time delays, and phases. Detailed knowledge of the environment (geometry described with the location of the facets and electrical properties of the building facets given with the permittivity, conductivity, and the reflection and transmission coefficients), the location of the radio nodes (transmitter and receiver), antenna patterns and frequency of operation are a required input for the ray-tracing method. The accuracy of the ray-tracing method depends on the accuracy of the description of the propagation environment. It is most accurate when the point of observation is many wavelengths from the nearest obstacles, and all obstacles are large compared to the wavelengths and smooth (the surface irregularities are small compared to the wavelength).

The concept of ray-tracing is based on the GO assumption that the energy can be considered to be radiated through infinitesimally small tubes along a certain straight line called ray. Therefore, high-frequency radio waves can be approximated with rays and the radio wave propagation can be modeled as ray propagation. The direction of the radio ray is perpendicular to the plane wave's electric field vector and magnetic field vector [3]. The rays are launched from a transmitter location and the interaction of the rays can be described with a theory of reflection and refraction. In GO only direct, reflected, and refracted rays are considered. The GO theory is complemented by the Geometrical Theory of Diffraction (GTD) and its uniform extension, the Uniform Theory of Diffraction (UTD) by introducing new types of rays, called diffracted rays [23], [24]. Diffracted rays are produced by incident rays that hit edges, corners, or vertices of boundary surfaces. Away from points of diffraction, the diffracted rays propagate as GO rays. The Fermat principle and the principle of the local field are two basic principles extensively used in the ray models [25]. The Fermat principle states that a ray follows the shortest path from a source point to a field point, while the principle of the local field states that high-frequency rays produce reflection, refraction, and diffraction when hit on a facet. This depends only on the electrical and geometrical properties of the obstacle at the point of interaction.

The principle of ray handling is an essential aspect of the ray-tracing algorithm. Based on the ray handling approach, there are two main methods: the brute force method and the method of images [15]. The methods are illustrated in Figure 2.2 and Figure 2.3.

The first method uses a brute force approach for tracing a large number of rays from the transmitter in all directions in the scene. The rays interact with the environment before they reach the receiver. The concept of a reception sphere with the adequate radius (depends on the path length of a particular ray and the angle between neighbouring shot rays) is used to detect the rays passing close to the receivers. The accuracy of the method is determined by the number of rays and the distance from the transmitter to the receiver location. The rays at the source point can be obtained using two-dimensional (2-D) and three-dimensional (3-D) methods [26], [27]. In two dimensions, the rays are launched along different directions in a plane with fixed angular spacing so that each ray carries similar power. The selection of the angle depends on the required accuracy and the computation time. In three dimensions, the rays are launched at an elevation and azimuth angle, and antenna patterns are incorporated to include the effects of antenna beam width. The ray-tracing algorithms that use this method refer to the principles of ray launching [28], pincushion method [29] or shooting and bouncing rays (SBR) [30]. The computational

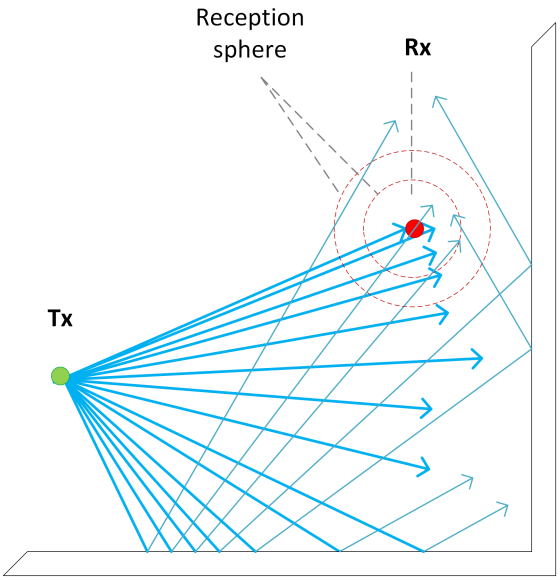

Figure 2.2: Shooting and bouncing rays in an indoor environment—presentation in two dimensions. Tx: transmitter, Rx: receiver.

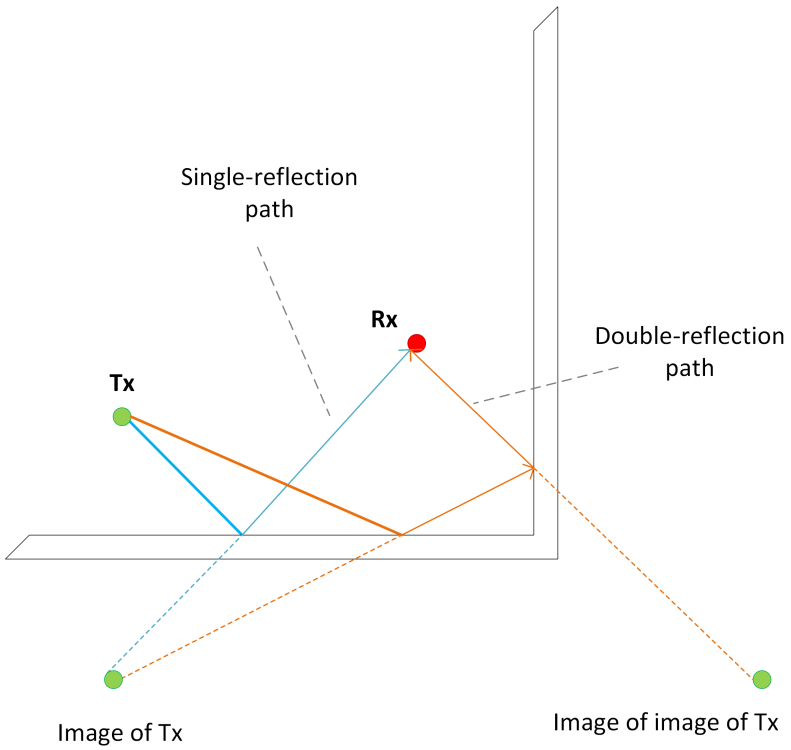

Figure 2.3: Method of images in an indoor environment—presentation in two dimensions. Tx: transmitter, Rx: receiver.

complexity of the brute force method depends on the number of facets and objects in the propagation environment for which an interaction test has to be made and the number of shot rays. The method of images is based on image theory. It assumes each facet of the indoor environment to be a mirror. The path of the ray reflected from a facet is

represented as a ray radiated directly from the virtual source which is a symmetrical image of the transmitter with respect to the plane that contains the facet.

The ray-tracing method is widely used for indoor propagation modeling. In this study, a 3-D propagation model (X3D) that runs on a graphics processing unit (GPU) and uses multi-threading to shorten the run times is considered [31]. The model includes reflections, transmissions, and diffractions along with frequency-dependent atmospheric absorption. After finding the SBR path, X3D implements Exact Path Calculator (EPC) algorithm [31] to find the paths that reach the exact receiver location. This correction reduces the errors in calculations without the longer run times required by the methods based on image theory.

2.2 Machine Learning

Machine learning (ML) is a field of artificial intelligence that lays at the intersection between computer science and statistics and is the core of data science [32], [33]. It is concerned with computer programs and systems that are able to leverage data to learn and automatically improve through the experience without being explicitly programmed. The ML field is intertwined with the field of data mining (also known as knowledge mining) which uses ML algorithms/methods/schemes to discover useful knowledge from a vast amount of data [34]. ML has progressed significantly over the past two decades in laboratory research but also in practical implementation in widespread commercial use [35]. The progress was driven by the development of theory in the field and the availability of open data and low-cost computation. Recently, methods based on ML have been successfully applied in diverse fields ranging from science, engineering, biology, medicine, finance, marketing, manufacturing, entertainment, agriculture, and more. ML is foreseen appropriate for the identification of indoor RE properties from RE signature due to (i) the correlation between the environmental information loaded in the signature and the indoor environment and (ii) the large amount of indoor propagation data in the next-generation wireless networks.

Inductive learning is an important concept in ML, where the experience given as an input to the ML algorithm is in the form of a training data set that contains the complete set of data available as a set of examples/instances/records. The output produced by the learning algorithm is a ML model [36]. The model is constructed based on available data that is part of all data and is expected to generalize (to be applicable) from new instances in data that were not seen during the learning. Each instance is an individual example of the object of interest and it is characterized by a number of attributes/variables that correspond to its properties. Since the size of the training data set and the predictive performance are positively correlated, the largest possible training data set is preferred. However, in practice, due to limited computational resources, scarcity of data, cost of annotation, or requirements of the particular application, the common practice is to limit the size of the training data set. Traditionally, the training data set is given in the structured form, most often as a flat table where each row is a training instance and each column is an attribute, or alternatively as a relational database. Recently, emerging data sources (such as Internet of things (IoT), multimedia, rich media, documents etc.) produce unstructured data (data that does not fit in a table and is a conglomeration of many varied types of data stored in different native formats) which requires specific processing to be converted to structured and fed to available tools or development of new tools that are capable of handling it in its native form.

The properties of an object can be represented with different types of variables that have to be considered in the selection of a ML algorithm. The attributes may belong to two

main types: (i) categorical/qualitative attributes that can usually be nominal, ordinal, or binary, and (ii) numeric/quantitative that can usually be discrete or continuous [34]. Data sets often contain variables that are categorized as ignore attributes, these are variables that are not significant for the application but because of some reason are kept in the data set.

In ML, data can be categorized into labeled and unlabeled. Labeled data is data that has been assigned a specific attribute or label that represents a property for which the ML system should learn to predict its value for unseen data samples. It is obtained through a process known as data annotation where unlabeled samples are labeled (also called marked/tagged) with outcomes that are used in the learning process. On the other hand, unlabeled data refers to data that has not been assigned any label or category.

Depending on the specific learning goal and the characteristics of the available data, the following ML approaches exist: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [34], [36], [37]. Supervised learning is a mature and widely researched area in ML that relates to concept learning [36]. The focus in supervised learning is on modeling the relationship between the input properties (also called input, descriptive, or independent attributes/features/variables) and output properties (also called output, target, response, or dependent attribute/variable). The objective is to train a model using a set of labeled examples to predict the target attribute value of new, previously unseen samples.

Based on the continuity of the target variable that has to be predicted, the learning task can be a classification task, where the property to be predicted is a category from a finite set of mutually exclusive categories (also called classes), or numerical prediction (also called regression) where the property to be predicted is a real value [34]. The classification approaches can be roughly grouped into three major groups: instance-based classifiers that use the observations directly and do not build models (e.g. k Nearest Neighbors) [38], classifiers based on generative models that build a generative statistical model (e.g. Naive Bayes) [39], and classifiers based on discriminative models that directly estimate a decision boundary (e. g. Decision Tree, Perceptron, Logistic Regression) [40].

Based on the number of target variables, the classification problem can be single-output/target classification with just one target variable or multi-output/target with more than one target variable. The single-output classification problems can be a binary classification or multi-class classification, while the multi-output classification problems can be multi-label classification [41] or multi-target classification [42]. The binary classification task is associated with a binary data set with only one target attribute that can take only two different values, usually known as positive and negative, but can also be interpreted as true and false, 1 and 0, or any combination of two values that are relevant for the application domain. In binary classification, the example does or does not have some property i.e. the presence or absence of a property has to be predicted. The multi-class classification task is associated with a multi-class data set with only one target attribute that can take any value from a finite set of predefined classes. In multi-class classification, the example can belong to only one of the multiple (equal or higher than three) possible classes. A number of multi-class classification algorithms rely on binarization, where a binary classifier is iteratively trained for each class against the others (One-vs-All (OVA) or One-vs-Rest (OVR) approach), or for each pair of classes (One-vs-One (OVO) approach). The multi-label classification task is associated with a multi-label data set that has multiple binary target variables. Here, a set of target attributes (also called labels) are predicted for each example. Finally, the multi-target classification task is associated with a multi-target data set that has multiple multi-class target variables. This task is concerned with predicting several targets, each of which can take a single value of several possible classes.

2.2.1 Multi-class classification

The goal of multi-class classification is to induce a classifier for a training set with input attributes and discrete target attributes. This is achieved with an entity referred to as an induction algorithm (also known as inducer or learner) that forms a model generalizing the relationship between the input attributes and target attribute. In prediction phase, the classifier outputs a class for new data samples. The multi-class data set has only one output attribute that can hold any of a set of predefined values. The meaning of each of the values and the value itself are specific to the particular application. The set of classes is finite and discrete.

2.2.1.1 Formal definition

The formal definition of a multi-class classification task is as follows. For a given:

- description space X, consisting of tuples of continuous values, $\forall X_i \in X, X_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$, where D is the size of the tuple, i.e., the number of descriptive variables,
- target vector y, representing the target variable, consisting of discrete value y_i representing the class, $\forall y_i \in y, y_i \in \{c_1 \cdots c_K\}$, where $K \geq 3$ is the number of classes,
- instance space I, set of all instances, where each instance is a pair of tuples from descriptive space and class value, respectively $I = (X_i, y_i) | X_i \in X, Y_i \in y, 1 \leq i \geq N$, where N is the number of instances in I,
- quality criterion q, which rewards models with high predictive accuracy and low complexity,

find a hypothesis $h : X \rightarrow y$ that maximizes q.

2.2.1.2 Multi-class classification methods

In this dissertation, four distinct learning methods are considered: tree-based methods, ensemble methods, kernel-based methods and neural networks. Implementations of Decision Tree[1], Random Forest[2], Support Vector Machine[3] and Multilayer Perceptron[4] from Scikit-learn Python library are used [43], [44]. It has been shown in the literature that these implementations produce reliable results in various applications [45].

Decision Tree Decision Tree (DT) is a non-parametric method that can handle both categorical and numerical data, as well as nonlinear relationships between features and the target variable. The algorithm constructs tree-like models with a hierarchical structure that can be used for both classification tasks (classification tree) and regression tasks (regression tree) [46]. The main advantage of these models is simplicity and comprehensibility, as the tree structure provides a clear representation of the decision-making process.

A classifier tree is obtained by the recursive partition of the instance space and consists of nodes (root node, intermediate nodes, and terminal/leaf nodes) and branches. The root node corresponds to the whole instance space, the intermediate nodes contain tests on input attribute with a branch for each outcome containing sub-space of the instance

[1] DecisionTreeClassifier
[2] RandomForestClassifier
[3] SVC
[4] MLPCLassifier

space, the leaf nodes contain only instances that belong to the same class, and each branch comprises the route from the root node to the leaf node.

Common algorithms for learning DTs are Iterative Dichotomiser 3 (ID3) [47], C4.5 [48], and Classification and Regression Trees (CART) [46]. The ID3 algorithm uses an iterative method based on subsets of the training data set to construct a tree. The test is made on a categorical attribute that maximizes the information gain. The C4.5 is an evolution of the ID3 algorithm with several improvements related to handling continuous attributes, attributes with missing values, and noisy data. The CART algorithm is very similar to C4.5 with the difference that it supports continuous target variables. It constructs binary trees by performing splits on attributes that result in the largest information gain at each node.

DT learning is based on a tree-construction algorithm known as Top-Down Induction of DTs (TDIDT) [34], which uses a divide and conquer approach for the greedy recursive growing of trees starting at the root node and proceeding down to the leaves. The growing of the tree stops when a stopping criteria is satisfied. An important element of the tree-growing process is the method to evaluate a certain split that will guide the learner to construct small trees with good predictive performance. Usually, a uni-variate splitting criterion is used, i.e., a node is split according to a value of a single attribute. Typically, the learner uses an impurity measure to find the best attribute upon which to perform the split. The main rationale behind this is to determine which attribute would split the data in a way that the under-laying child nodes are most homogeneous or pure. Common impurity-based criteria are information gain and gini index of diversity [34], [49]. The former uses entropy as an impurity measure, while the latter measures the divergence between the probability distributions of the target attribute values. Other, less-frequently used, uni-variate splitting criteria exist, such as orthogonal criterion [50], Kolmogorov–Smirnov [51], and mean posterior improvement [52].

Pruning of tree models is a common practice used to improve the predictive performance of the models on unseen data. Large and fully grown trees may have limited generalization capabilities. Pruning removes the knowledge captured by the tree that is very specific to the training data and protects against overfitting the tree. Pruning methods that are integrated into the growing process as stopping criteria are called pre-pruning methods. The methods that prune the tree after it is constructed are called post-pruning methods. In this dissertation, we use pre-pruning by specifying the maximum depth of the tree and a minimum number of samples per leaf as pruning parameters. Maximum depth is an integer value that is used to control the depth of the tree, i.e. the tree is grown until the maximum depth is reached. The minimum number of samples per leaf is an integer value that is used when considering tests at a specific node. The given test can be considered for further evaluation if it produces subsets with more instances than the specified value of this parameter. If there is no test that satisfies this condition, then the node is replaced with a leaf. New instances are classified by navigating them from the root of the tree down to a leaf node according to the tests along the path. When the instance reaches a leaf, it is classified into the leaf class.

The main advantages of DT are related to the non-parametric induction process (no requirements for prior assumptions about the probability distribution of variables in input and output space), the small computational requirements even on large data sets, the robustness of the learning process against noise and outliers, fast training and prediction, comprehensibility, and good predictive performance.

Random Forest Random Forest (RF) is a tree-based ensemble method [53]. The algorithm can be used for both classification and regression problems and can handle both

categorical and numerical data. The underlying intuition of the algorithm is to create a collection of DTs, each of which makes its own prediction. The final prediction is made by aggregating the individual predictions of each tree. The trees in the ensemble are different and each captures a different aspect of the data. Therefore, by combining the results of multiple trees, the forest of trees provides a more reliable and robust prediction than a single tree.

Ensemble methods are a category of ML techniques that involve constructing multiple independent predictive models and integrating their predictions into a single prediction. The individual predictive models are referred to as base predictive models, and the set of these models is referred to as an ensemble. The concept of ensemble can be applied to both classification and regression models. Roughly speaking, the basic hypothesis in ensemble learning is that the predictive performance of an ensemble is better than that of a single base model.

The most widely used methods for constructing ensembles are based on learning heterogeneous predictive models, manipulating the training set, or manipulating the learning algorithm. For instance, RF is constructed based on changes in the instance and feature space. This approach is convenient in combination with unstable predictive models, i.e., models that experience great challenges in their structure with small changes in the training set, such as DT. In general, the manipulation is performed by manipulating the training instances, the feature space, or both using various techniques such as bootstrapping, boosting, or importance sampling. With manipulation, many sub-spaces are created from the original instance or feature space and each of the base predictive models is learned using a different sub-space. Manipulation of the instance space is expected to perform well for data sets with a large number of instances, while manipulation with the feature space is expected to perform well with high-dimensional data sets having a small number of instances.

One of the most important aspects in ensemble learning is the scheme for combining the predictions of the base predictive models into a single prediction [54]. Two approaches exist to obtain a single prediction from an ensemble: model selection and model combination. The model selection approach first evaluates the performance of each base predictive model and the prediction of the ensemble is the prediction of the best-performing model. Although the final model is simple, understandable, and fast, the main drawback is that the predictive performance is limited by the performance of the best predictive model. The model combination approach overcomes this drawback by combining the prediction of all base predictive models in an overall final prediction of the ensemble. A common scheme for a combination of the base models is the voting scheme, with the majority vote and probability distribution vote schemes being the most widely used. The generalization performance of a forest of tree classifiers depends on the performance of the individual trees in the forest and the correlation between them.

RF combines bootstrapping with feature sub-space selection. Bootstrap creates several bootstrap replicas of the training data set by random selection with replacement [55]. Each tree in the ensemble is constructed using bootstrap sample of the training data set. Furthermore, during tree construction, at each node of the tree, it considers splitting on a different randomly selected subset of features.

In the implementation of the algorithm considered in the dissertation, the RF classifier is a meta-learner that uses DT base-learner to grow a tree on bootstrap sub-sample of the training data set [43]. The implemented algorithm is based on the perturb-and-combine technique. The random instance and feature sub-spaces are used to decrease the variance of the estimator. The randomness is used for constructing uncorrelated individual trees with decoupled prediction errors. The main parameters that should be adjusted by the user

are the number of trees in the forest and the number of features considered when splitting a node. Increasing the number of base learners improves the predictive performance of the ensemble model until some critical number of trees is reached. The values of these parameters affect the computation time and memory consumption. Additionally, parameters that control the constituent trees, especially their sizes, such as the maximum depth of the tree and the minimum number of samples per leaf, should be set. Here, the ensemble prediction is obtained by averaging the probabilistic predictions of the base models.

The main advantages of RF are its versatility in terms of learning tasks and data types it can handle, the ability to run the processes on multiple machines, good performance on high-dimensional data due to sub-sampling the feature space, low bias, and moderate variance. The disadvantages of RF are considerable computational cost with large data sets, risk of overfitting due to too many deep trees or inadequate pruning, bias toward the majority class in imbalanced data sets, and sensitivity to outliers and noisy data. In terms of comprehensibility, while RF is not a complete black box, it can be difficult to interpret the results.

The choice between DT and RF depends on the specific requirements of the problem. In general, the DT is a good choice in situations where simplicity and interpretability are paramount, while RF is a good choice when accuracy and robustness are paramount. The randomization process used in RF makes the algorithm less prone to overfitting and can improve the overall performance of the model. The DTs are easier to interpret than RF, because the tree structure shows the conditions and decisions that lead to a particular prediction, while the RF can be more difficult to interpret because it makes the final prediction by combining the predictions of many trees. The large number of trees used in RF makes this algorithm computationally intensive for large data sets. On the other hand, the DT is relatively fast for training, but if the constructed tree model is deep and has many branches, it can be slow in prediction phase.

Support Vector Machine Support Vector Machine (SVM) is a parametric, kernel-based method [56]. It can solve both linear and non-linear problems for classification and regression tasks. The main intuition behind the algorithm is to find the best boundary that separates the classes in the data in such a way that the margin between the classes is maximized, considering only the most influential data points. In the case of linear data, the classes are linearly separable in the low dimension, while in the case of non-linear data, there is not a linear boundary that separates the classes but the algorithm can be still used by further manipulating the data using kernel function. The kernel function is used for mapping the input vectors from the training data set in higher-dimensional space. The main idea is that in this higher-dimensional space, a linear boundary that separates the classes and ensures high generalization can be found. The separation boundary in the original input space corresponds to a non-linear decision function determined by the kernel. Consequently, the boundary in the original space can then be obtained by mapping the higher-dimensional boundary back to the original space. The use of kernels allows all necessary computations to be performed directly in the input space. The main advantage of this type of model is that they are complex enough for real-world applications and still simple enough to be analyzed mathematically.

The use of the kernel function is essential for transforming the input space into higher-dimensional space where the data may become linearly separable. There are several commonly used kernel functions in SVM, including the linear kernel [57], polynomial kernel [58], radial basis function (also known as Gaussian) kernel [59], and sigmoid kernel [60]. The choice of kernel depends on the nature of the data and the problem being solved. In general, the linear kernel is a good starting point and more complex kernels can be tried

if the linear kernel does not produce satisfactory results.

In the context of SVM, the decision boundaries used to separate the classes are referred to as hyper-planes that divide the space into disconnected parts. For separating the classes, there may exist many separation hyper-planes, so the optimal one has to be selected. The optimal hyper-plane is defined as the one with a maximal margin of separation between classes and is called the maximal-margin hyperplane. The maximal margin is the distance between the hyperplane and the closest samples in the data, called support vectors. Only the support vectors are relevant in defining the hyperplane. The margin ensures that the boundary between the classes is well-defined and robust. The hyperplane is learned from training data using an optimization procedure that maximizes the margin. For messy real-world data that cannot be separated perfectly with a hyperplane, the concept of soft margin is used. In such cases, tolerance for margin violations is introduced. The objective is to find a trade-off between keeping the margin as wide as possible and limiting the margin violations. The degree of tolerance for margin violations is controlled by a regularization parameter called C. A lower value of the parameter allows more margin violations but results in a wider margin, while a larger value of C results in fewer margin violations but also a narrower margin, and the goal is to find the right balance that will prevent overfitting. Learning an SVM model is achieved using optimization procedures. Numerical optimization procedures for searching the coefficients of the hyperplane are inefficient. Thus, procedures that reformulate the problem as a quadratic programming problem are rather used, such as the Sequential Minimal Optimization (SMO) algorithm [61].

One of the prominent advantages of the SVM is its scalability, which enables it to handle high-dimensional data effectively. Additionally, it possesses non-linear capabilities, allowing it to address non-linear dependencies between variables. The algorithm is also capable of handling imbalanced data sets, where one class has more samples than the other. Its consideration of only the most influential data points, referred to as support vectors, in the optimization process leads to robustness in the presence of irrelevant or noisy data. Furthermore, the algorithm demonstrates versatility through the ability to specify different kernel functions, both common and custom, thereby providing flexibility to the model. Despite its advantages, there are also some drawbacks associated with the algorithm. It can be computationally expensive and memory-intensive for large datasets. Additionally, SVM is highly sensitive to parameter selection, such as the regularization parameter selection, which balances the model's complexity and performance, and can significantly affect the algorithm's overall performance.

Multilayer Perceptron A perceptron is a single-neuron linear model that combines the given inputs and weights in a weighted sum and produces output if the sum exceeds a predefined threshold [62]. The neurons are also known as linear threshold units. The threshold in this case represents the activation function [63]. A perceptron can be trained by calculating the errors between predicted outputs and true outputs and adjusting the weights. The perceptron cannot be applied to non-linear data [64]. In order to tackle this limitation the MLP is developed [65].

Multilayer Perceptron (MLP) is a type of feed-forward neural network in which the mapping between inputs and outputs is non-linear. The main intuition behind MLP is to model complex relationships in data by combining multiple simple functions through the use of weights and biases, and optimizing the parameters of these functions through the training process so that the network can accurately predict outputs based on the input data. Essentially, the network can be viewed as a non-linear function approximator where the input data is transformed into an output prediction through a series of transformations in the hidden layers. The weights and biases in the network determine the strength of each

activation function and the relationships between the inputs and the outputs. The training process of an MLP involves adjusting the weights and biases. This is usually done using optimization and automatic differentiation in an iterative manner.

The MLP network consists of a large number of interconnected artificial neurons arranged in a layered architecture. Each artificial neuron within the network represents a computational unit that computes the weighted inputs and applies a non-linear activation function to produce an output. The first layer in the MLP network is referred to as the input or visible layer. This layer receives the input data and serves as an entry point for information into the network. Typically, the input layer has one neuron per input variable in the data set. The final layer in the MLP network is referred to as the output layer and is responsible for generating the network's predictions. The information within the MLP network flows in the forward direction, from the input layer to the output layer. The number of neurons in the output layer depends on the nature of the problem being modeled. There may be one or more hidden layers between the input and output layers. These layers process the data and make intermediate predictions that are critical to the overall functionality of the network. The number of hidden layers in the MLP network depends on the complexity of the task and can be increased arbitrarily. When the number of hidden layers is significant, the network is called a deep neural network [66]. The neurons within the MLP network are interconnected, and each connection is assigned a weight that dictates the importance of the relationship between the neurons.

The role of the weights associated with each connection and the activation function used in each neuron are crucial for the modeling capabilities of the network. The weights dictate the importance of a particular connection, while the activation function affects the non-linearity of the network and its ability to prevent under-fitting in the case of non-linear data. Some commonly used activation functions in MLP networks are the sigmoid function, the hyperbolic tangent function (Tanh), the reflected linear function (ReLU), the leaky-reflected linear function (Leaky ReLU), and the the Softmax function [63]. Each activation function has different properties, strengths, and weaknesses, and the choice of activation function depends on the specific problem being addressed. The sigmoid function maps values from a nearly infinite range to probabilities in the range between 0 and 1. The Tanh function is similar to the sigmoid function, but maps values in the range between -1 and 1. The ReLU function outputs zero for inputs below zero and follows a linear relationship as the input increases above zero. The Leaky ReLU function is a modification of the ReLU function that addresses the dying problem of the ReLU activation function by following a linear equation with a small gradient for negative input values, instead of setting them to zero. The Softmax function maps real values to probabilities in the range between 0 and 1 and is often used in the output layer for multi-class classification tasks.

The training process of an MLP network involves several phases aimed at finding the optimal values for the network's parameters in order to minimize the loss function and thus, reduce the prediction error of the model. The phases of the training process include feed-forward, cost computation, back-propagation, and weight update. In the feed-forward phase, the training data is passed through the network in order to generate predictions for each input example. The values of the network parameters are initially random, although other initialization techniques can be used. Once predictions have been generated, a loss function is used to measure the match of the predictions against the correct target. The loss function is then used to calculate the gradients of the loss with respect to the weights of the network through the back-propagation phase. Optimization algorithms, such as stochastic gradient descent, are then applied to update the weights of the interconnections using the calculated gradients. The steps of feed-forward, cost computation, back-propagation, and weight update are repeated multiple times until the lost function reaches a minimum and

the model has been sufficiently trained.

The training of a neural network involves the consideration of several hyper-parameters, including the network architecture, learning rate, and the number of epochs. The network architecture refers to the number of hidden layers, the number of neurons in each layer, and the type of activation function used, while the learning rate determines the step size of weight updates and the speed of learning. The number of epochs defines the number of times the training data is passed through the network. The selection of hyper-parameters before the training process is important for adjusting the algorithm so that the models train better and faster. The choice of hyper-parameters should be guided by prior knowledge of the problem and the data, as well as experimentation and comparison of different configurations. The ultimate goal is to select optimal hyper-parameters that result in an MLP network that generalizes well to new, unseen data while minimizing overfitting.

Neural networks offer several advantages in the analysis of data, such as the ability for universal approximation, the ability to handle non-linear relationships, proficiency in learning from large data sets, and stability in the presence of high-dimensional data. However, they also have some disadvantages, such as the risk of overfitting, high computational cost, difficulty in interpreting results and understanding the reasoning behind predictions, sensitivity to initialization, and the requirement for large amounts of data.

2.2.2 Multi-label classification

Multi-label classification is a learning task in which a predictive model is trained from a data set with multiple binary target variables. The model can predict the presence or absence of multiple labels simultaneously [41]. Multi-label classification differs from single-target classification in the number of outputs expected from the trained models. While the single-target classifier returns only one output value, the multi-label classifier returns multiple output values. Therefore, the multi-label classification task is more difficult because of the additional complexity in the output space. The multi-label classification can be viewed as a collection of several binary classification tasks and is different from the multi-target classification task [42] where several targets have to be predicted, each of which can take only one value of several possible classes.

The multi-label learning is associated with several specific aspects, such as dependencies between labels, high dimensionality in the output space, and label imbalance, that affect the learning complexity. The first aspect refers to the relation between the labels and how they can be used for designing better multi-label classifiers. The second aspect refers to the high number of dimensions in the output space as a result of having one output for each label. Finally, the imbalance is even more emphasized because when a large number of labels exist, it is usual that some labels are more frequent than others (for instance, some environment properties are more common than others) and also when the original multi-label data set is transformed, the imbalance is increased (due to taking all examples where one label appears as positives and all other examples as negatives).

2.2.2.1 Formal definition

Multi-label classification tasks can be formally defined as follows. For a given:

- description space X consisting of tuples of continuous values, $\forall X_i \in X$, $X_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$, where D is the size of the tuple, i.e., the number of descriptive variables,
- label space $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_Q\}$ which is a set of Q possible labels. $P(\Lambda)$ denotes the powerset of Λ, containing all possible combinations of labels $\lambda \in \Lambda$ including the

empty set and λ itself,

- target space Y consisting of tuples of binary values, $\forall Y_i \ \in Y, Y_i = (y_1, y_2, \cdots, y_Q)$ where each element is 1 if the label is relevant and 0 otherwise,
- set of instances I where each instance is a pair of tuples from the description and the target space, respectively $I = (X_i, Y_i) | X_i \in X, Y_i \in Y, 1 \leq i \geq N$, where N is the number of instances in I, and
- quality criterion q, which rewards models with high predictive performance and low complexity,

find a function h: $X \rightarrow Y$, such that h maximizes q.

Therefore, for a given input instance, the multi-label classifier will return a binary vector indicating the relevant labels.

2.2.2.2 Multi-label classification methods

The methods for learning from multi-label data can be grouped into: problem transformation methods and algorithm adaptation methods [67].

The first group of methods aims to solve the complex multi-label task by transforming the original multi-label classification problem into several simpler single-target classification sub-problems that can be solved using existing algorithms. The multi-label data set is transformed into several single-target data sets (one for each label in the target space). The data sets share the same descriptive space. These data sets are then approached with single-target binary classification methods and individual classifiers are built for each data set. At prediction phase, all binary classifiers are invoked and their individual predictions are joined to obtain the final set of labels. In this approach, complete independence between the labels is assumed. However, when the presence of one label could determine whether another label is likely to be present, the exploitation of the dependencies between labels may improve the performance of the learning method.

The second group of methods is based on adaptation techniques that aim to adapt the proven algorithms to tackle the multi-label classification problem. The adaptation complexity depends on the type of the original method and the way in which the several labels will be considered. Multi-label classification methods adapted from tree-based (Multi-label C4.5 [68], Multi-label Alternate DTs [69], Predictive Clustering Trees [70]), kernel-based (Twin Multi-Label Support Vector Machine [71]), instance-based (Multi-label k-Nearest Neighbor [72]), and neural networks (Multi-label Back-propagation [73], Multi-label Radial Basis Function Network [74]) are reported in the literature. For the experiments conducted in this dissertation, problem transformation approach is considered and the classification methods described in Section 2.2.1 are used.

Chapter 3

Related Work

In this chapter, we present a detailed review of the state-of-the-art literature in the key research areas related to our work: convergence of sensing and communication, characterization of indoor environments, identification of indoor environment and electromagnetic characterization of materials; and identify gaps in the literature on environmental awareness.

3.1 Convergence of Sensing and Wireless Communication

Among the many visionary projections about next-generation networks (e.g., beyond 5G and 6G), a significant paradigm shift is anticipated to support ubiquitous sensing, connectivity, and intelligence [75], [76]. Radio sensing, which uses the communications system as a sensor that exploits the propagation of radio waves (transmission, reflection, and scattering) to obtain information about the surrounding, has the potential to become an essential component of the solution for creating authentic digital twin representations of the physical world [77]. In addition, the sensing capability can be combined with artificial intelligence to enable a fusion of the physical environment with its digital counterpart [78].

To meet all the sensing needs in the next-generation communications era, solutions must be developed that combine radio sensing with non-radio sensing by other sensors, such as those ubiquitously used in devices like accelerometers, gyroscopes, and cameras, to create multi-layered maps of the environment. This has attracted the research interest from academia and the wireless industry in the convergence of communications and sensing [79]. The concept has been explored for point-to-point communications used primarily in vehicular networks with applications for autonomous driving [80] and cellular networks [81]. Introducing sensing functionality into the communication network provides an additional dimension to its capabilities, allowing it to evolve into a perceptual network that provides sensing along with uncompromised communication [82], [83]. Large-scale sensing is important to industry and society, enabling IoT applications and a range of innovative initiatives such as smart city and smart transportation [84].

Although the two operations of sensing and communication are based on similar physical phenomena, both were developed in parallel for decades. Communication systems aimed to accurately convey information to a receiver [85], while radar systems aimed to sense (i.e., obtain information about) a target [86]. Initially motivated by the similarities between the two systems in terms of underlying phenomena, hardware, signal processing, and working bandwidth, and later by the advances in the next-generation networks, the introduction of meta-materials [87], massive reconfigurable smart antennas [88], [89], reconfigurable surfaces [90], and high carrier frequencies [91], to name a few, the research focus shifted over the years to different levels of integration of the two functionalities. Various

research communities have discussed the idea under different names, e.g., radar communications [92], joint communications and radar [93], joint radar and communications [94], [95], and dual-functional radar-communications [96], [97].

The integration of communication and sensing functionalities is a paradigm shift in which the two operations are no longer viewed as separate end goals, but are optimized together in a spectral, energy, and cost-efficient manner for mutual benefit. The systems communicate and sense over a common waveform, a single hardware platform, a signal processing framework, and a network infrastructure. This addresses the problem of spectrum congestion while reducing the hardware and signaling costs, referred to as integration gain. By exploiting the ability to design the two functions together, communication-assisted sensing and sensing-assisted communication are enabled. This can significantly improve sensing and communication performance, which is referred to as coordination gain. The ultimate degree of integration of both functionalities will be driven by the limited spectrum, which is a scarce resource to be dedicated just for a single functionality, a dense network with ubiquitous infrastructure over a large area, wide signal bandwidth that provides high-resolution sensing capabilities, massive multiple-input multiple-output (MIMO), and spatial processing techniques [98].

Based on the degree of integration between sensing and communication systems, two categories of approaches can be distinguished: coexistence and joint (integrated) design [99], [100]. The coexistence approaches consider separate systems for sensing and communication that are mutually interfering [101]. Depending on the amount of information shared between the systems, coexistence can be cooperative [102], [103] or non-cooperative [104]–[106]. Cooperative coexistence can be considered as the initial stage in the development of jointly designed systems, where the cooperative performance depends on the characteristics of both systems. Several coexistence schemes have been proposed in the literature, the most common of which are coexistence with spectral overlap and coexistence by transmitting separate signals in time, frequency, or code. The main drawback of the coexistence with spectral overlap is the strong mutual interference and the need for interference mitigation techniques to ensure satisfactory performance for both functions, while the main drawback of the coexistence by transmitting separate signals is the low spectral efficiency. The systems can negotiate for resources using cognitive techniques [107]. However, the complexity of finding gaps in the spectrum and the non-guaranteed performance of the secondary system are drawbacks of the coexistence scheme. A more recent architecture is functional coexistence based on the information embedding strategy using waveform diversity [108], phase modulation [109], side lobe amplitude modulation [110], and multi-waveform amplitude shift keying [111]. The joint design category includes approaches in which the two systems are designed together so that they behave as a unified system [112]. The main advantages of these approaches are high spectral efficiency, simultaneous operation without interference, and mutual benefits from information sharing.

For several years, a great research effort has been undertaken to investigate various aspects of coexistence, focusing on several main themes: (i) underlying operational method—investigation of different underlying schemes, the most common of which are waveform design [113], software-defined networks [114], and reinforcement learning [115]; (ii) mutual impact—analysis of the effects of radar systems on communication performance [116]–[118] and the effects of communications systems on radar performance [119], [120]; (iii) performance—estimation of the performance of the coexisting system [121]; and (iv) mutual interference—development of techniques to cope with and suppress mutual interference [106], based on estimation of the interference channel using classical methods such as least squares estimation and minimum mean squared error, coordination with a dedicated control center connected to both systems [122], and use of pilot signals [123].

In the area of cooperatively coexisting systems, most of the research focused on the analysis of performance tradeoffs [99], [124] and methods to achieve cooperation using various approaches such as beam control [111], waveform diversity [108], frequency diverse array [125], and modulation techniques [126], [127].

Based on the design approach, the jointly designed systems can be divided into the following three categories: (i) radar-centric systems, where the communication functionality is integrated into a primary radar system [128], [129]; (ii) communication-centric systems, where the sensing functionality is integrated into a primary communication system [76], [130]; and (iii) jointly designed systems without an underlying system [131], [132]. Research in radar-centric systems is mainly concerned with embedding information in pulsed or continuous radar signals and, more recently, in novel radar waveforms, e.g., MIMO orthogonal frequency-division multiplexing (OFDM) [133] and frequency-hopping radar [134], with a signal format similar to modern communication systems, usually with index modulation that does not change the basic radar waveform and signal structure [135]. Other important areas not covered in depth in the literature include the development of communication protocols for radar-centric systems and receiver signal processing techniques to extract the embedded information from the signal. The literature on communication-centric systems mainly reports studies on the application in vehicular networks [136], wireless local area networks [137], and large cellular networks [82]. The fundamental issues concern full-duplex operation in a monostatic setup [138] and mitigating the effects of clock asynchronization between spatially separated transmitters and receivers [139]. Preliminary research on joint design of communication and sensing systems without constraints on existing communication or sensing systems addresses high-frequency systems, particularly millimeter and terahertz waves, which provide both high data rate communication and high accuracy sensing [77], [140]. Initial studies focused primarily on the feasibility and potential of beamforming [141] and uniform waveform design [142].

Although the vision and benefits of integrating the communication and sensing functionalities are clearly laid out in many preliminary studies that are paving the way to a true digital twin of the physical environment, there are still many open challenges related to the theoretical and technical foundations, the interplay with other emerging technologies, and practical implementation [143].

3.2 Characterization of Indoor Environments

Knowing indoor scenes with high fidelity is highly important in architecture, engineering, and construction domains. A precise description of the buildings is needed for spatial understanding, automation, localization, and positioning of robots and mobile terminals.

Research in the field of indoor characterization has been active for decades and is largely based on advances in various fields, such as sensor technologies [144], photogrammetry [145], laser scanning [146], computer vision [147], robotics [148], massive antennas [149], [150], radar [151], to name a few. For many years, the focus of indoor mapping has been on describing the geometry of the indoor structure. The geometry can be represented in a variety of formats depending on the level of detail: 2-D building drawings [152], [153], 3-D models [154], point clouds [155], [156], meshes [157], [158], and graphs [159]. Indoor characterization methods have evolved over time in response to the development of new technologies. Initially, manual methods are used to create 2-D drawings and 3-D models [160], [161]. With the development of digital tools, models are created using specialized software and stored in an electronic format [162]. Later, an environmental reconstruction based on light or radio technologies emerges as a suitable method for characterizing existing buildings [155], [163]. In addition, the literature reports a combination of indoor scene

estimation and localization, an approach known as simultaneous localization and mapping (SLAM) [164], [165]. The simultaneous creation of an environment model and localization is particularly associated with the use of robots. Traditionally, SLAM involves simultaneously estimating the state of a robot equipped with on-board sensors, and building a model of the environment sensed by the sensors. Recently, wireless sensing has been introduced as a method to characterize a remote object by analyzing the received signal [166], [167].

For many years, construction drawings, a term that generally refers to any floor plan in the blueprint, whiteprint, or digital drawing file format, have served as the primary means of describing building design [161]. Building plans are generally part of construction documentation and are prepared during the design phase of a construction project to document and visually represent the building design, including the layout, dimensions, and structural components. While construction drawings have been suitable for communicating information between engineers, architects, and other parties involved in construction, they are unsuitable for emerging applications in various fields that rely on accurate and precise descriptions of the building [168]. This is due to their 2-D format, which cannot represent the full complexity of indoor spaces, lack of information about actual conditions in the physical building, e.g., materials used in construction, presence of obstacles and hazards, potential inaccuracies due to changes made during construction or modifications made to the building over time, and limited accessibility and updatability.

With the development of digital technologies and tools, 3-D models have become common, usually created manually using computer-aided design software (CAD) [169]. The 3-D models have several advantages over conventional drawings, including better visual representation and greater accuracy. In the construction industry, building information modeling (BIM) is used to create comprehensive digital models that are used throughout the life cycle of a building, from design to construction, operation, and maintenance. These models are created using software that enables the creation of highly detailed and accurate building models that contain information about the architectural, structural, and mechanical designs, as well as cost estimates.

Manual modeling is typically used for new buildings during the design and construction phases. Because of the slow, lengthy, and costly creation process, creating manual models for existing buildings is a challenging task, especially for large or complex buildings. The description of existing buildings is often captured by manually surveying the site. This involves collecting and documenting data about the physical features of the building, such as its size, shape, and location, which usually involves sending a team of surveyors to the site to conduct a physical inspection of the building. Manual surveying is very time-consuming, labor-intensive, and of limited accuracy. Therefore, many studies have focused on an automated and efficient approach for reconstructing the environment based on environment scanning with various technologies [170], [171]. The conventional reconstruction process consists of two main steps: environment sensing and environment identification. In the first step, indoor data is acquired using various environment scanning methods, while in the second step, the raw data from the acquisition step is processed to obtain a reconstructed scene of the indoor geometry. In addition, topological, e.g., adjacency, connectivity, containment, and semantic, e.g., type, function, material, information can be added to the reconstructed features [172].

The methods of environment acquisition based on the underlying wave can be divided into two main groups: light-based and radio-based methods [146], [173]. The light-based methods have been considered for a long time because of the laser properties such as narrow beam, stability, longevity, etc. The main problem with these methods is scanning under variable light conditions, which is often the case indoors. On the other hand, radio-based methods can be used in poor visibility conditions [174]. The studies in the field of

light-based environment scanning mainly investigate the use of range sensors such as Light Detection and Ranging (LiDAR) [175] and infrared lasers [176] as well as photographic sensors such as read-green-blue (RGB) cameras combined with a depth sensor [177]. LiDAR is a remote sensing method that uses light in the form of a pulsed laser to measure the distance to the target. It is used to create high-resolution maps of indoor environments. Laser scanners use infrared light to estimate the distance to an object based on the time-of-flight principle. Infrared scanners supplemented with software tools for processing points are commercially available [1]. RGB depth sensors combine RGB colour information with depth information per pixel. Depth per pixel is determined using an infrared camera with time-of-flight measurement. The low-cost Microsoft Kinect camera is commonly used to capture point clouds in research projects [178].

Much of the research in indoor mapping considered SLAM [179]. Currently, the focus of the fundamental research is on the robot/environment/performance combination, since the available SLAM algorithms can easily fail if either the robot's motion or the environment is too challenging [180]. The progress in SLAM is related to the availability of new sensors such as 3-D LIDAR, depth, light-field, and event-based cameras. Recently, the consideration of high frequencies in the millimeter-wave and terahertz bands, and the integration of a large number of directional antennas with an ability to electronically steer narrow beams has led to active research in radio scanning [181], [182]. The characteristics of the millimeter waves, such as the wide bandwidths, the quasi-optical propagation pattern with predominant LOS component, and the sparse angular spectrum that allows separation of MPCs, provide the technical basis for an accurate sensing [149], [183]. Early studies are paving the way for terahertz sensing to potentially become a central topic in wireless communications and sensing in the coming years [184].

The broad term wireless sensing refers to various methods of gathering information about distant objects by analyzing their effects on propagating EM waves. In most studies, wireless sensing is investigated in terms of using existing wireless links [185]. The low-cost and non-intrusive sensing enabled with this approach has attracted much attention, and various solutions based on different wireless technologies with versatile application scenarios have been proposed [94]. Most of the solutions are based on physical properties such as received signal strength indicator (RSSI), CSI, frequency shift for frequency modulated carrier wave (FMCW), and Doppler shift. These studies examine the effects of the propagation environment on signal strength, channel conditions or wireless link characteristics, and frequency shift. The widespread use of unlicensed wireless local area network (WLAN) technologies and wireless sensor networks in the IoT paradigm has led to a growing number of publications addressing theoretical foundations, methods, and techniques, use of off-the-shelf and custom hardware, limitations, and more [185]. As the propagation data generated by the dense wireless networks becomes high-dimensional and complex, a need arises for intelligent sensing that takes advantage of ML techniques [186]. In the literature, intelligent sensing is highlighted as one of the central topics in 6G networks with artificial intelligence; however, this topic is not yet mature [187].

The development of other fields that use spatial information has shifted the research focus to the development of methods to capture different aspects of space. The active research resulted in several research areas that address different topics related to environmental characterization, including (i) building content retrieval systems—to capture the objects that make up the environment, including furniture, fixtures, and other relevant objects [188]; (ii) human detection systems—to detect the presence and position of people [189], [190]; (iii) topological modeling—to capture the topological relationships between elements in space [191]; and (iv) semantic modeling—to semantically label elements with relevant attributes [192].

A considerable amount of research has already been published on conventional building reconstruction techniques and indoor mapping. However, recent technological advances have created the opportunity to explore unconventional approaches, particularly those that are data-driven and intelligent, that can provide a comprehensive description of indoor structures, and that can enable the combined use of different sensing methods. The next decade is likely to see a surge in research into intelligent and efficient sensing techniques that do not require human intervention, are inexpensive, and do not require specialized equipment. These methods are expected to gain increasing attention due to their convenience in terms of effortless sensing. While ML offers a wide range of techniques that have proven successful in a variety of research areas, the workflow for intelligent detection of indoor environments has yet to be determined. In addition, a taxonomy of the types of indoor structures in future smart buildings may be important for developing optimal indoor sensing workflows. Developing accurate methods for detailed characterization of the environment will pave the way for creating a digital twin of the physical environment [193] that encompasses all aspects of the environment relevant to the emerging use cases.

3.3 Identification of Indoor Environment

The identification of indoor environments using radio measurements is a new area of research on which only a few studies have been conducted [194], [195]. However, research interest in this topic is increasing due to the need for accurate indoor environment information obtained using a fast, efficient, and data-driven approach. The great success of ML in various fields additionally encourages studies that propose methods that use the techniques of ML to build models that can identify the environment with high accuracy. Several initial publications on this topic focused on the identification of the indoor environment category defined in terms of obstacles in the room. The identification was studied in the context of classifying different environment classes using ML models [196]. In the published studies, the environment classes represent the indoor structure as a whole. Although this approach may be practical for some application scenarios, it is limited to predefined environment classes. Approaches to identifying the properties of the facets that compose the space as well as the objects within the space are still lacking.

Several publications are documenting the identification of a few environment classes that differ from each other in the level of clutter [194]–[196]. In the aforementioned studies, the authors investigate the use of several properties of the received signal and ML algorithms for classifying the environment in the WiFi band associated with IEEE 802.11g standard from Vector Network Analyser (VNA) measurements data obtained in campus building under stationary channel conditions. In the experimentation, the use of RSSI, Channel Transfer Function (CTF), and Frequency Coherence Function (FCF) for feature engineering as well as DT, SVM, k-Nearest Neighbor (k-NN), and convolutional neural networks (CNN) as algorithms for constructing predictive models were reported. The main motivation behind the particular research was the development of a time-efficient solution that is suitable for real-time deployment scenarios in the Internet of Things.

The earliest work published on this topic investigates the suitability of CTF and FCF as a signature for environment classification considering the spatial correlation coefficient [194] and evaluates the proposed approach for only two environments: a highly cluttered laboratory, and a sport hall using the SVM algorithm with different kernel functions. Later, the authors extend their work by investigating different combinations of RSSI, CTF, and FCF as signatures, additional algorithms (DTs and k-NN), and a wider variety of indoor environments: highly cluttered—laboratory, moderately cluttered—narrow corridor, low cluttered—lobby, and open space—sports hall [196]. Their results show that a combina-

tion of CTF and FCF is suitable to achieve the satisfactory performance of the learned classifiers. Finally, in their latest work [195], they explore the power of deep learning for extracting features from measured data instead of manually selecting signatures and build on previous findings to develop an indoor localization approach that takes advantage of environmental classification.

3.4 Electromagnetic Characterization of Materials

The necessity for proper indoor planning is confirmed with the studies showing that most of the wireless connections are performed indoors [197]. Detailed EM characterization of the materials in a particular indoor structure is one of the crucial requirements for planning and deploying an indoor radio system. Since all conventional building materials are non-magnetic and non-ionised, only the dielectric properties of the materials have to be estimated [198]. Permittivity characterization of materials is extensively studied in the literature and the proposed methods for obtaining the permittivity of a material can be divided into two groups: permittivity characterization of material samples and permittivity characterization of materials inside buildings [199], [200].

The working principle for permittivity characterization of a given material sample includes: (i) isolated measurements with measurement systems that include network analyzers (to display the reflection and transmission measurement response as a function of frequency), impedance analyzers, and inductance-capacitance-resistance (LCR) metres for measuring material properties at lower frequencies, (ii) measurement fixtures suitable for the chosen measurement technique and the physical properties of the material to ensure that the electromagnetic field is applied in a predictable manner, and (iii) specialized software to convert the measurement data from the instrument to appropriate terminology and format. Depending on the measurement principle described in the literature, the available methods can be divided into five groups: (i) coaxial probe method—best suited for liquids and semi-solids; (ii) transmission line method—best suited for low-loss magnetic and anisotropic materials; (iii) free space method—best suited for large and flat samples; (iv) resonant cavity method—best suited for small samples with low loss; and (v) parallel plate method—best suited for thin, flat sheets [199]. The methods stated previously are not specific to building materials, but each of them is limited to certain types of materials and has advantages in terms of accuracy, cost of equipment, the frequency range of measurements, etc. Of all the methods mentioned, the free-space methods, in which a material plate is placed between the transmitting and receiving antennas, are the most suitable for characterising conventional building material samples [201]. System calibration, minimization of spurious reflections from other materials, and multiple measurements are required to obtain good results with this method.

Isolated measurements of material samples are inappropriate for determining the relative permittivity of materials used for indoor structures because of the cost of the methods (use of high-quality equipment and other costs associated with laboratory measurements) and the properties of the materials incorporated into indoor structures (which may differ from the isolated material samples depending on the size and type of structure). Several conventional studies have made on-site measurements of the reflection coefficient and estimated the relative permittivity of the materials based on the reflection loss [202]–[204]. In recent work, the inverse reflection problem was combined with the identification of the reflecting surfaces from a point cloud to obtain a 3-D permittivity map of an empty office environment at 60 GHz [200].

The critical analysis of the most relevant publications in the literature on EM characterization of materials shows that the reported methods focus on characterization of building

material samples with specialized equipment and on-site measurements. The studies found in the literature focus mainly on frequency bands with practical relevance in the period when the experimental studies were performed. The methods used to characterize material samples are not suitable for EM characterization of indoor propagation environments in the next generation scenarios. Various approaches to on-site estimation of EM properties of materials have been described in detail in the literature. In general, the methods are based on procedures for measuring the reflection loss.

3.5 Environmental Awareness Research Gap

Wireless communications in future smart buildings can take advantage of environmental awareness to dynamically adapt and meet the various demands that the wide range of applications may pose. Accurate, detailed maps of buildings are one of the essential requirements for achieving state-of the-art performance in future indoor wireless communication networks. In the context of wireless communications, environmental awareness is at the propagation level and provides insight into how the environment affects radio propagation. Awareness of the RE properties, the site geometry, the surface roughness, and the EM properties of the materials, serves as foundation for an improved radio channel prediction and plays important role in the development of the future intelligent wireless networks. The EM properties of the materials used for the surfaces that bound the space are essential because they directly determine the propagation mechanisms and the extent to which each of the mechanisms occurs. The state-of-the-art literature on building characterization lacks parsimonious methodologies, simple and elegant, yet accurate, for identifying materials in a given room. This underscores the need for a new methodology that provides comprehensive and accurate materials characterization of indoor structures.

An accurate model of the physical structure in the form of a multi-layer map, where each layer describes a different aspect of the building, is of utmost significance, primarily to enable a more holistic understanding of the space, but also to enable the development of solutions in different areas that expand the building's ability to operate more efficiently, flexibly, interactively, sustainably, and improve occupants comfort. In the future, the creation of these maps will require a combination of advanced sensing methods and technologies capable of capturing various aspects of the structure and making efficient use of time and resources. For improved indoor wireless communications, the map of the building should include a thorough and accurate description of the propagation characteristics of the site. This level of environmental awareness will be critical to future advances in indoor environment-aware communications, as it will provide the necessary ground for developing unprecedented approaches to radio channel prediction beyond what can be achieved with current resources (environmental information) and will enable next-generation wireless networks to adapt their configuration to the specific environment.

We propose to use the radio communication network and the methods for a comprehensive characterization of the propagation level properties of an inner building structure. Future networks using heterogeneous wireless technologies will be deployed everywhere in buildings. Therefore, next-generation networks will represent a data source that contains a signature of propagation environment characteristics left on signals from a wide range of typical environments in residential, office, and commercial setups, at various frequency carriers traveling between radio nodes at different locations relative to the bounding structure. The availability of large amounts of versatile data in terms of the communication system setup and propagation environment properties opens up the possibility of using ML techniques to discover knowledge about the materials in the surrounding of a wireless link with the use of ML techniques. In essence, the large number of connected radio de-

vices forms the basis for collecting huge data sets, which are essential for learning models capable of making accurate predictions, with generalization capabilities to unseen data and robustness to noise in the data due to outliers and random fluctuations. When a large database of sensing data is available, the ML can automate the process of data analysis and find the relationship between the radio environment signatures in the received signal and the materials in the radio environment. The ML-based methodology reduces the need for manual intervention and increases efficiency. In addition, it can scale up to large datasets containing data from numerous indoor setups, allowing the algorithms to learn complex relationships between the material and the propagation, which would otherwise be difficult. Finally, the use of ML techniques is well-suited to dynamic and changing scenarios, such as indoor environments, as the ML models can adapt and improve over time as they are exposed to new data. By using wireless technologies and ML together, the complementary strengths of both, i.e., the ability of the wireless network to sense and the ability of the ML algorithms to learn from sensing data can be leveraged for improved environmental awareness.

Chapter 4

Methodology for Intelligent CSI-Based Characterization of Indoor Radio Environment

In the dissertation, we propose a novel methodology for the characterization of indoor RE based on the RE signature embedded in the received signal using ML. In order to tackle the indoor RE characterization problem, we propose to formalize the problem as a supervised learning task. A predictive model is developed to make predictions about the RE properties based on the input RE signature represented by the propagation characteristics of a wireless link. RE characterization based on RE signatures using ML models is a novel approach presented in this dissertation and, to the best of our knowledge, it has not been studied, published, or patented up to now.

4.1 Concept

We propose to use a data-driven methodology to address the indoor RE characterization problem. It is formalized as a supervised learning task, where the algorithms learn a model based on the relationship between the distortion of the received signal and the RE. The adoption of ML automates the environment characterization process, leading to a reduction in the time required to characterize a particular RE. In contrast to existing approaches, the proposed methodology obviates the necessity for dedicated equipment, by leveraging the existing indoor radio infrastructure as a source of data for learning a model. A crucial aspect of the methodology is the incorporation of knowledge and experience in the field of indoor radio propagation at all stages, to learn accurate models that are relevant to the problem being addressed.

The proposed concept relies on two assumptions: (i) inside buildings, the received signal conveys a RE signature, and (ii) the CIR can be accurately estimated for each wireless link. The environmental information is in the form of environment-specific distortion of the received signal. This information is included in the received signal as a consequence of the interaction of the transmitted waves and the surrounding, resulting in a discernible signature left on the propagating waves. The CIR is used for the calculation of other propagation characteristics that represent the multipath effect of the propagation environment and the RE signature.

A corpus of diverse RE signatures structured in format appropriate for ML and annotated with environmental information is used as training data. The RE signature can be associated with any property of the RE that affects the radio propagation, such as the

shape of the room, size, material of the facets, and surface roughness. A suitable state-of-the-art algorithm uses the data to (i) analyse the complex relationships between the RE signatures and the properties of the indoor RE and (ii) learn a model capable of making predictions about the RE properties when a new RE signature is input. This dissertation focuses on RE properties with discrete values. The identification of RE properties inside buildings with a ready-to-use ML model that has been trained, evaluated, and optimized for the specific application is illustrated in Figure 4.1.

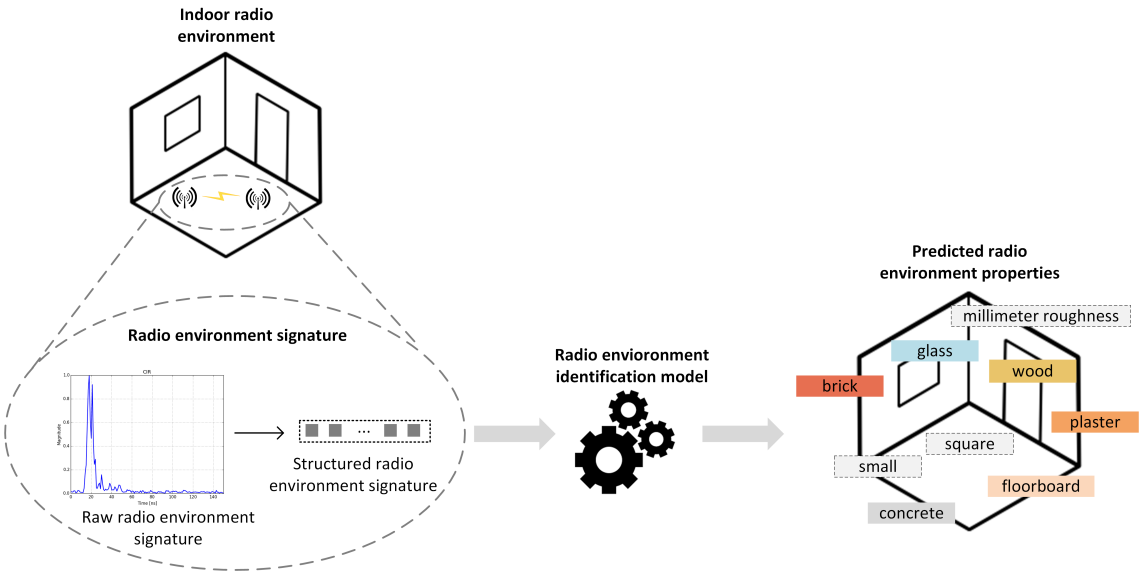

Figure 4.1: Identification of indoor RE properties. The illustration shows how a ready-to-use model predicts the RE properties when fed with RE signature. This illustration serves to clarify the concept of RE identification using the ML model.

4.2 Framework

To streamline the procedure for characterization of indoor RE based on RE signature we developed a framework. A schematic representation of the framework architecture is presented in Figure 4.2. The framework enables the construction of RE identification model using expertise from the wireless communication area and approaches from wireless communications and ML. The radio identification model outputs a subset from a set of predefined RE properties for an input RE signature example. The framework has four components, i.e., modules:

- Domain knowledge,
- RE acquisition,
- Propagation characteristic processing and storing,
- ML-based modeling.

The domain knowledge has a crucial role in all stages since it provides expertise to ensure that the real-world problem of indoor RE characterization is well defined and aligned with the needs for improved environmental awareness, the workflow is effective, and the resulting model is addressing the indoor environment characterization problem. First, a

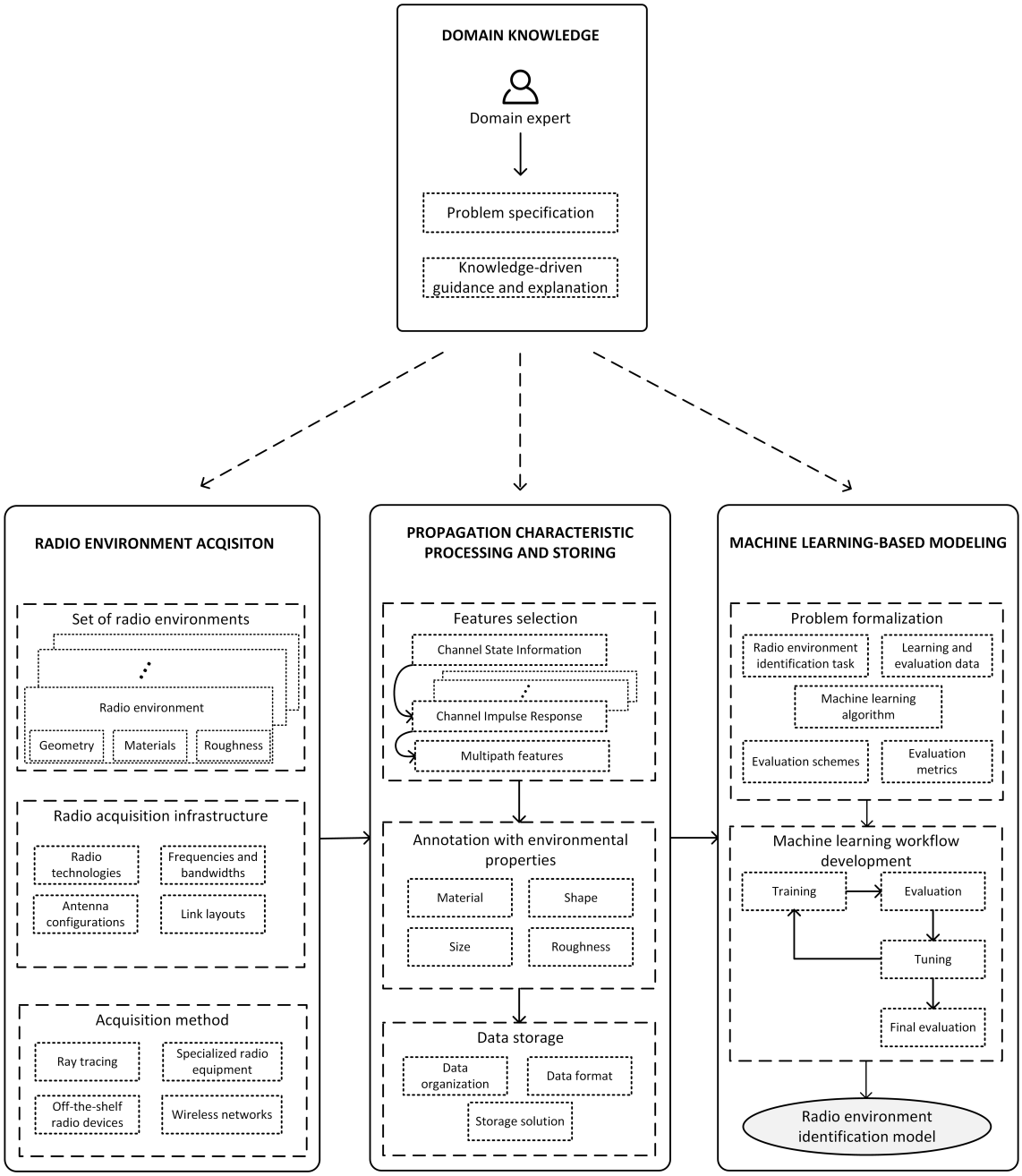

Figure 4.2: Schematic diagram of the framework for the construction of RE identification model.

large amount of representative and diverse radio propagation data is acquired for training and evaluating the models. Next, the raw propagation data is processed to a structured form that is suitable for ML analysis, annotated with environmental properties, and stored in a structured and organized manner. Finally, the indoor environment characterization problem is formalized as a ML problem, a suitable ML workflow is selected, and a predictive model capable to predict the properties of the RE for a given RE signature is constructed.

4.2.1 Domain knowledge

This module is responsible for incorporating domain knowledge provided by radio communication experts in the methodology for indoor RE characterization for ensuring that the developed model is relevant to the problem and useful for the intended application, i.e., RE characterization. Specifically, the domain knowledge is used for learning problem specification, guidance in various stages of the methodology, and results interpretation.

Important aspects of this module are the source of the knowledge and its points of use. The source of the knowledge are experts with specialized knowledge in the field of radio propagation and channels who have an in-depth understanding of how (i) the radio signals behave in indoor environments, (ii) the physical characteristics of indoor environments can affect and attenuate the radio signal propagation due to absorption, reflection, diffraction, and scattering, (iii) the radio waves interact with the obstacles, and (iv) the communication systems have to be configured to be able to estimate the multipath propagation characteristic of a wireless link.

The expertise is used for providing directions about the acquisition of propagation characteristics, such as which (i) REs to be included (in terms of geometry, roughness, irregularities, building materials, and obstacles), (ii) configurations of the radio communication systems to be considered, and (iii) methods for estimating the propagation characteristics to be used. Additionally, it is used for selecting meaningful features from the raw propagation data, while preserving most of the environmental information originally embedded in the received signal. Finally, it is used for providing a deeper understanding of the factors that may drive the model outcomes.

4.2.2 Radio environment acquisition

The purpose of the module is to acquire a large amount of indoor radio propagation data that is representative and diverse in terms of propagation environment characteristics and wireless system configuration.

This module is of pivotal significance for the success of data-driven RE characterization, since it provides data for learning and evaluating ML models. The quality of the propagation data acquired is a major limiting factor for the capabilities of the learned models. The data should be representative and diverse enough in terms of the properties of the RE and configuration of the communication system used to capture the propagation characteristics and to cover the various scenarios that the model is likely to encounter in a real-world use cases. Such data allows the model to capture enough variability in the correlations between the propagation characteristics and propagation environment without overfitting. The module includes three steps: (i) selecting REs, (ii) specifying acquisition infrastructure, and (iii) performing an RE acquisition campaign.

In the first step, a set of indoor scenes in residential and office environments is selected for RE signature acquisition. The set includes rooms with different levels of complexity, geometry, materials, and roughness. In the selection process, the real-life construction practices are considered. It is important that the set is large enough in order to provide sufficient experience for the ML algorithms.

In the next step, a set of system setups used for obtaining the propagation characteristics is selected. The radio system setup includes radio technologies, carrier frequencies and bandwidths, antenna configurations, and link layouts. The setup enables estimation of the propagation characteristics of the MPCs, therefore wide-band technologies should be considered. For instance, ultra wideband (UWB) communication systems are able to estimate multipath channel power delay distribution. The link layout is selected in a way that allows the multipath effect of the environment to be captured in a wide range of posi-

tions. Thus, we propose placing the radios on a grid with uniform-length cells covering the whole room. The distance between the radio nodes is selected according to the wavelength of the radio wave, to ensure that the small-scale variations are captured.

In the final step, an acquisition campaign is performed for estimating the propagation characteristics of spatially distributed radio links in the pre-specified set of REs using the pre-specified setups for the radio acquisition infrastructure. Synthetic and experimental methods, such as ray tracing, specialized radio equipment, off-the-shelf radio devices, and wireless networks, can be used to estimate the propagation characteristics of the wireless link. The methods can be used individually, or several methods can be combined.

The output of the module is a large corpus of propagation characteristics, i.e. spatially distributed CIRs from various indoor REs acquired with various communication system configurations. The raw propagation data is an input for the propagation characteristic processing and storing module.

4.2.3 Propagation characteristic processing and storing

The purpose of this module is to process and annotate the raw propagation characteristics provided by the RE acquisition module and to provide structured, labeled data sets that are suitable as input for the ML algorithms. The module includes three steps: (i) selecting features from the raw propagation data, (ii) annotating the formatted RE signatures with environmental properties, and (iii) storing the data.

In the first step, relevant features are identified from the raw propagation characteristics based on domain knowledge. The propagation characteristics of the MPCs in the CIR are considered as features. The number of descriptive variables and the input features of the i-th data point is given in Eq. 4.1 and Eq. 4.2, respectively. In particular, three propagation characteristics, (i) the received power P, (ii) the phase shift Φ, and (ii) the excess delay τ of the R MPCs reaching the receiver are considered as input features of the data point.

The number of descriptive variables is

$$D = kR, \tag{4.1}$$

where k is the number of propagation characteristics per MPC. In this study, k is three. The input features of the i-th data point can be expressed as

$$X_i = (P_{i,1}, \Phi_{i,1}, \tau_{i,1}, \cdots, P_{i,R}, \Phi_{i,R}, \tau_{i,R}). \tag{4.2}$$

The data point is considered to be a formatted representation of the signature that the RE has left on the particular radio link. Looking only at a subset of R informative MPCs that convey most of the energy, and thus, hold most of the environmental information, leads to reduced dimensionality of the descriptive space. Consequently, the time and resources needed for training are reduced.

In the next step, the formatted RE signatures are associated with categorical environmental properties (material, shape, size, and roughness). The annotation is a crucial step as it provides the ground truth, i.e., correct output for the input RE signatures data used for training a classification model that learns the mapping between the RE signature and the RE properties, and a benchmark for evaluation of the model's performance. In order to represent the output space as a matrix, each column in the output corresponds to a label, i.e., environment property. Practically, after the annotation, the input feature vector is associated with a binary indication vector that indicates the presence or absence of a particular label in the surrounding of the wireless link. For instance, when the materials used for the surfaces that bound the space around the radio link are in focus, each RE signature is associated with the materials in the surrounding. The label space is

$$\Lambda = \{M_1, \cdots, M_Q\}, \tag{4.3}$$

where M is the material label and Q is the number of all possible materials that can be used for constructing the surfaces. The source of labels depends on the radio acquisition method. When a synthetic method is used, a description of the environment can be found in a file used for environment modeling. On the other hand, when radio measurements are conducted in a realistic indoor environment, the information about the properties of the space can be found in the building construction documentation. In both cases, a solution has to be developed to annotate the propagation on a specific radio link with the properties of the surrounding.

In the final step, the data is stored to ensure that it can be easily accessed and analyzed with ML tools. A suitable data organization, data format, and storage solution have to be selected. The former two are concerned with how the data is structured and represented, while the latter is concerned with where the data is physically stored. Regardless of the specific choice for storing the data, it is important to ensure that the data is accurate, organized in a consistent way, and easily accessible. This ensures that the ML algorithms can effectively process the data. We suggest the use of a flat tabular form for organizing the data, represented in the form of comma-separated values (CSV). The data is stored on a local hard drive and available in open databases [205]. The tabular data organization has been identified as a well-suited approach due to its ease of manipulation and processing. The CSV format is widely recognized for its simplicity, efficiency, and compatibility with a broad range of tools and libraries for ML analysis. Each row represents one instance of RE signature from a particular radio link operating at a particular frequency and placed inside a building labeled with binary values indicating the presence/absence of RE properties. The columns in the descriptive space represent the particular property of the multipath propagation while the columns in the output space represent the categorical property of the RE. The output of the module serves for training, validation, and testing in the ML-based modeling module.

4.2.4 Machine learning-based modeling

This module is responsible for outlining the stages of the ML-based process for constructing the RE identification model. The purpose is to provide a clear and structured approach to the ML-based modeling that includes the best practices appropriate for the RE characterization problem.

The module includes two components: problem formalization and ML workflow development. The components form the core of the model-building process. The former is responsible for the indoor RE characterization problem to be adequately defined as an ML problem that can be solved using ML algorithms, and appropriate learning and evaluation approaches are selected; while the latter provides a structured workflow for developing a model ensuring that the set of tasks related to training, evaluation, refinement and final evaluation are performed sequentially, and the final model is accurate and reliable.

To be more specific, the problem formalization is responsible for (i) defining the RE identification task, (ii) selecting data for learning and evaluation, (iii) selecting a ML algorithm, and (iv) selecting evaluation metrics and evaluation schemes; while the ML-workflow development is responsible for the training, evaluation, tuning, and final evaluation of a model. In the training stage, the model is trained on a labeled data set. The goal is to learn relationships between the input and output space in the training data so that the model is able to make predictions on new, unseen data. Once the model is trained, it is evaluated following the selected evaluation scheme to measure its performance. State-of-the-art evaluation measures are considered [34]. Based on the evaluation results, in the tuning stage, the model is fine-tuned by adjusting the hyper-parameters. This workflow is iterative and the tuning and evaluation stages are repeated several times until satisfactory

performance is achieved. Finally, the fine-tuned model is evaluated on a separate test set that has not been used in the training and tuning stages, to obtain a final estimate of the model's performance. The goal of this workflow is to develop a model that is not overfitted to the train data and can accurately make predictions on new CIR data.

The task of RE properties identification belongs to supervised learning and is formalized as multi-label classification where a model outputs a subset from a predefined set of RE properties for a given RE signature as input. The task is approached with problem-transformation methods where the separate binary classifier is trained for each label. Algorithms from various algorithm families (tree-based (DT), ensemble-based (RF), neural networks (MLP), and kernel-based (SVM)) are considered as base learners. The proposed ML workflow is depicted in Figure 4.3.

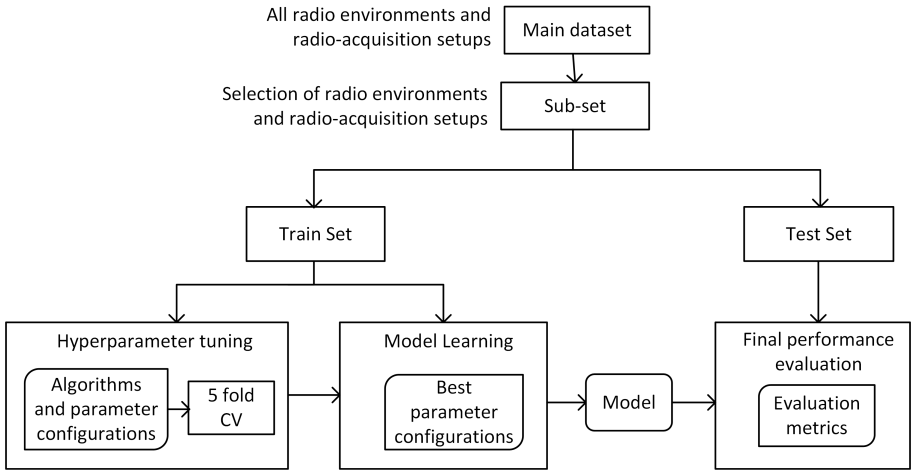

Figure 4.3: The design of the ML workflow.

The multi-label data set that includes all indoor REs and radio acquisition setups also referred to as the main data set is sampled in sub-sets that contain a selection of REs and acquisition setups. The train and test set are created by selecting instances from the subset using the following strategy. Each room is divided into 1 m^2 regions. From each region, one radio node position is selected for training and one radio node is selected for testing. Consequently, both sets include radio nodes that uniformly cover the room (one radio node from each region in the room). The distribution of samples per class is the same in the train and test set since all the rooms that are included in the original sub-set are represented in both sets (train and test).

In the second stage, referred to as hyper-parameter tuning, a 5-fold-cross validation is performed on the train set to find the optimal hyper-parameter configuration for each of the classifiers. The technique involves defining a search space as a set of possible values for each hyper-parameter that we want to consider and exhaustively searching through all possible combinations of the hyper-parameter values in the space to find the combination that yields the best performance. A detailed description of the specific hyper-parameters values evaluated in this study is provided in Table 4.1.

In the third stage, the selected optimal hyper-parameter combination is used for learning a model on the whole train set. In the final stage, we evaluate the predictive performance using the test set. The test set is used only for evaluation of the predictive performance of the trained model and it has not been used at any other stage. For learning, hyper-parameter tuning and evaluations we used MultiOuptutClassifier as meta-learner with DecisionTreeClassifier, RandomForestClassifier, MLPClassifier, and SVC as base-

Table 4.1: Hyper-parameter search space used in the tuning stage of the ML workflow for the base algorithms (Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), and Support Vector Machine (SVM)).

Algorithm	Implementation*	Hyper-parameters and their values
DT	DecisionTreeClassifier	max_depth=[2, 5, 10], min_samples_leaf=[25, 50, 75], criterion=['gini', 'entropy']
RF	RandomForestClassifier	max_depth=[2, 5, 10], min_samples_leaf=[25, 50, 75], n_estimators=[50, 100, 200, 300]
MLP	MLPClassifier	hidden_layer_sizes=[(25,), (22, 11), (32, 8), (16,8)], activation=['logistic', tanh', 'relu'], solver=['sgd', 'adam'], learning_rate=['constant', 'adaptive'], max_iter=[4000, 5000, 6000]
SVM	SVC	kernel=['linear', 'rbf'], C=[0.01, 0.1, 1, 10, 100], max_iter=[1000, 2000]

*The implementations are part of the scikit-learn Python library [44].

learners, GridSearchCV and scores implemented in Scikit-learn Python library [43]. The output of the module is a trained RE identification model that can be used for predicting the properties of the RE on new RE signatures from radio link positions and propagation environments that were not represented in the training process.

4.3 Methodology Use Case

The trained RE model can be used independently for identifying the properties of new indoor RE, or it can be integrated as a component in emerging methodologies for comprehensive indoor environment characterization to provide predictions about the RE that enriches the digital twin of the physical space. Consequently, the RE model is assumed to be one of the most important enablers of environmentally aware indoor wireless communications. The RE identification models can be used in design and optimization of indoor wireless communication networks, as indoor structure geometry and materials are an imperative requirement when deploying wireless systems inside buildings. The use of predictive models shortens the time to characterize (in terms of properties that affect the indoor radio propagation) a particular indoor RE in next-generation scenarios, which is essential for the dynamic adaptation of the next-generation wireless systems to the propagation environment. Additionally, the RE identification model can be used as part of other methodologies for the comprehensive characterization of indoor structures. Due to the versatility of the requirements for spatial description, the emerging methodologies integrate

heterogeneous approaches to provide multi-layered description of the indoor structure.

The procedure for the characterization of particular indoor RE with RE identification model is shown in Figure 4.4. It includes the following steps:

- RE signature acquisition,
- Propagation characteristic processing, and
- RE properties identification.

In the first step, the RE signature is acquired. The RE signature can be obtained from a radio propagation characteristic database or with a RE acquisition (when propagation data for that RE is not available). The propagation characteristic database contains radio propagation characteristics of indoor environments with different geometry and materials for different configurations of the wireless communication system (frequency, antenna configuration, polarization, and position of the radio nodes). The database initially contains data from ray tracing simulations and on-site measurements with equipment based on radio devices that can estimate the CIR. In the future, data from next-generation indoor wireless networks capable to estimate the CIR can also be included. When there is no available data for the particular RE, the RE signature can be acquired with ray tracing tools or measurement campaigns. The data included in the propagation characteristics database, as well as the new data obtained from site-specific simulations or measurements is raw.

In the second step, the raw propagation characteristic is processed and organised in tabular format to be appropriate as input for ML models. In the next-generation scenarios, solutions for the transformation of the raw CIR estimates (from emerging radio technologies) and extraction of relevant features (propagation characteristic per path), need to be developed. The output of the propagation characteristic processing step is formatted RE signature.

In the third step, the trained RE identification model is used to predict the RE properties (shape category, size category, material category, and roughness category) based on the input RE signature.

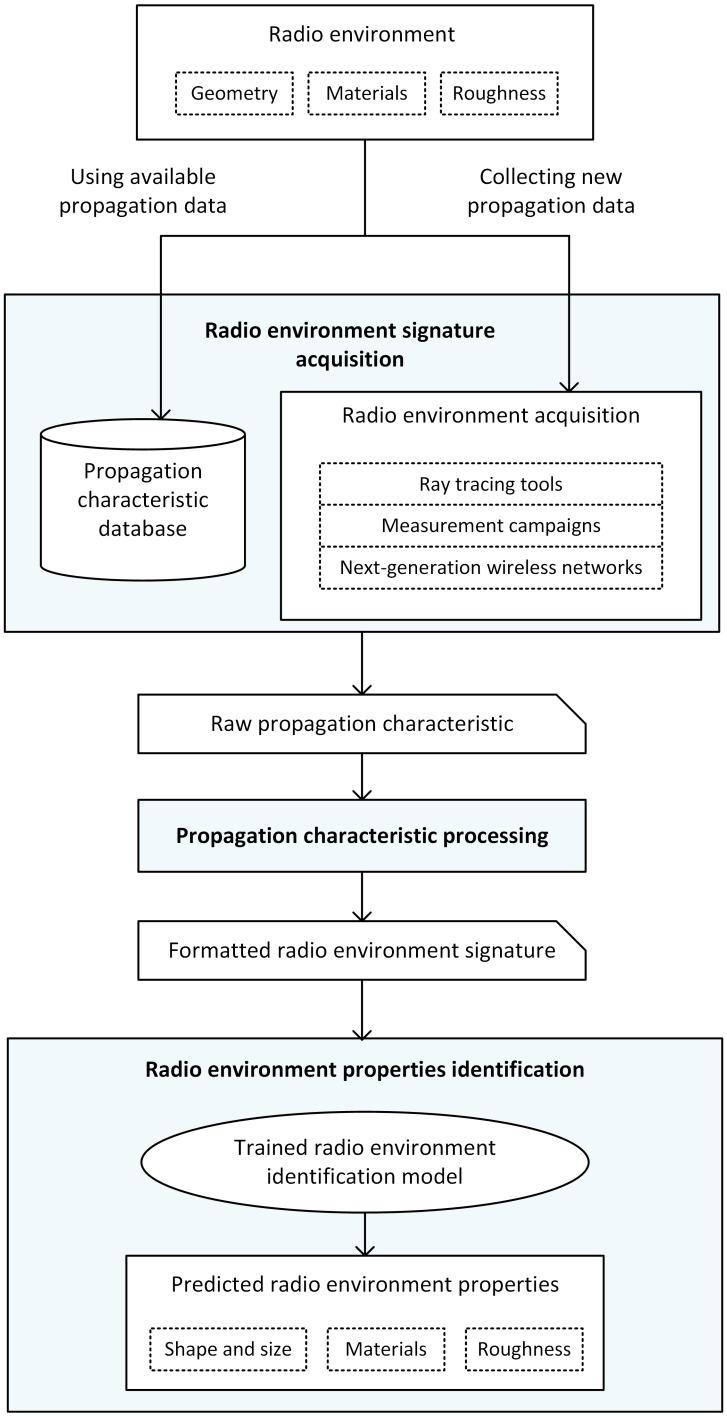

Figure 4.4: Procedure for indoor RE characterization with RE identification model.

Chapter 5

Evaluation Procedure

In this chapter, we describe the procedure for experimental evaluation of the proposed methodology for material characterization in indoor RE. We evaluate the methodology for predicting indoor materials based on CIR. First, we collect CIR data from many different indoor REs that differ in the materials of the facets using computer simulations based on ray tracing. We then conduct initial experiments with an initial set of rooms, in which we build ML models that predict the material of a particular facet in a room based on CIR. Finally, we perform exhaustive experiments with a larger number of rooms with different combinations of materials. In these experiments, we build ML models that predict all materials in a room based on CIR.

5.1 Experimental Questions

In this study, we investigate the ability of ML algorithms to learn predictive models from CIR data to predict the materials in indoor RE from unseen CIR. We investigate the predictive performance of models in indoor environments with varying complexity with respect to the material used. By experimenting with several ML algorithms and different train/test settings in terms of the size of the rooms and the position of the radio links, we compare different strategies for approaching the indoor material prediction problem and investigate the effects of room size and link position on prediction performance. Specifically, we consider the following questions:

- *Capability to predict the material of single facet:* What is the predictive performance of ML models in predicting the material of a single wall?
- *Capability to predict all the materials in a room:* What is the predictive performance of ML models in predicting all the materials present in a room?
- *Specific or more general approach to train the predictive models:* What is the optimal approach for constructing a material prediction model: training a model tailored to a specific room size/layout of radio links, or training a more general model for multiple room sizes/layouts of radio links? Under what conditions should each approach be preferred?
- *Impact of the room size and position of radio links on the ability of the models to accurately predict the materials:* Do the room size and the position of the radio links considered in the training phase affect the ability of the trained models to make accurate predictions about the materials for data from link positions and room sizes that were not included in the training process?

5.2 Indoor Environment

We considered the following factors to describe the indoor propagation environment:

- Geometry:

  Geometry refers to the overall shape, dimensions, and spatial layouts of the indoor space. It includes the position, dimensions, shapes, and relation between all architectural features, such as floors, ceilings, walls, columns, windows, and doors. In general, the space is enclosed with cuboids and the thickness of each cuboid is also the geometrical property of the structure important in through-the-wall radio propagation. However, since we are not interested in the wave energy that penetrates the structure as a result of the transmission mechanism, we approximate the cuboids with facets.

  We look at common and basic indoor structures (rooms), defined with four walls forming an enclosure around the floor and ceiling. Each room is considered as a separate entity and it is not related to an entire floor composed of many rooms. This geometry is basic, yet representative of the most common rooms inside residential and office buildings. The geometry of an empty room is schematically represented in Figure 5.1. The facet's names, the number of facets per room, and their position within the room layout, considered in the experiments, are summarized in Table 5.1. The overall geometrical properties of the rooms are summarized in Table 5.2.

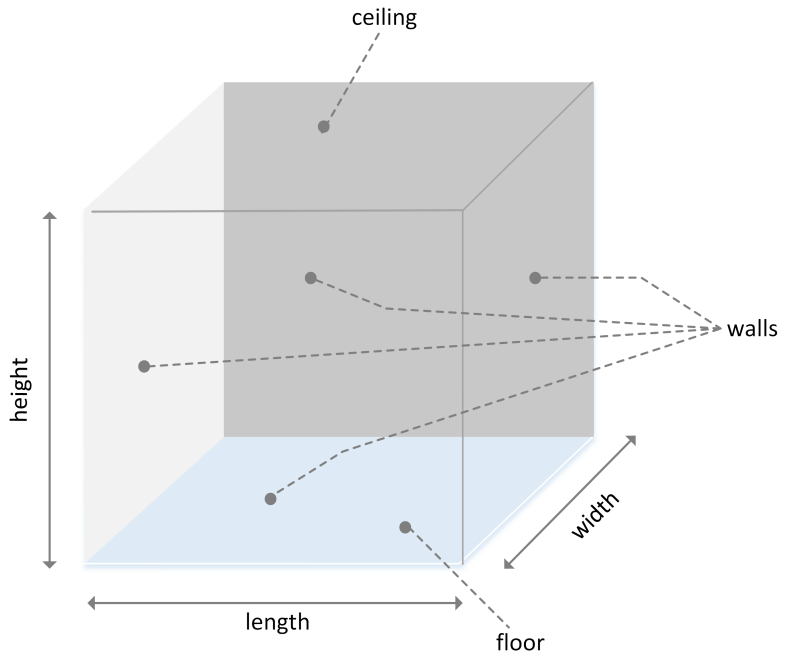

Figure 5.1: Schematic representation of the room geometry. This scheme represents the facets (floor, ceiling, and walls) that define the room, the dimensions (length and width) of the floor/ceiling, and the vertical distance between the floor and ceiling (height).

Table 5.1: Description of the facets that enclose the indoor space.

Facet name	Total number of facets per room	Position in the room layout
floor	1	horizontal-bottom
ceiling	1	horizontal-top
wall	4	vertical-lateral

Table 5.2: Summary of the geometrical properties of the rooms. S: small, M: medium, L: large, w: width, l: length, h: height.

room-size	floor/ceiling surface	floor/ceiling dimensions (w × l)	wall dimensions (l × h)	floor/ceiling shape	walls shape
S	9 m^2	3 m × 3 m	3 m × 3 m	square	square
M	25 m^2	5 m × 5 m	5 m × 3 m		rectangular
L	49 m^2	7 m × 7 m	7 m × 3 m		rectangular

With respect to the shape, we looked at rooms with a square shape in terms of the floor shape. With respect to the size of the room, we looked at three room-size categories. Each category is determined by the area of the floor and the vertical distance between the floor and ceiling. The area of the floor in rooms with different sizes is 9 m^2, 25 m^2, and 49 m^2. The area corresponds to floor dimensions of 3 m by 3 m, 5 m by 5 m, and 7 m by 7 m, consequently. In the later text, the rooms with floor surface equal to 9 m^2 are referred to as small (S) rooms, the rooms with floor surface equal to 25 m^2 are referred to as medium (M) rooms, and the rooms with floor surface equal to 49 m^2 are referred to as large (L) rooms. The vertical distance between the floor and ceiling is 3 m, for all rooms. The dimensions of the walls depend on the overall dimensions of the room. Therefore, in S rooms, the walls have a square shape, length of 3 m and height of 3 m; in M rooms, the walls have a rectangular shape, length of 5 m and height of 3 m; and in L rooms, the walls have a rectangular shape, length of 7 m and height of 3 m.

- Roughness of the surfaces:

  The irregularities on the surface of the facets scatter the radio wave in multiple directions. This propagation mechanism should be considered in high-frequency regimes because even physically small variations on the surfaces can scatter the radiation [206]. We assume that the surface of the facets is smooth, i.e., the surface's variations are so small compared to the wavelength that they do not scatter the waves.

- Materials:

  The materials present in a room belong to several categories, such as structural materials used to construct the room, surface materials used to cover the facets and other structural components in the room, and materials of the objects present in the room. We focus on the materials used to construct the facets. More specifically,

we look at rooms where the same material that is used for constructing the facets appears on the surface, i.e., there is no additional cover on the surface made of a different material. The materials are observed from the perspective how they affect the propagation.

Each material is described with its constitutive parameters (electric permittivity, magnetic permeability, and conductivity), and the reflection coefficient [206]. The constitutive parameters describe how EM waves behave as they interact with the specific material and propagate through it, while the reflection coefficient expresses the ratio of the reflected electric field to the incident field on the boundary between two media with different permittivities and permeabilities. The electric permittivity and magnetic permeability determine the speed at which EM waves travel through the medium, as well as the amount of energy that is absorbed or reflected by the medium [6]. The conductivity determines the amount of energy that is lost due to electrical resistance in the medium [6]. Other intrinsic properties of the material itself are not considered. The rationale behind this comes from fundamental electromagnetism theory and Maxwell's equations showing the medium properties impact on the strength of the fields [6]. The electric permittivity is expressed relative to the permittivity in free space as

$$\epsilon = \epsilon_0 \epsilon_r F m^{-1}, \tag{5.1}$$

where ϵ_0 is the permittivity in free space ($\epsilon_0 = 8.854 \times 10^{-12} F m^{-1}$), and ϵ_r is the relative permittivity. The relative permittivity is a dimensionless, complex-valued quantity:

$$\epsilon_r = \epsilon_r^{'} + j\epsilon_r^{''}, \tag{5.2}$$

where $\epsilon_r^{'}$, and $\epsilon_r^{''}$ are the real and imaginary parts, respectively. The imaginary part is:

$$\epsilon_r^{''} = \frac{\sigma}{\epsilon_0 \omega_c}, \tag{5.3}$$

where σ is conductivity measured in Siemens per meter $[S m^{-1}]$, and

$$\omega_c = 2\pi f_c rad/s \tag{5.4}$$

is the angular frequency. The magnetic permeability is expressed relative to the permeability in free space as

$$\mu = \mu_0 \mu_r H m^{-1}, \tag{5.5}$$

where μ_0 is permeability in free space ($\mu_0 = 4\pi \times 10^{-7} H m^{-1}$), and μ_r is relative permeability. The reflection coefficient Γ of a material is a function of the relative permittivity of the material ϵ_r and the angle of incidence α. When the angle of incidence is equal to the angle of reflection, the reflection coefficient for horizontal polarization (HP) and vertical polarization (VP) can be calculated as

$$\Gamma_{HP} = \frac{\cos\alpha - \sqrt{\epsilon_r - \sin^2\alpha}}{\cos\alpha + \sqrt{\epsilon_r - \sin^2\alpha}}, \tag{5.6}$$

$$\Gamma_{VP} = \frac{\epsilon_r \cos\alpha - \sqrt{\epsilon_r - \sin^2\alpha}}{\epsilon_r \cos\alpha + \sqrt{\epsilon_r - \sin^2\alpha}} \tag{5.7}$$

Although there are various materials available for indoor construction, some of them are more frequently used in practice, and usually specific facets tend to be associated with particular materials. In the general indoor construction practice, the floors are commonly constructed of hardwood, tile or concrete, and the ceilings are typically

constructed of concrete, but also gypsum boards or suspended ceilings are used often, and when it comes to walls, brick, concrete, plaster, and glass are used. The choice of materials depends on various factors, such as function, need for durability, aesthetic, etc., and also on the specific regulations and building codes that must be followed.

In this study, we considered a set of five materials with different permittivity: brick, concrete, glass, plaster, and wood. The frequency-dependent values of the relative permittivity and conductivity of the materials are summarized in Table 5.3. We used the material electrical properties at the frequencies of interest provided by the International Telecommunication Union (ITU-R) [13]. The materials are considered to be non-ionized and non-magnetic. Therefore, the density of free charge is set to zero and the permeability for all materials is set to the permeability of free space. The reflection coefficients of the materials for vertical and horizontal polarization are plotted in Figure 5.2. The plot shows the reflection coefficient in decibels as a function of the incidence angle.

Table 5.3: Electrical properties of materials [13].

material	relative permittivity	conductivity	frequency range [GHz]
brick	3.75	0.038	1-40
concrete	5.31	0.120	0.001-100
glass	6.27	0.029	0.1-100
plaster	2.94	0.036	1-100
wood	1.99	0.026	0.001-100

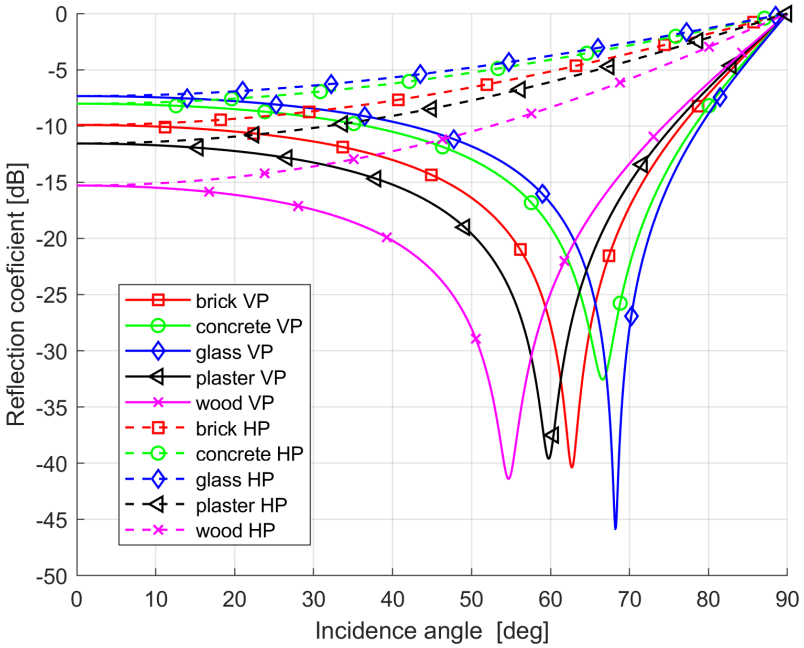

Figure 5.2: Reflection coefficients of brick, concrete, glass, plaster, and wood. The solid lines represent the reflection coefficients in the case of vertical polarization, while the dashed lines represent the reflection coefficients in the case of horizontal polarization. VP: vertical polarization, HP: horizontal polarization, f=3494.4 MHz.

Within the selected set of materials, we identified three groups of materials that are commonly used for particular facets (floors, ceilings, and walls). We refer to the materials used for constructing the floors as floor-materials, to the materials used for constructing the ceilings as ceiling-materials, and to the materials used for constructing the walls as wall-materials. In Figure 5.3, the relation between the material groups and the facets of an empty room is schematically represented. The floor-materials group includes concrete and wood; the ceiling-materials group includes concrete, plaster, and wood; and the wall-materials group includes brick, concrete, glass, plaster, and wood.

- Other obstacles:

  While large objects in the building, such as the floor, ceiling, and walls, have a dominant effect on radio propagation, the obstacles with smaller sizes, such as tables, chairs, cabinets, and people, can also affect the propagation. The frequency of the radio signal is a major factor that determines the extent to which the small objects interact with the propagating wave. Also, the effect of a specific object depends on the size and shape of the object and its material, as well as its position relative to the transmitter and receiver.

  We consider only the reflected paths that are the result of the reflection that occurs when the waves are incident onto smooth, large (relative to the wavelength) facet. Therefore, we look only at empty rooms. By simplifying the environment, the propagation characteristics of the paths that convey most of the environmental information can be obtained, without the need to process propagation characteristics that are not relevant to the learning task.

In our study, we included 11,250 rooms in total. In particular, based on the materials used for constructing the facets in the room, these rooms belong to 3,750 distinct room

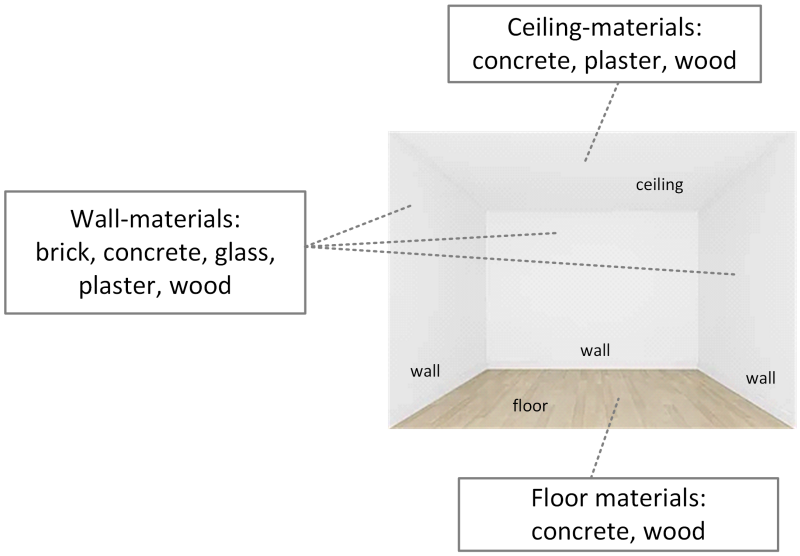

Figure 5.3: Schematic representation of the materials considered for constructing floors, ceilings, and walls.

types and each room type is considered in the three room sizes defined earlier in this section. The number of distinct room types comes from the materials used for the floor and ceiling, having six floor-ceiling material combinations; and the materials used for the walls, having 625 wall materials combinations.

5.3 Communication System

The communication system has a crucial role in the RE acquisition process. Appropriate configuration is essential to ensure that the propagating waves convey sufficient environmental information for environment characterization. The configuration refers to the selection of system parameters, such as frequency, bandwidth, transmission power, and receiver sensitivity, as well as the number of antennas and their characteristics, and geometry of antenna arrays.

The operating frequency is a fundamental parameter since it determines the behavior of the radio waves as they travel through and interact with the indoor environment. The bandwidth has a significant impact on the capability to distinguish MPCs. A wider bandwidth allows more precise time-domain resolution, which means that the receiver can distinguish between the different MPCs arriving at different times. This is important for extracting features for each propagation path that is used for training the ML models. The transmission power affects the distance that the wave can travel and the strength of the received signal while the receiver sensitivity affects the receiver's ability to detect weak signals. These two parameters have to be set to achieve the desired range and received signal strength while minimizing the noise.

The parameters of the communication system are specified so that the study can be upgraded with the experimentally measured RE signatures in a real environment. We consider UWB radio technology according to 802.15.4-2011 standard, thus we selected carrier frequency equal to 3494.4 MHz [207].

Omnidirectional antennas are used due to their equal radiation and reception in all directions, which makes them adequate for distributing the signal through the space and acquiring the properties of the indoor RE. Additionally, the omnidirectional antennas re-

quire minimal setup and calibration which makes them a suitable choice for collecting large amounts of propagation data for ML analysis. We select the mounting height according to the common real-life positioning of the radio devices, thus antenna is mounted 1.5 m above the floor. The input power is set to 0 dBm, the polarization is vertical and the threshold for the receiver is set to -250 dBm. The configuration of system parameters is summarized in Table 5.4.

Table 5.4: Wireless system parameters configuration. Tx: transmitter, Rx: receiver.

system parameter	configuration
frequency	3494.4 MHz
bandwidth	UWB
polarization	vertical
antenna type	omnidirectional
Tx/Rx height	1.5 m

One fixed radio node and a portable node that can change its location are used in the standard channel measurement procedure for estimating indoor and outdoor CSI estimation. A uniform grid is defined in room, indicating the possible location of the portable node, while the three possible topologies of the fixed nodes in the room are considered, namely (1) one node in the center of the room, (2) one node in the center of the room but equipped with multiple antennas arranged in a circle, and (3) four nodes located near the room corners. The first two fixed node topologies correspond to CSI estimation for training purposes, while the last one is very common for wireless systems in operation, where the access points or the base stations are located at the room corners. In order to study the impact of CSI estimation for training and testing, the three link layouts are specified, namely L1 corresponding to the node topology 1, L2 corresponding to topology 2, and L3 corresponding to topology 3.

In Figure 5.4, a schematic representation of the radio nodes positions relative to the room geometry is given. The layout-specific and grid positions of the nodes are presented in two separate top views of the room. In L1, the layout-specific radio node is placed in the center of the room. In L2, the layout-specific radio nodes are evenly spaced in a circular pattern around a central point which is in the center of the room. The distance from the center of the room to the circumference of the circle is 0.5 m, and the spacing between the radio nodes is $\pi/4$ rad. In L3, the layout-specific nodes are placed close to each corner of the room. The coordinates of the layout specific nodes for the defined room sizes are given in Table 5.5.

The portable nodes are arranged in a grid pattern with equal distance between the nodes and covering the whole room. The corners of the grid are 0.25 m apart from the walls and the distance between the nodes is also 0.25 m. Since the grid size is defined relatively to the room size, the total number of grid nodes is different in rooms with different sizes. For rooms belonging to S, M, and L room sizes, the total number of grid nodes is 121, 361, and 729, respectively. The number of radio links per room size is given in Table 5.6.

An example of the 15 strongest rays between the layout-specific radio nodes in L1, L2, and L3 and one grid node in M room are illustrated in Figure 5.5, Figure 5.6 and Figure 5.7.

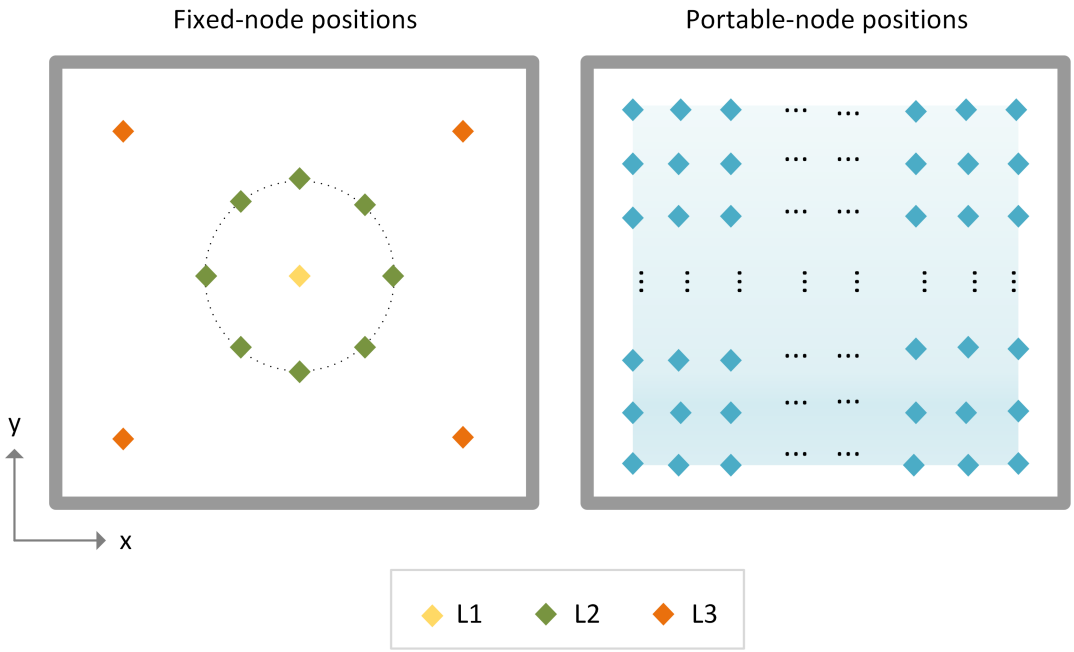

Figure 5.4: Schematic representation of the locations of fixed and portable radio node.

Table 5.5: Cartesian coordinates (x, y) of the positions of layout-specific nodes. The coordinates are in meters.

radio-link layout	room size					
	S		M		L	
	x	y	x	y	x	y
L1	1.50	1.50	2.50	2.50	3.50	3.50
L2	2.000	1.500	3.000	2.500	4.000	3.500
	1.850	1.850	2.850	2.850	3.850	3.850
	1.500	2.000	2.500	3.000	3.500	4.000
	1.150	1.850	2.150	2.850	3.150	3.850
	1.000	1.500	2.000	2.500	3.000	3.500
	1.150	1.150	2.150	2.150	3.150	3.150
	1.500	1.000	2.500	2.000	3.500	3.000
	1.850	1.150	2.850	2.150	3.850	3.150
L3	0.375	0.375	0.375	0.375	0.375	0.375
	2.625	0.375	4.625	0.375	6.625	0.375
	2.625	2.625	4.625	4.625	6.625	6.625
	0.375	2.625	0.375	4.625	0.375	6.625

Table 5.6: Number of radio links.

radio-link layout	room size		
	S	M	L
L1	121	361	729
L2	968	2888	5832
L3	484	1444	2916
Total number of radio-links per room:	**1573**	**4693**	**9477**

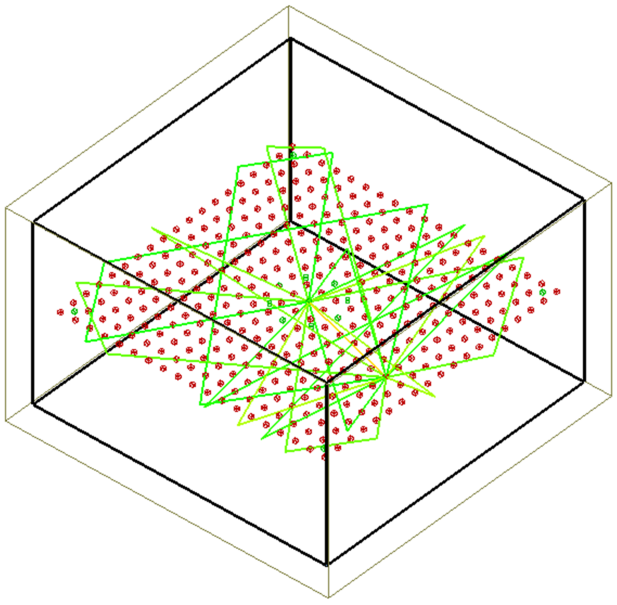

Figure 5.5: Strongest radio rays between single radio node of layout L1 and a single grid node.

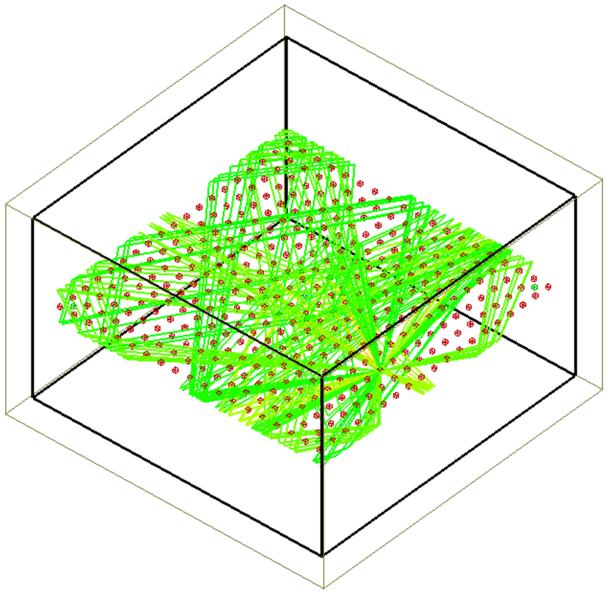

Figure 5.6: Strongest radio rays between eight radio nodes of layout L2 and single grid node.

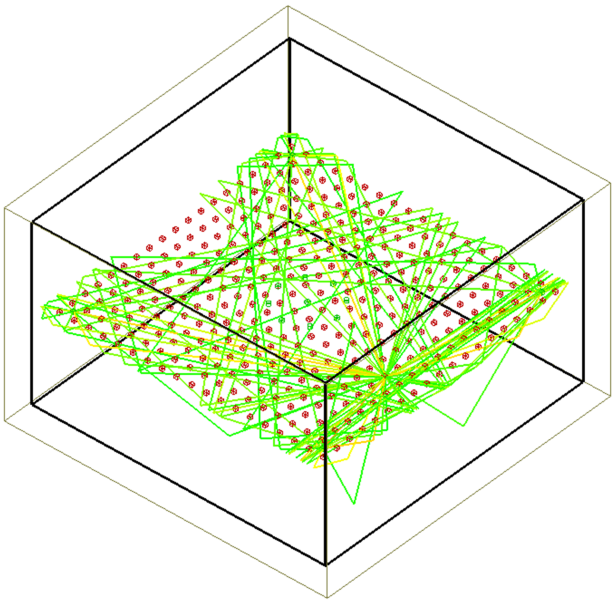

Figure 5.7: Strongest radio rays between four radio nodes of layout L3 and single grid node.

5.4 Data Collection

The workflow for building a data set with CIR values for radio links in different positions within an indoor environment is depicted in Figure 5.8. It included the following steps:

- Design of simulation scenarios.

- Development of a solution for automation of the simulation input parameters modification and simulation execution.
- Ray-tracing and CIR calculation.
- Integrating and processing the simulation output.

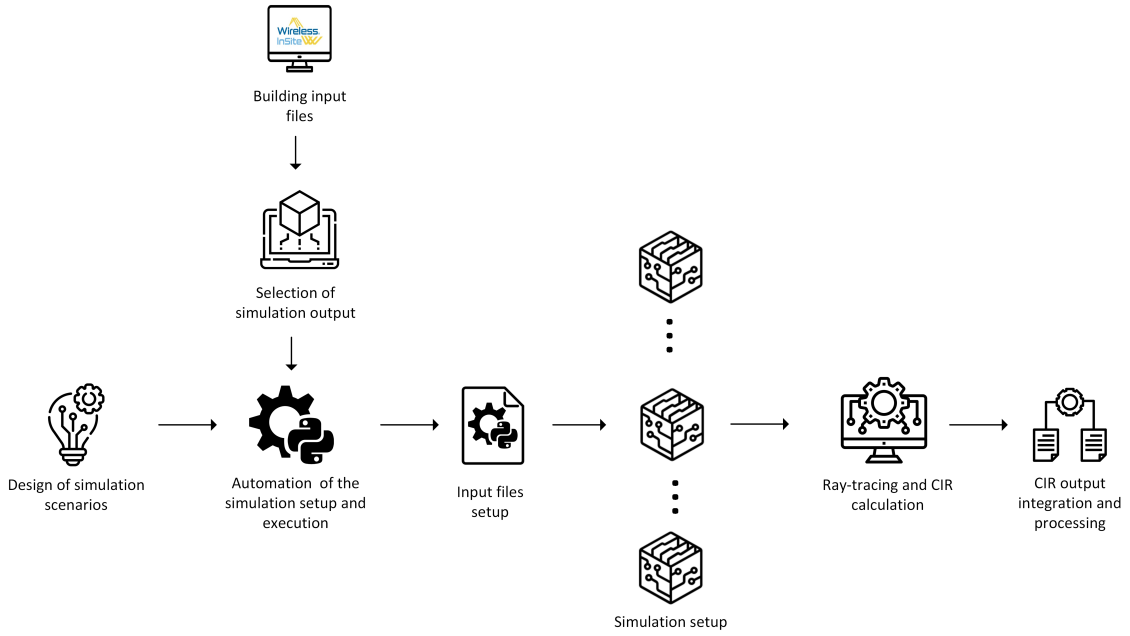

Figure 5.8: Workflow for building CIR data set.

In the first step, a tool for indoor radio simulation is selected and the simulation scenarios are defined according to the description of the rooms given in Section 5.2, and the communication system given in Section 5.3. A commercially available 3-D wireless prediction software, Remcom Wireless InSite, is chosen for predicting the effects of buildings on the propagation of the EM waves and generating accurate values for the specific propagation characteristics [31]. Wireless InSite is a 3-D wireless prediction software for analysis of site-specific radio wave propagation. The calculations are based on shooting rays from the transmitters and propagating them through the indoor geometry, evaluating the complex field of each ray, and combining the contributions of arriving ray paths to determine the propagation quantities. The field estimates, when the parameters of the propagation environment and the wireless system are given, are representative of the fields associated with the ray paths in a real-life scenario with the corresponding parameters. The key features of this tool suitable for our study are (i) modeling capabilities—accurate ray-tracing models and support for indoor environments; (ii) site-geometry editors—creation and modification of indoor environments; (iii) import functionality—import of indoor geometries from CAD files and antenna patterns from common formats; (iv) databases of waveforms, antennas, and materials; (v) large set of outputs that can be calculated and written to ASCII-based files; (vi) support for facet-level material definition; (vii) GPU acceleration and multithreading; (viii) support for user-defined input files; and (viii) command-line execution of the calculation engine. The description of the simulation scenarios is stored in scenario files. The scenario files include a (i) description of the propagation environment—geometry and materials; and (ii) transmitter/receiver configuration—location relative to the geometry of the room, carrier frequency, bandwidth, antenna type, mounting height, orientation, and

configuration; as necessary inputs for the simulation tool. The Wireless Insight workflow includes a setup and calculation phase as shown in Figure 5.9 [31]. In the setup phase, the RE, transmitter/receiver location, antennas, waveforms, and propagation model are defined and the output type is selected, while in the calculation phase, the geometry is processed, the propagation paths are determined, the field associated with each path is evaluated, the requested propagation quantities are calculated and the generated output is written to files.

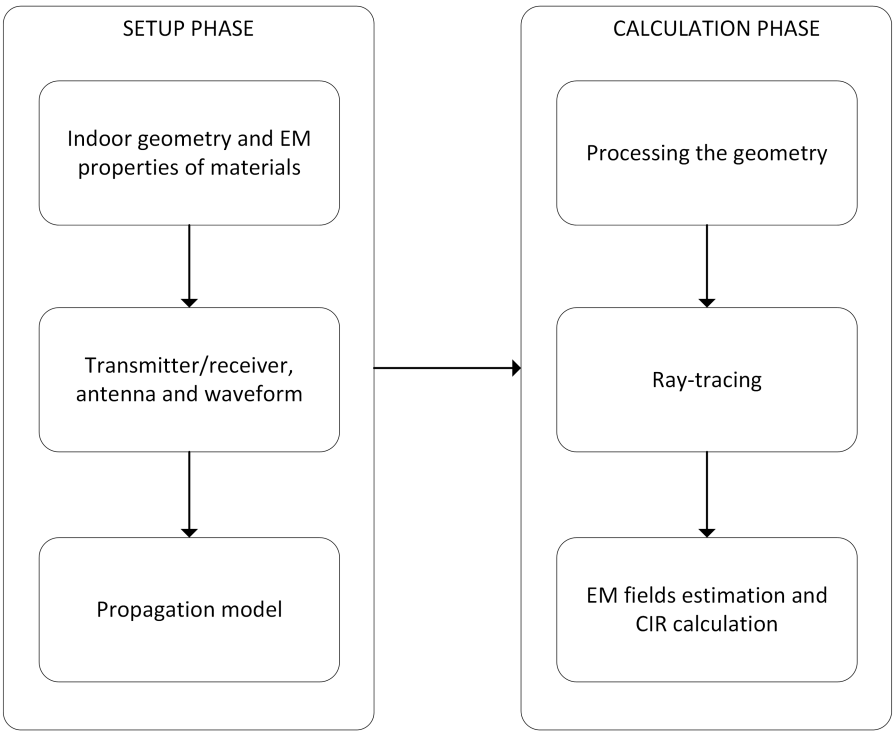

Figure 5.9: Workflow for propagation characteristic estimation based on ray-tracing.

In the second step, since data from a large number of scenarios has to be collected, a solution is developed to control the tool, i.e., the input parameter modification, running the simulation's engine from the command line and output storing. In particular, we develop a Python script for automatic change of the Wireless InSite input parameters, simulation execution, and output storage [208]. Data and parameters that should be configured to perform the simulation refer to the floor plan, materials, antennas, waveforms, transmitters, receivers, and outputs.

In the third step, the simulation project is set using the Python script for input parameter setup, the ray paths from the transmitter to the receiver are found using ray-tracing, and the CIR is calculated. The indoor environments are defined with coordinates of the vertices and materials for each facet in an ASCII floor plan file that can be exported for further use. The floor plan file contains four sections:

- File header—specifies the type of geometry;
- Location section—specifies the origin of the floor plan in Cartesian coordinates;
- Material section—specifies the materials that are referenced to the facets with their parameters; and

- Structure-groups section—specifies the facets that define the geometry. Each facet is described with a set of vertices and an assigned material.

The materials are stored as part of the floor plan in the floor plan file. Each material is defined with its EM properties (conductivity and relative permittivity), from which the reflection coefficients are calculated.

The waveform properties are defined in a waveform file format that includes the type of the waveform and its specific values, like the waveform shape. We consider the sinusoidal type of waveform which is described by the carrier frequency, bandwidth, and initial phase shift. The omnidirectional antennas are defined by specifying the antenna type, the reference to the waveform assigned to the antenna, gain, polarization, transmission line loss, receiver threshold, and voltage standing wave ratio. The properties of the transmitter and the receiver are stored in transmitter and receiver files, respectively. The transmitter/receiver files have four sections: header, location, antenna, and additional properties.

The region in which to perform simulations, the propagation model to be used and the outputs to be generated are defined with the Wireless InSite study area. The size and the location of the study area are selected to include the entire indoor environment. The selected propagation model is X3D [31]. This is 3-D propagation model with no restrictions on geometry shape or transmitter/receiver height. It includes reflections, transmissions, and diffractions along with atmospheric absorption and diffuse scattering and supports frequencies up to 100 GHz. The selected ray-tracing method is SBR with exact path correction which reduces error in calculated power and phase [31], [209].

All the predictions for the CIR values per ray are generated by the Wireless InSite calculation engine. The electric field is evaluated in the far zone of the transmitting antenna. In 3-D calculations, the polarization of the incident electric field E^i is neither completely parallel nor completely perpendicular to the plane of incidence at each reflection, so the reflected field E^r is then calculated from the two reflection coefficients $\Gamma_\parallel$ and $\Gamma_\perp$ where the components of the reflected electric field parallel $E^r_\parallel$ and perpendicular $E^r_\perp$ to the reflection plane are calculated as follows

$$\begin{pmatrix} E^r_\parallel \\ E^r_\perp \end{pmatrix} = \begin{pmatrix} \Gamma_\parallel & 0 \\ 0 & \Gamma_\perp \end{pmatrix} \begin{pmatrix} E^i_\parallel \\ E^i_\perp \end{pmatrix}. \tag{5.8}$$

At each receiver site, the CIR component is calculated per ray path and written to ASCII-based files. Each file has header rows that describe the transmitter and receiver and per-ray CIR values. Each row has a separate column for each of the three CIR parameters (time of arrival in seconds, phase in degrees, and received power in dBm). We keep the CIR for the 15 strongest paths, including the direct path, single-bounced paths, and paths with multiple bounces.

The CIR of the i-th MPC s_i is

$$s_i = P_i e^{j\Psi_i}, \tag{5.9}$$

where P_i is the power carried by the i-th MPC and

$$\Psi_i = \tan^{-1}(\frac{Im(V_i)}{Re(V_i)}) \tag{5.10}$$

is the phase of the i-th MPC [31]. The complex voltage at the feed point of the receiving antenna due to the i-th MPC V_i is proportional to

$$V_i \propto E_{\theta,i} g_\theta(\theta_i, \phi_i) + E_{\phi,i} g_\phi(\theta_i, \phi_i), \tag{5.11}$$

where θ and ϕ are theta and phi coordinates in global spherical coordinate system, $E_{\theta,i}$ and $E_{\phi,i}$ are the θ and ϕ components of the electric field of the i-th MPC at the receiver point, and g_θ and g_ϕ are the direction of arrival in terms of θ and ϕ. The θ component of the direction of arrival is given by

$$g_\theta(\theta,\phi) = \sqrt{|G_{\theta(\theta,\phi)}|}e^{j\psi_\theta}, \tag{5.12}$$

where G_θ is the θ component of the receiving antenna gain and ψ_θ is the relative phase of the θ component of the far zone electric field. The ϕ component of the direction of arrival is calculated analogously. The CIR from radio links in two different rooms in terms of the materials used is shown in Figure 5.4.

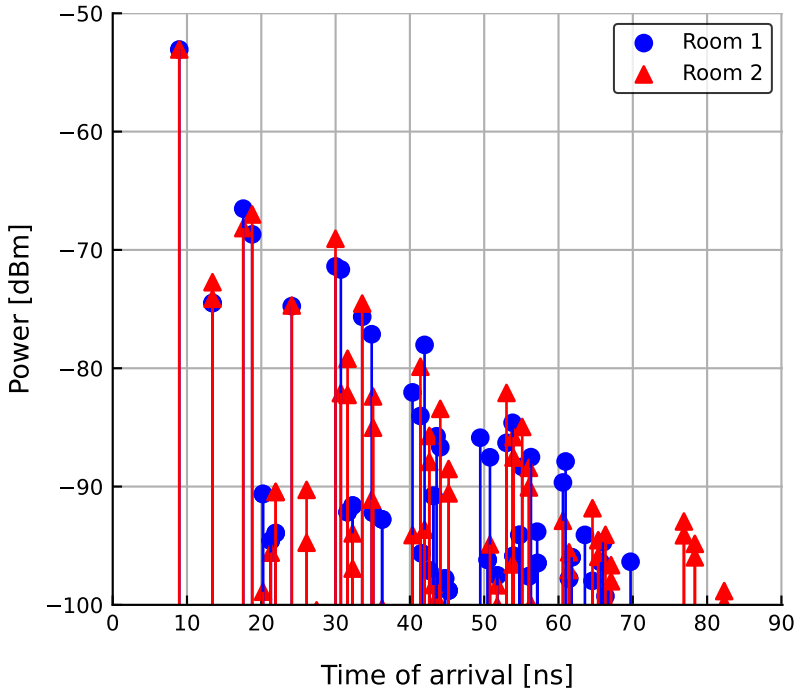

Figure 5.10: CIR of radio link in two example rooms. Room 1—floor: concrete, ceiling: concrete, wall 1: concrete, wall 2: concrete, wall 3: concrete, wall 4: brick. Room 2—floor: wood, ceiling: plaster, wall 1: wood, wall 2: glass, wall 3: brick, wall 4: concrete.

Finally, the simulation output is integrated and organised in a tabular data set. Organising the simulation output, which is stored in multiple proprietary file formats, into a structured data set appropriate for ML analysis is a crucial step for leveraging the simulation data for ML modeling. The relevant CIR values are extracted from the simulation output files and integrated into a single, structured data set using Python to read, process, and integrate data.

5.5 Experimental Scenarios

The study includes both initial and advanced experimental scenarios in terms of goal, indoor environment complexity, and evaluation setup (radio links and rooms in the training and testing process). The implementation of both scenarios enables the development of an

informed approach to data-driven indoor environment characterization based on CIR. The iterative process of defining and refining the experimental scenarios is essential because the problem is previously unexplored and there are no approaches in the literature for using CIR to predict the materials in the indoor propagation environment with ML models.

An initial experimental scenario, referred to as *Baseline Scenario* is defined, to

- develop a foundational understanding of how to approach the problem,
- investigate the feasibility of the concept, and
- study the potential of ML models to predict materials in basic propagation environments based on CIR.

This scenario serves as the foundation for refining the subsequent experimental design. Specifically, data is collected from an initial, carefully selected set of rooms, chosen based on the materials used, to enable a more focused analysis of the research problem. The decision to use a reduced number of rooms is guided by pragmatic and methodological considerations. In particular, a smaller data sample facilitates a more efficient and manageable validation of the concept's feasibility, while requiring fewer computational resources.

A more detailed experimental scenario, referred to as *Comprehensive Scenario* is defined, to

- study the ML-based material prediction in a larger set of rooms with varying complexity in terms of the number and combination of materials, and
- assess the generalization performances of the models in various evaluation setups in terms of link positions and room sizes.

When a problem that has not been previously addressed using ML techniques is approached, it is important to evaluate different learning algorithms. The effectiveness and suitability of different learning methods for constructing models that capture the relationships between the CIR and the materials in the propagation environment is enabled by the evaluation of multiple algorithms. Several factors, such as the complexity and nature of the problem, the characteristics of the data, the available computational resources, and the requirement for model interpretability, determine the selection of a suitable ML algorithm.

In both experimental scenarios, four different classification algorithms, as representatives of tree-based (DT), ensemble (RF), neural network (MLP) and kernel-based (SVM) learning methods, are used. Given the novelty of the problem and, consequently, the lack of sufficient data and the interpretability requirement, deep learning methods are not considered. The algorithms and their implementations are described in Section 2.2.1.2. In order to assess the ability of the models to generalize to new CIRs, independent test data that has not been used in any of the previous phases of the model development process is considered. All trained models are tested with new data from radio links other than those used in the training and tuning phases.

5.5.1 Baseline scenario

In this scenario, a subset of 48 rooms from the set of rooms described in Section 5.2 is selected for the initial investigation of indoor material identification. The set comprises 16 rooms for each of the three room size categories (S, M, and L). Careful consideration is given to the effects of different materials on indoor radio propagation when selecting the propagation environments for the initial experiments. Rooms are chosen where concrete is the predominant building material for the floor, ceiling, and three walls, while one wall is made of a different material from a predefined set of materials, including brick,

glass, plaster, and wood. More detailed explanations of the selection of materials are given in Section 5.2. The objective is to identify the material of the different wall, given that concrete is present in all rooms. The learning task is formalized as a multi-class classification and a predictive model is learned that outputs a material class representing the material of the wall for the input CIR. The formal description of the task and the methods for multi-class classification are provided in Section 2.2.1.

During the predictive model developing, the selection of the data for training the model is crucial. Initially, it is important to investigate how the problem can be approached, either by (i) training specific models for different room sizes and radio link layouts, or (ii) constructing more general models that can be used for a wide range of rooms and layouts.

Both specific and general approaches in terms of link location and room size are being considered. In terms of the room size,

- specific models – using data from a single room size, i.e., S, M, and L, and
- more general models – using data from different (two and three) room sizes, i.e., SM, ML, SL, and SML, are being trained.

The specific approach is employed to examine how the model performance is affected by room size, whereas the general approach is utilized to investigate the effects of combining data from different room sizes. To construct a more general model that is not tailored to a single room size, the inclusion of room sizes in the training phase is an important consideration, as the performance of the model can be influenced by the range of sizes represented in the training process. Different combinations of room sizes are being considered to examine the effects of combining different sizes on model performance. In terms of the link layout,

- specific models – using data from a single link layout, i.e., L1, L2, and L3, and,
- more general models – using data from multiple layouts, i.e., L1+L2+L3 (ALL), are being trained.

The specific approach is used to study how the link layout affects model performance, while the general approach is used to study the effect of combining data from different link layouts on model performance.

The model performance is assessed using an initial evaluation scheme referred to as *EvalSch–Init*, where different locations of the portable node are used for testing resulting in different link locations and orientations for training and testing. The room sizes and link layouts included in the test set are represented in the train set. The scenario where the model is expected to make predictions for new data collected using the same acquisition layout as for collecting training data is represented by this evaluation scheme. The definition of the evaluation settings in terms of the room size(s) and link layout(s) included in the train and test set is as follows:

- ***1RS–train/1RS–test:*** Single room size for training and testing.

  Models are trained with data of each room size (S, M, and L), and they are tested on the same size. Therefore, the room sizes used for training and testing are S/S, M/M, and L/L.

  - ***1Lyt–train/1Lyt–test:*** Single link layout for training and testing.

    Models are trained using data acquired with single layout (L1, L2, and L3), and tested on data from the same layout.

- ***mLyt–train/mLyt–test:*** Multiple link layouts for training and testing.
  Models are trained with data acquired with several layouts (ALL), and tested on data obtained with the same layout.

- ***2RS–train/2RS–test:*** Two room sizes for training and testing.

  Models are trained for each pair of room sizes (SM, ML, and SL), and evaluated on the same sizes.

  - ***1Lyt–train/1Lyt–test:*** Single link layout for training and testing.
    Models are trained using data acquired with single layout (L1, L2, and L3), and tested on data from the same layout.
  - ***mLyt–train/mLyt–test:*** Multiple link layouts for training and testing.
    Models are trained using data acquired with several layouts (ALL), and tested on data obtained with the same layout.

- ***3RS–train/3RS–test:*** Three room sizes for training and testing.

  Models are trained for three room sizes (SML), and evaluated on the same sizes.

  - ***1Lyt–train/1Lyt–test:*** Single link layout for training and testing.
    Model is trained using data acquired with single layout (L1, L2, and L3), and tested on data from the same layout.
  - ***mLyt–train/mLyt–test:*** Multiple link layouts for training and testing.
    Models are trained on data acquired using several layouts (ALL), and tested on data from the same layouts.

5.5.2 Comprehensive scenario

In this scenario, the full set of 11,250 rooms described in Section 5.2 is considered. The set includes 3,750 rooms for each room size (S, M, L). The rooms have different numbers and combinations of materials from a predefined set, including brick, concrete, glass, plaster, and wood. More detailed explanations of the selection of materials are given in Section 5.2. This set represents the wide variety of rooms (in terms of material used for the facets) that exist in a real office and residential buildings. The objective is to use the CIR to identify all surface materials in the indoor environment. Therefore, the goal is to learn a predictive model, using CIRs from spatially distributed radio links, which outputs the subset of materials that exist in the RE for unseen CIR. The learning task is formalized as multi-label classification and the problem transformation approach is used to learn from multi-label data [41], [67]. The formal definition of the task and the description of the problem transformation method are given in Section 2.2.2.

One of the most important aspects of any ML problem is the choice of the evaluation scheme that determines the selection of data used in the training and testing phase. In particular, in the evaluation setup for the problem of indoor material prediction, one of the most important aspects is whether or not the size of the rooms and the layout of the radio links included in the train set are also present in the test set. We use the initial evaluation scheme *EvalSch–Init*, and define three additional evaluation schemes: *EvalSch–Lyt*, *EvalSch–RS* , and *EvalSch–RS–Lyt*, to investigate the generalization of the models to links, room sizes, and both links and room sizes that are not included in the train set, respectively.

5.5.2.1 Evaluation scheme *EvalSch–Init*

The initial evaluation scheme *EvalSch–Init* is used to investigate the ability of predictive models to generalize to unknown link locations. Thus, the same fixed node locations and the same room sizes are applied for the training and testing, but the different portable nodes are used for testing; that results in different link locations and orientations for training and testing. The models are trained and tested on a single room size (S, M, L) or several room sizes (SM, ML, SL and SML) and apply the same fixed link layouts for training and testing (L1, L2, L3, ALL), but different location of the portable nodes for testing. Specifically, the evaluation settings in terms of the room sizes and link layouts included in the train and test set are defined as follows:

- ***1RS–train/1RS–test:*** Single room size for training and testing.

  Models are trained separately for each room size (S, M, and L), and tested on the same size, as follows S/S, M/M, and L/L.

  - ***1Lyt–train/1Lyt–test:*** Single link layout for training and testing.

    Models are trained using data acquired with L1, L2, and L3 layouts, separately; and the same layout is used for testing, as follows L1/L1, L2/L2, and L3/L3.
  - ***mLyt-train/mLyt-test:*** Multiple link layouts for training and testing.

    Models are trained using data acquired with several layouts (ALL), and the layouts used for training are also used for testing, thus the layouts in training and testing are ALL/ALL.

- ***2RS–train/2RS–test:*** Two room sizes for training and testing.

  Models are trained for pairs of room sizes (SM, ML, and SL), and the same room sizes are used for testing, as follows SM/SM, ML/ML, and SL/SL.

  - ***1Lyt–train/1Lyt–test:*** Single link layout for training and testing. Models are trained using data from single acquisition approach (L1, L2, and L3) and tested using data obtained with the same layout. Therefore, the layouts for training and testing are L1/L1, L2/L2, and L3/L3.
  - ***mLyt-train/mLyt-test:*** Multiple link layouts for training and testing.

    Models are trained using data from several acquisition approaches (ALL) and tested using data acquired with the same layouts, thus the layouts in training and testing are ALL/ALL.

- ***3RS–train/3RS–test:*** Three room sizes for training and testing.

  Models are trained for three room sizes (SML), and the same room sizes are used for testing the models. Thus, the room sizes in training and testing are SML/SML.

  - ***1Lyt-train/1Lyt-test:*** Single link layout for training and testing.

    Models are trained using data from single acquisition layout (L1, L2, and L3), and tested using data from the same layout, as follows L1/L1, L2/L2, and L3/L3.
  - ***mLyt-train/mLyt-test:*** Multiple link layouts for training and testing.

    Models are trained using data obtained with several acquisition approaches (ALL), and tested using data from the same layouts, i.e., the layouts for training and testing are ALL/ALL.

5.5.2.2 Evaluation scheme *EvalSch–Lyt*

The evaluation scheme *EvalSch–Lyt* is used to investigate the ability of predictive models to generalize to data acquired a with different approach (i.e., layout) than the one used for acquiring training data. This evaluation scheme represents the real scenario, where the model is making predictions on data acquired with different approach, without knowledge how the training data is obtained. Namely, the training data is acquired using one or several positions of fixed node, while portable node is moved randomly over the grid covering the room; and the model is applied on data using different positions of the fixed node. Different locations of fixed node, arranged in different layouts, as well as of the portable node are used for training and testing. Thus, the models are trained either with data of a single room size (S, M, L) or several room sizes (SM, ML, SL, SML) and evaluated with the same room sizes represented in the training process, but using the different link layouts for training and testing, namely L1/L2, L1/L3, L2/L1, L2/L3, L3/L1, L3/L2. Specifically, the evaluation settings in terms of the room sizes and link layouts included in the train and test set are defined as follows:

- ***1RS-train/1RS-test:*** Single room size for training and testing.

  Training a model on a single room size (S, M, and L) and evaluating the model on the same room size. In terms of the room sizes, the models are trained and tested as follows: S/S, M/M, L/L.

  - ***L1–train/diffLyt–test:*** L1 for training and different layout (L2, and L3) for testing.

    Training a model using data collected with a fixed node positioned in the center of the room (L1); and evaluating the model separately on data collected with fixed nodes in L2 positions on a circle centered in the room and data collected using nodes positioned near the corners (L3). The portable node is randomly moved over the grid in the room for collecting training and testing data. Consequently, the layouts for training and testing are L1/L2 and L1/L3.
  - ***L2–train/diffLyt–test:*** L2 for training and different layout (L1, and L3) for testing.

    Training a model using data acquired with eight fixed nodes positioned on a circle centred in the room (L2); and evaluating the model separately on data collected with a fixed node in the center of the room (L1) and data collected with fixed nodes placed close to the corners (L3). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Therefore, the layouts used for training and testing are L2/L1 and L2/L3.
  - ***L3–train/diffLyt–test:*** L3 for training and different layout (L1, and L2) for testing.

    Training a model using data obtained with fixed nodes positioned near the corners of the room (L3); and evaluating the model separately on data acquired with a fixed node in the center of the room (L1) and fixed nodes on a circle in the center of the room (L2). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Thus, the layouts for training and testing are L3/L1, and L3/L2.

- ***2RS–train/2RS–test:*** Two room sizes for training and testing.

  Training a model on a pair of room sizes (SM, ML, and SL) and evaluating the model on room sizes that are used in the training. The room sizes used for training and testing are SM/SM, ML/ML, and SL/SL.

- ***L1-train/diffLyt–test:*** L1 for training and different layout (L2, and L3) for testing.

  Training a model using data collected with a fixed node positioned in the center of the room (L1); and evaluating the model separately on data collected with fixed nodes in L2 positions on a circular pattern close to the center of the room and data collected using nodes positioned near the corners (L3). The portable node is randomly moved over the grid in the room for collecting training and testing data. Consequently, the layouts for training and testing are L1/L2 and L1/L3.

- ***L2-train/diffLyt–test:*** L2 for training and different layout (L1, and L3) for testing.

  Training a model using data acquired with fixed nodes positioned in eight positions on a circle centred in the room (L2); and evaluating the model separately on data collected with a fixed node in the center of the room (L1) and data collected with fixed nodes placed close to the corners (L3). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Therefore, the layouts used for training and testing are L2/L1 and L2/L3.

- ***L3-train/diffLyt–test:*** L3 for training and different layout (L1, and L2) for testing.

  Training a model using data obtained with fixed nodes positioned near the corners of the room (L3); and evaluating the model separately on data acquired with a fixed node in the center of the room (L1) and fixed nodes on a circle in the center of the room (L2). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Thus, the layouts for training and testing are L3/L1, and L3/L2.

- ***3RS–train/3RS–test:*** Three room sizes for training and testing.

  Training a model for three room sizes (SML) and evaluating the model on sizes that are considered in the training. The room sizes in training/testing are SML/SML.

  - ***L1–train/diffLyt–test:*** L1 for training and different layout (L2, and L3) for testing.

    Training a model using data collected with a fixed node positioned in the center of the room (L1); and evaluating the model separately on data collected with fixed nodes in L2 positions on a circular pattern close to the center of the room and data collected using nodes positioned near the corners (L2). The portable node is randomly moved over the grid in the room for collecting training and testing data. Consequently, the layouts for training and testing are L1/L2 and L1/L3.

  - ***L2–train/diffLyt–test:*** L2 for training and different layout (L1, and L3) for testing.

    Training a model using data acquired with fixed nodes positioned in eight positions on a circle centred in the room (L2); and evaluating the model separately on data collected with a fixed node in the center of the room (L1) and data collected with fixed nodes placed close to the corners (L3). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Therefore, the layouts used for training and testing are L2/L1 and L2/L3.

- ***L3–train/diffLyt–test:*** L3 for training and different layout (L1, and L2) for testing.

  Training a model using data obtained with fixed nodes positioned near the corners of the room (L3); and evaluating the model separately on data acquired with a fixed node in the center of the room (L1) and fixed nodes on on a circle in the center of the room (L2). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Thus, the layouts for training and testing are L3/L1, and L3/L2.

5.5.2.3 Evaluation scheme *EvalSch–RS*

The evaluation *EvalSch–RS* scheme is considered to investigate the ability of the trained models to generalize to data from unknown room sizes when the layout for obtaining CIR is represented in the training set. Including all possible room sizes in the training data may be impractical or impossible if the range of sizes is very large or if certain sizes are very rare. In such cases, it is important to select a subset of sizes that are representative of all the sizes that are likely to occur in the intended use case and to which the model is expected to generalize to. First, the ability of the models trained with data of single room size (S, M, L) to make predictions about the materials in rooms not included in training set is investigated. The usability of the model in larger rooms than the ones for which it is trained (S/ML), smaller rooms (L/SM), and both smaller and larger rooms (M/SL) is studied. The same location of link layout is applied for training and testing, but different portable nodes are used for testing. Next, expecting that only one size of the room is not sufficient to build a model, we train the model on two room sizes and test on the remaining room size not used in training. Thus, studying the generalization to larger rooms (SM/L), smaller rooms (ML/S), and rooms whose size lies between the smallest and the largest sizes included in the train set (SL/M). To summarize, in *EvalSch–RS* evaluation scheme, we evaluate the ability of the models to predict the materials based on CIR data of rooms with sizes different (S/ML, M/SL, L/SM, SM/L, ML/S, SL/M) from the sizes of the rooms represented in the training process, obtained with radio links in the same layout (L1/L1, L2/L2, L3/L3, ALL/ALL). Specifically, the valuation settings are:

- ***1RS–train/2RS–test:*** Single room size in train set and two room sizes in test set.

  Models are trained using data of single room size (S, M, and L), and evaluated on data from two room sizes not included in the training. Thus, the room sizes in training and testing are S/ML, M/SL, and L/SM.

  - ***1Lyt–train/1Lyt–test:*** Single link layout for training and testing.

    Models are trained using data acquired with L1, L2, and L3 layouts, separately; and the same layout is used for testing, as follows L1/L1, L2/L2, and L3/L3.

  - ***mLyt–train/mLyt–test:*** Multiple link layouts for training and testing.

    Models are trained using data acquired with several layouts (ALL), and the layouts used for training are also used for testing, thus the layouts in training and testing are ALL/ALL.

- ***2RS–train/1RS–test:*** Two room sizes for training and single room size for testing.

  Models are trained using data of two room sizes (SM, ML, and SL), and evaluated on data from one room size that is not included in the training. Therefore, the room sizes in training and testing are SM/L, ML/S, and SL/M.

- ***1Lyt–train/1Lyt–test:*** Single link layout for training and testing.

  Models are trained using data acquired with L1, L2, and L3 layouts, separately; and the same layout is used for testing, as follows L1/L1, L2/L2, and L3/L3.

- ***mLyt–train/mLyt–test:*** Multiple link layouts for training and testing.

  Models are trained using data acquired with several layouts (ALL), and the layouts used for training are also used for testing, thus the layouts in training and testing are ALL/ALL.

5.5.2.4 Evaluation scheme *EvalSch–RS–Lyt*

The *EvalSch–RS–Lyt* evaluation scheme is the most general and challenging among the considered evaluation alternatives, since both the layouts of the links and the sizes of the rooms are not represented in the training phase. This evaluation scheme examines the ability of the models to generalize to both unseen link layouts (L1/L2, L1/L3, L2/L1, L2/L3, L3/L1, L3/L2) and room sizes in train/test (S/ML, M/SL, L/SM, SM/L, ML/S, SL/M). Specifically, the evaluation settings are:

- ***1RS–train/2RS–test:*** Single room size for training and two room sizes for testing.

  Models are trained using data of single room size (S, M, and L), and evaluated on data from two room sizes not included in the training. Thus, the room sizes in training and testing are S/ML, M/SL, and L/SM.

  - ***L1–train/diffLyt–test*:** L1 for training and different layout (L2, and L3) for testing.

    Training a model using data collected with fixed node positioned in the center of the room (L1); and evaluating the model separately on data collected with fixed nodes in L2 positions on a circular pattern close to the center of the room and data collected using nodes positioned near the corners (L3). The portable node is randomly moved over the grid in the room for collecting training and testing data. Consequently, the layouts for training and testing are L1/L2 and L1/L3.

  - ***L2–train/diffLyt–test*:** L2 for training and different layout (L1, and L3) for testing.

    Training a model using data acquired with a fixed nodes positioned in eight positions on a circle centred in the room (L2); and evaluating the model separately on data collected with fixed node in the center of the room (L1) and data collected with fixed nodes placed close to the corners (L3). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Therefore, the layouts used for training and testing are L2/L1 and L2/L3.

  - ***L3–train/diffLyt–test*:** L3 for training and different layout (L1, and L2) for testing.

    Training a model using data obtained with fixed nodes positioned near the corners of the room (L3); and evaluating the model separately on data acquired with a fixed node in the center of the room (L1) and fixed nodes on a circle in the center of the room (L2). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Thus, the layouts for training and testing are L3/L1, and L3/L2.

- ***2RS–train/1RS–test:*** Two room sizes for training and single room size for testing.

Models are trained using data of two room sizes (SM, ML, and SL), and evaluated on data from one room size that is not included in the training. Therefore, the room sizes in training and testing are SM/L, ML/S, and SL/M.

- ***L1–train/diffLyt–test*:** L1 for training and different layout (L2, and L3) for testing.

  Training a model using data collected with a fixed node positioned in the center of the room (L1); and evaluating the model separately on data collected with fixed nodes in L2 positions on a circular pattern close to the center of the room and data collected using nodes positioned near the corners (L3). The portable node is randomly moved over the grid in the room for collecting training and testing data. Consequently, the layouts for training and testing are L1/L2 and L1/L3.

- ***L2–train/diffLyt–test*:** L2 for training and different layout (L1, and L3) for testing.

  Training a model using data acquired with fixed nodes positioned in eight positions on a circle centred in the room (L2); and evaluating the model separately on data collected with a fixed node in the center of the room (L1) and data collected with fixed nodes placed close to the corners (L3). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Therefore, the layouts used for training and testing are L2/L1 and L2/L3.

- ***L3–train/diffLyt–test*:** L3 for training and different layout (L1, and L2) for testing.

  Training a model using data obtained with fixed nodes positioned near the corners of the room (L3); and evaluating the model separately on data acquired with a fixed node in the center of the room (L1) and fixed nodes on a circle in the center of the room (L2). The portable node is randomly moved over the grid in the room for acquiring both the training and testing data. Thus, the layouts for training and testing are L3/L1, and L3/L2.

5.6 Description of the Data Sets

The data sets used in this study are divided into two groups based on their output:

- DS-1T – data sets with a single discrete target. Data from a subset of 48 rooms, 16 from each size, is included in these data sets. Concrete is used for the floor, ceiling and three walls, while the fourth wall is built using one material from a predefined set of materials, including brick, glass, plaster, and wood. This data is used for the experiments in the *Baseline Scenario*.

- DS-MT – data sets with multiple binary targets. Data from 11,250 rooms, 3,750 from each size, is included in these data sets. The rooms are built using different combinations of materials from a predefined set, including brick, concrete, glass, plaster, and wood. This data is used for the experiments in the *Comprehensive Scenario*.

In both DS-1T and DS-MT data sets, the descriptive space consists of 45 descriptive variables with continuous values. The selected features of the CIR for the 15 strongest MPCs, including received power, excess delay, and phase shift, are represented by the descriptive variables. The power values are given in logarithmic scale. In DS-1T data sets, the material of the different wall in the room is represented by the discrete target variable. Each CIR is associated with one material class from the four possible classes in DS-1T data sets. In DS-MT, the possible materials are target labels, and binary values for each of the material labels are associated with the CIRs to indicate the presence/absence of the material in the room where the CIR is obtained.

The name of the data sets is structured with three hyphen-separated parts. The first part represents the number of target attributes, either 1T denoting one discrete target or MT representing multiple binary targets. The second part denotes the size of the rooms where the data included in the data set is collected, ranging from a single room size (S, M, or L), two room sizes (SM, ML, or SL), or three room sizes (SML). The third part indicates the positions of the fixed node used for acquiring the data and may be a single layout (L1, L2, or L3) or ALL, indicating the use of all layouts for data collection.

The CIR data, acquired with the fixed node positioned in L1, L2, and L3 layout-specific locations and the portable node in 121 positions in each of 16 rooms with size S, 361 positions in each of 16 rooms with size M, and 729 positions in each of 16 rooms with size L, is stored in data sets 1T-S-ALL, 1T-M-ALL, and 1T-L-ALL. The data sets' description can be found in Table 5.7. Similarly, the CIR data obtained with the fixed node positioned in L1, L2, and L3 layout-specific locations and the portable node in 121 positions in each of 3,750 rooms with size S, 361 positions in each of 3,750 rooms with size M, and 729 positions in each of 3,750 rooms with size L, is stored in data sets MT-S-ALL, MT-M-ALL, and MT-L-ALL. The data sets' description can be found in Table 5.8. The product of the number of links and the number of rooms included in the data sets equals the number of instances. Table 5.6 describes the number of radio links in rooms with S, M, and L sizes.

Three new data sets are created from each data set described in Table 5.7 and Table 5.8, by splitting the original data set based on the layout that is used for collecting the data. Each of the new data sets is created to contain data from a particular room size (S, M, or L) that is collected using a specific layout (L1, L2, or L3). The description of these data sets is given in Table 5.9 and Table 5.10, respectively.

Table 5.7: Description of single-target data sets with CIR data acquired from S, M, and L rooms using all layouts. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the data set, *Layout* shows the layout used for acquiring the data in the data set, and *Room Size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layouts	Room size
1T-S-ALL	25168	ALL	S
1T-M-ALL	75088		M
1T-L-ALL	151632		L

Table 5.8: Description of multi-label data sets with CIR data acquired from S, M, and L rooms using all layouts. *Data set name* indicates the data set name, *Number of instances* shows the number instances in the data set, *Layout* shows the layout used for acquiring the data in the data set, and *Room Size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layout	Room size
MT-S-ALL	5898750	ALL	S
MT-M-ALL	17598750		M
MT-L-ALL	35538750		L

Table 5.9: Description of single-target data sets with CIR data from single room size acquired using single link layout. *Data set name* indicates the data set name, *Number of instances* shows the number instances in the data set, *Layout* shows the layout used for acquiring the data in the data set, and *Room Size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layout	Room size
1T-S-L1	1936	L1	S
1T-S-L2	15488	L2	
1T-S-L3	7744	L3	
1T-M-L1	5776	L1	M
1T-M-L2	46208	L2	
1T-M-L3	23104	L3	
1T-L-L1	11664	L1	L
1T-L-L2	93312	L2	
1T-L-L3	46656	L3	

Table 5.10: Description of multi-label data sets with CIR data from single room size acquired using single link layout. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the data set, *Layout* shows the layout used for acquiring the data in the data set, and *Room Size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layout	Room size
MT-S-L1	453750	L1	S
MT-S-L2	3630000	L2	
MT-S-L3	1815000	L3	
MT-M-L1	1353750	L1	M
MT-M-L2	10830000	L2	
MT-M-L3	5415000	L3	
MT-L-L1	2733750	L1	L
MT-L-L2	21870000	L2	
MT-L-L3	10935000	L3	

Two new data sets are created by selecting CIR instances based on the position of the portable node from each data set described in Table 5.9 and Table 5.10. A description of the new data sets created from the data sets in Table 5.9 is given in Table 5.12 and a description for the new data sets created from the data sets in Table 5.10 is given in Table 5.13. The procedure involves dividing each room into $1m^2$ regions and randomly selecting one location of the portable node from each region, resulting in 9 positions in S rooms, 25 positions in M rooms, and 49 positions in L rooms. The CIR instances obtained with the portable node positioned in the selected locations are used to create a new data set from the original data set. This same procedure is used to create the second data set. The first data set is used for training and tuning the models and the second data set is used for testing. The new data sets have the same number of instances and include links between fixed nodes in layout-specific locations and portable nodes in positions that uniformly cover the room. The number of instances in the new data sets is equivalent to the product of the number of links per room and the total number of rooms included in the data sets. The number of links in rooms with S, M, and L size is summarized in Table 5.11.

Table 5.11: Number of radio links between a fixed node in layout-specific locations and randomly selected locations for the portable node.

radio-link layout	room size		
	S	M	L
L1	9	25	49
L2	72	200	392
L3	36	100	196
Total number of radio-links per room:	**117**	**325**	**637**

Table 5.12: Description of the proposed single-target train/test sets with CIR data acquired from rooms with a single size. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the train/test set, *Layout* shows the layout used for acquiring the data in the train/test set, and *Room size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layouts	Room size
1T-S-L1-train/test	144/144	L1	S
1T-S-L2-train/test	1152/1152	L2	
1T-S-L3-train/test	576/576	L3	
1T-S-ALL-train/test	1872/1872	ALL	
1T-M-L1-train/test	400/400	L1	M
1T-M-L2-train/test	3200/3200	L2	
1T-M-L3-train/test	1600/1600	L3	
1T-M-ALL-train/test	5200/5200	ALL	
1T-L-L1-train/test	784/784	L1	L
1T-L-L2-train/test	6272/6272	L2	
1T-L-L3-train/test	3136/3136	L3	
1T-L-ALL-train/test	10192/10192	ALL	

Table 5.13: Description of the proposed multi-label train/test sets with CIR data acquired from rooms with a single size. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the train/test set, *Layout* shows the layout used for acquiring the data in the train/test set, and *Room size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layouts	Room size
MT-S-L1-train/test	33750/33750	L1	S
MT-S-L2-train/test	270000/270000	L2	
MT-S-L3-train/test	135000/135000	L3	
MT-S-ALL-train/test	438750/438750	ALL	
MT-M-L1-train/test	93750/93750	L1	M
MT-M-L2-train/test	750000/750000	L2	
MT-M-L3-train/test	375000/375000	L3	
MT-M-ALL-train/test	1218750/1218750	ALL	
MT-L-L1-train/test	183750/183750	L1	L
MT-L-L2-train/test	1470000/1470000	L2	
MT-L-L3-train/test	735000/735000	L3	
MT-L-ALL-train/test	2388750/2388750	ALL	

For scenarios where models are trained or tested with data from two or three room sizes, data sets are created by merging data from individual room sizes. Table 5.14 and Table 5.15 describe the single-target and multi-label data sets with data from two room sizes, used for training and testing. Similarly, Table 5.16 and Table 5.17 describe the single-target and multi-label data sets with data from three room sizes, respectively.

Table 5.14: Description of the proposed single-target train/test sets with CIR data acquired from rooms with two sizes. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the train/test set, *Layout* shows the layout used for acquiring the data in the train/test set, and *Room size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layouts	Room size
1T-SM-L1-train/test	544/544	L1	SM
1T-SM-L2-train/test	4352/4352	L2	
1T-SM-L3-train/test	2176/2176	L3	
1T-SM-ALL-train/test	7072/7072	ALL	
1T-ML-L1-train/test	1184/1184	L1	ML
1T-ML-L2-train/test	9472/9472	L2	
1T-ML-L3-train/test	4736/4736	L3	
1T-ML-ALL-train/test	15392/15392	ALL	
1T-SL-L1-train/test	928/928	L1	SL
1T-SL-L2-train/test	7424/7424	L2	
1T-SL-L3-train/test	3712/3712	L3	
1T-SL-ALL-train/test	12064/12064	ALL	

Table 5.15: Description of the proposed multi-label train/test sets with CIR data acquired from rooms with two sizes. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the train/test set, *Layout* shows the layout used for acquiring the data in the train/test set, and *Room size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layouts	Room size
MT-SM-L1-train/test	127500/127500	L1	SM
MT-SM-L2-train/test	1020000/1020000	L2	
MT-SM-L3-train/test	510000/510000	L3	
MT-SM-ALL-train/test	1657500/1657500	ALL	
MT-ML-L1-train/test	277500/277500	L1	ML
MT-ML-L2-train/test	2220000/2220000	L2	
MT-ML-L3-train/test	1110000/1110000	L3	
MT-ML-ALL-train/test	3607500/3607500	ALL	
MT-SL-L1-train/test	217500/217500	L1	SL
MT-SL-L2-train/test	1740000/1740000	L2	
MT-SL-L3-train/test	870000/870000	L3	
MT-SL-ALL-train/test	2827500/2827500	ALL	

Table 5.16: Description of the single-target train/test sets with CIR data acquired from rooms with three sizes. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the train/test set, *Layout* shows the layout used for acquiring the data in the train/test set, and *Room size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layouts	Room size
1T-SML-L1-train/test	1328/1328	L1	SML
1T-SML-L2-train/test	10624/10624	L2	
1T-SML-L3-train/test	5312/5312	L3	
1T-SML-ALL-train/test	17264/17264	ALL	

Table 5.17: Description of the proposed multi-label train/test sets with CIR data acquired from rooms with three sizes. *Data set name* indicates the data set name, *Number of instances* shows the number of instances in the train/test set, *Layout* shows the layout used for acquiring the data in the train/test set, and *Room size* shows the size of the rooms where the data is collected.

Data set name	Number of instances	Layouts	Room size
MT-SML-L1-train/test	311250/311250	L1	SML
MT-SML-L2-train/test	2490000/2490000	L2	
MT-SML-L3-train/test	1245000/1245000	L3	
MT-SML-ALL-train/test	4046250/4046250	ALL	

According to the experimental scenario, which specifies the number of targets (single-target or multi-label), and the evaluation scheme, which describes the room sizes and link layouts represented in the train and test sets, the train/test set is selected. For instance, in the *Comprehensive Scenario* when the model is trained using data from rooms with S size acquired with layout L1 and tested using data from M and L rooms obtained with layout L2, i.e., S-L1/ML-L2, the MT-S-L1-train set, described in Table 5.13, is used for training the model, and MT-ML-L2-test set, described in Table 5.15, is used for testing the model.

In the experiments of the *Baseline Scenario*, data from an adequate data set from the data sets described in Table 5.12, Table 5.14, or Table 5.16 is used when one, two, or three room sizes are considered for training and testing, respectively. On the other hand, in the experiments of the *Comprehensive Scenario*, when considering the *Eval–Init* and *Eval–Lyt* evaluation schemes, data from an adequate data set from Table 5.13, Table 5.15, or Table 5.17 is applied for training and testing the models with data from one, two, or three room sizes. In the evaluation schemes *Eval–RS* and *Eval–RS–Lyt*, data from an adequate data sets from Table 5.13 is used when a single room size is considered for training or testing, and from Table 5.15 when two room sizes are represented in training or testing.

5.7 Evaluation Metrics

To quantitatively measure the performance of a classification model in predicting outcomes, multiple performance measures (metrics) can be calculated. When evaluating models, other criteria are also important, such as computational cost, efficient use of the machine resources, and comprehensibility [34].

In this dissertation, we focus on performance metrics. The performance metrics are calculated based on the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These values are obtained based on the classes that the classification model has predicted against the actual classes associated with the data sample. An overview of the predicted versus true classes is presented in the confusion matrix. An example of a confusion matrix for binary classification is shown in Table 5.18. TP indicates the number of samples correctly predicted in the first (positive) class, while TN indicates the number of samples correctly predicted in the negative class (or all other classes in cases with more than two classes). On the other hand, the FP denotes the number of data samples incorrectly classified as positive, and FN denotes the number of samples belonging to the positive class but incorrectly classified as negative (or as any other class in the cases with more than two classes).

Table 5.18: Confusion matrix for binary classification.

		Predicted class	
		P (Positive)	N (Negative)
True class	P (Positive)	TP (True Positive)	FN (False Negative)
	N (Negative)	FP (False Positive)	TN (True Negative)

The F1 score is commonly used for classification tasks [34], it combines precision

$$Precission = \frac{TP}{TP + FP}, \tag{5.13}$$

and recall

$$Recall = \frac{TP}{TP + FN} \tag{5.14}$$

in single metric as

$$F1 = \frac{2 * Precission * Recall}{Precission + Recall}. \tag{5.15}$$

The binary metrics are extended for multi-class cases by considering each class separately and constructing a binary matrix for each class using the one-vs-rest approach. The metrics defined for binary classification, when extended to multi-class problems, represent metrics for only one class. All the per-class metrics jointly give a detailed insight into the classification performance of the models. Usually, a single value is used to characterize the classification results. For that purpose, the per-class metrics are averaged over all classes. Similarly, the metric defined for binary classification can be extended to multi-label problems [41]. In this case, the binary metrics represent per-label metrics. The final score is calculated by averaging per-label metrics over all labels.

The averaging can be done in a micro, macro, or weighted manner. The difference between the averaging strategies is emphasized for imbalanced classification. When the data set is balanced, all three strategies give a similar result. Macro-averaged metrics are unweighted mean of the metric across all classes, meaning that each class is represented equally in this metric. By macro-averaging, the smaller classes, which usually have lower scores, influence the final score to a great extent. Weighted averaging accounts for class imbalance by computing the metric average by assigning weight to each class that correlates with the number of samples of each class. The last strategy, micro-averaging, performs averaging on the instance level, with each instance contributing equally to the final score, which makes it similar to accuracy calculation. Therefore, micro-averaged metrics have the same drawbacks as accuracy when dealing with imbalanced datasets. As a result, since we do not want the score to be dominated by a few large classes macro-averaged metrics are preferred when presenting indoor material prediction results, more specifically, macro F1 score. For calculating the macro F1 score we adhere to the approaches recommended in [210].

Chapter 6

Results and Discussion

The discussion of the experimental results is divided into two parts. The first part discusses the results of the *Baseline Scenario*, which focuses on (i) examining the feasibility of the concept, (ii) investigating specific and general approaches to train the model in terms of the room sizes in which the training data is collected and the acquisition layout used, and (iii) assessing the capability of different learning algorithms to extract knowledge from CIR and building material identification models.

The second part discusses the results from the *Comprehensive Scenario*, where the focus is on (i) studying the ML-based material prediction in rooms with different combinations of materials and (ii) investigating the capability of models to generalize to new locations of the portable node, link layouts used to estimate CIR, room sizes where CIR is acquired, and both room sizes and link layouts.

The analysis of the results provides insights into the (i) capability of ML models to predict the material of a single wall, as well as all materials in the room, (ii) suitability of specific and general approaches in terms of the links and room sizes incuded in the train set for building a model, and (iii) impact of the room size and position of the radio links used for acquiring CIR data on the generalization performance of the models.

6.1 Results from Baseline Scenario

The results from the initial experimentation, where the learning task is to predict the material of a single surface in a room from CIR, are presented and discussed in this section.

6.1.1 Evaluation setting *1RS–train/1RS–test*

We present and discuss the results of models trained using data from a single room size; and we examine the impact of the room size and the layout-specific location(s) of the fixed node relative to the room geometry on the model's performance. The results in terms of macro-F1 of models trained with data from one room size are shown in Figure 6.1. In particular, it shows the scores of models evaluated using the *EvalSch–Init* evaluation scheme; trained and tested using single room size, considering the *1RS–train/1RS–test* evaluation setting (S/S, M/M, and L/L), and different link layouts, with evaluation settings *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3), and *mLyt–train/mLyt–test* (ALL/ALL).

The results are showing that scores above 0.9 for all room sizes are achieved by most of the models, regardless of the layout used (L1, L2, L3, or ALL), with the exception of the DT models for all three room sizes when L3 is used, and RF and SVM models for S size when L1 is considered. Slight variations in the performance of the models can be noticed when different layouts are used. The differences are more pronounced when using L1 or

L3 compared to when L2 or ALL are used. Specifically, when L2 is used, the performance of the models is similar to when ALL layouts are applied.

Most models attain scores above 0.9 when using the L1 layout, with the exception of RF and MLP which have scores ranging from 0.8 to 0.9 when using S rooms. Similar performance is exhibited by SVM models irrespective of the room size. Slight variations are observed across different room sizes for DT, RF, and MLP models, with similar performances for M and L sizes (above 0.9) and lower scores (approximately 0.8) for S room size.

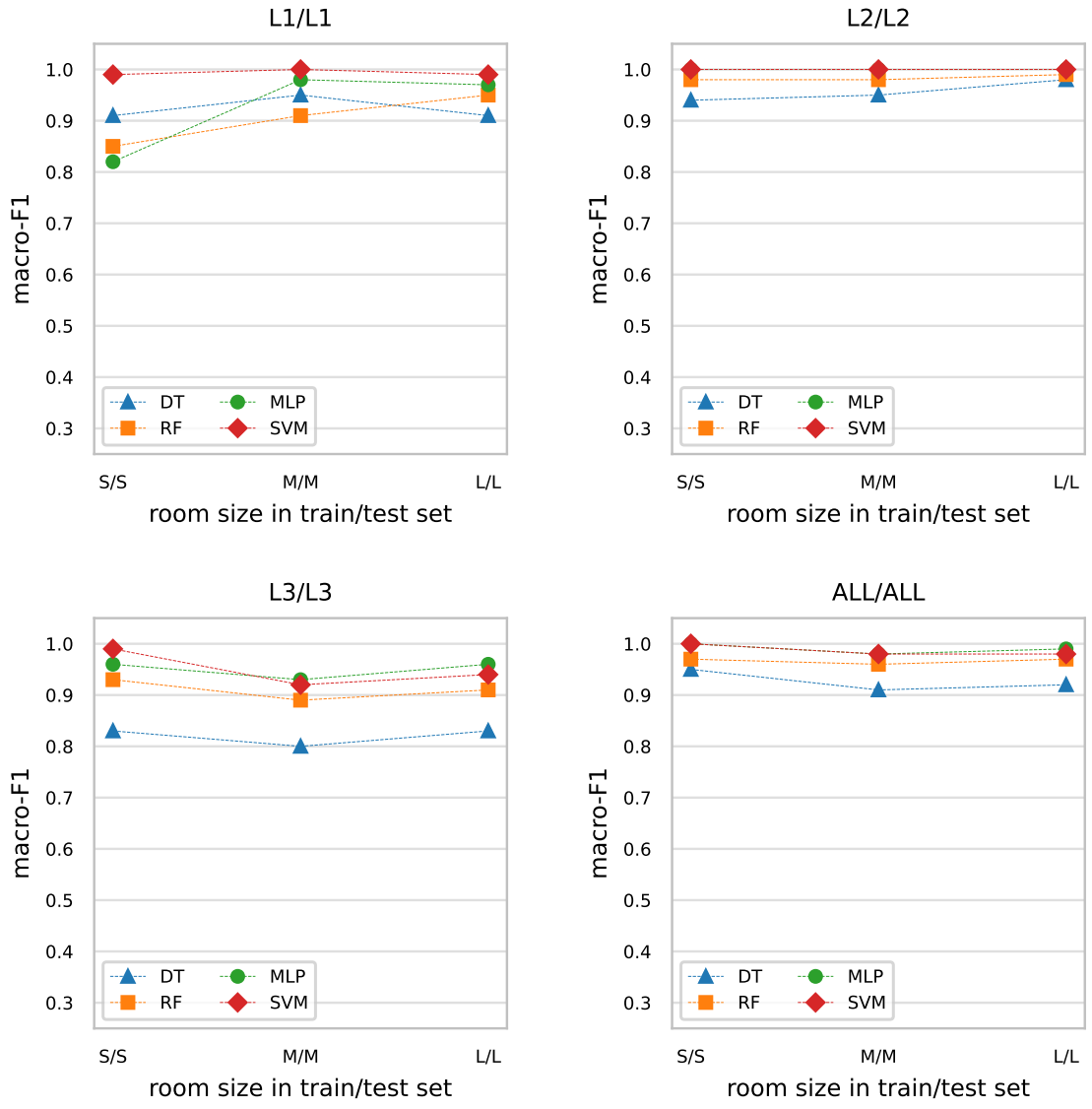

Figure 6.1: Predictive performance of models in *Baseline Scenario* trained/tested on single room size for different link layouts. The evaluation scheme is *EvalSch–Init*. The evaluation settings are *1RS–train/1RS–test* (S/S, M/M, and L/L) in terms of room size in train/test set, and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

When the L2 layout is employed, similar scores are observed across all models (above 0.9), regardless of the room size, with the exception of DT models exhibiting slightly lower

scores compared to other models. A minor distinction can be discerned in the case of DT models. Specifically, when the L2 layout is employed, the scores exhibit a decrease as the size of the room is smaller (lowest scores for S rooms and highest for L rooms). Conversely, when ALL layouts are taken into account, the DT models exhibit the highest scores for S rooms and comparable, albeit slightly lower, scores for M and L rooms, due to a larger number of links used.

The scores for all room sizes exhibit a resemblance in the settings where L3 is applied, although a minor decline is observed for M rooms. MLP and SVM models demonstrate highly comparable scores, whereas RF models exhibit lower scores compared to them, for all room sizes. Lower scores are noticed for DT models as compared to the other models, for all room sizes.

The results confirm that indoor material prediction can be tackled with the room size specific modeling approach and single layout (L1, L2, or L3), or several layouts (ALL) approaches in terms of the layout used for collecting the training CIR data. Considering multiple positions for the fixed node spatially distributed around the room is preferred for building more general models that can be applied to data from different acquisition strategies. A small number of fixed node-positions is a good choice when the resources are limited and shorter training time is needed.

Based on the results, it can be inferred that the size of the room, and the layout-specific location(s) of the fixed node relative to the room have an impact on the performance of the models. Therefore, the placement of fixed nodes should be carefully considered when developing a model for indoor material prediction. In particular, when using the L1 layout, with a single position of the fixed node and thus a small number of links, to obtain CIR, the effect of the room size is more pronounced compared to when L2, or ALL layouts, both with several positions of the fixed node and thus, a large number of links, is used. A similar conclusion is drawn for the case when L3 is considered.

The learning method also has an impact on the performance of the models. This effect is emphasized when L1 or L3 are used, thus the number of links is small or the location of the fixed node is challenging. However, when the fixed node is placed in L2-specific locations, due to multiple positions, placed centrally and thus symmetrically to the room geometry the effect of the room size and algorithm are very small. Similarly, when ALL layouts are used, due to many locations of the fixed node (one central, several on a circle centered in the room, and several close to the corners) and consequently a large number of links, the impact of both, the room size and learning method, is small.

6.1.2 Evaluation setting *2RS–train/2RS–test*

The results of models trained using data from two room sizes are presented; and the effect of combining two room sizes on model's performance is analysed. The results in terms of macro-F1 are shown in Figure 6.2. Specifically, the figure shows the scores of models evaluated using the *EvalSch–Init* evaluation scheme; trained and tested using two room sizes, considering the *2RS–train/2RS–test* evaluation setting (SM/SM, ML/ML, and SL/SL), and different link layouts, with evaluation settings *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3), and *mLyt–train/mLyt–test* (ALL/ALL).

The performance scores manifest greater variations with the usage of L1 or L3 layouts in comparison to L2 or ALL layouts, regardless of the combined room sizes. It is notable that the utilization of L2 layout results in similar scores across all models.

In the cases where L1 is employed for both training and testing, the RF, SVM, and MLP models demonstrate scores exceeding 0.9 for all room size pairs. The MLP and SVM models exhibit similar scores, whereas the RF model exhibits slightly lower scores ranging between 0.8 and 0.9. On the other hand, the DT models exhibit the lowest performance

with scores being the lowest for SM room sizes, and exhibiting slight improvements upon combining L rooms with S and M rooms.

In the context of utilizing L2, all models achieve high performance scores, around 0.95 or above, and exhibit similar performance across all room size settings (SM, ML, and SL), except for DT models which achieve slightly lower scores. Similarly, high scores exceeding 0.9 are obtained with minimal performance differences among the models when considering ALL layouts. However, these differences are more noticeable compared to the cases where L2 is used. Regarding the ML and SL sizes, SVM, RF, and MLP models demonstrate similar scores, when ALL layouts are considered. In contrast, when SM sizes are utilized, the SVM models exhibit comparable scores to the DT models, whereas the RF and MLP models display higher scores, with MLP achieving the best performance scores.

In settings where L3 is employed, the differences in performances among the models are more pronounced. MLP and RF models attain scores exceeding 0.9 across all size settings, whereas SVM models demonstrate such scores only when ML and SL are taken into account and lower scores similar to DT models when SM is used. DT models exhibit the lowest scores, approximately 0.8, across all room size settings.

The results show that the material of a single surface can be predicted with models trained with two room sizes. Thus, they confirm that the indoor material prediction can be approached more generally in terms of the room sizes included in the train set. Training one model per room size-pair is preferred over training several separate models, one per each size, in cases where the model is intended to make predictions for rooms in both sizes.

In terms of the impact of the location(s) of the fixed node, relative to the room, on models' performances; similar conclusions are drawn as in the *1RS–train/ 1RS–test* evaluation setting. Specifically, when L2 or ALL layouts are employed, the variations in scores are small across different room size settings and learning methods, whereas the variations in scores are more noticeable, compared to the former when L1 or L3 are used.

With respect to the impact of joining data from different room sizes, the results show small variations in the scores for different pairs of room sizes. These results indicate that when the models are evaluated in rooms with sizes represented in the training process, the effect of the different sizes on the models' performance is not pronounced. However, we think that in more challenging evaluation scenarios, where the models are expected to generalize to sizes that were not represented in the training process; the impact of the room sizes included in the training set is emphasized. To investigate the impact of the room sizes considered for training on the capability of models to generalize to new, i.e., not represented in the train set, sizes, we extend our experiments and we discuss the results in Section 6.2.

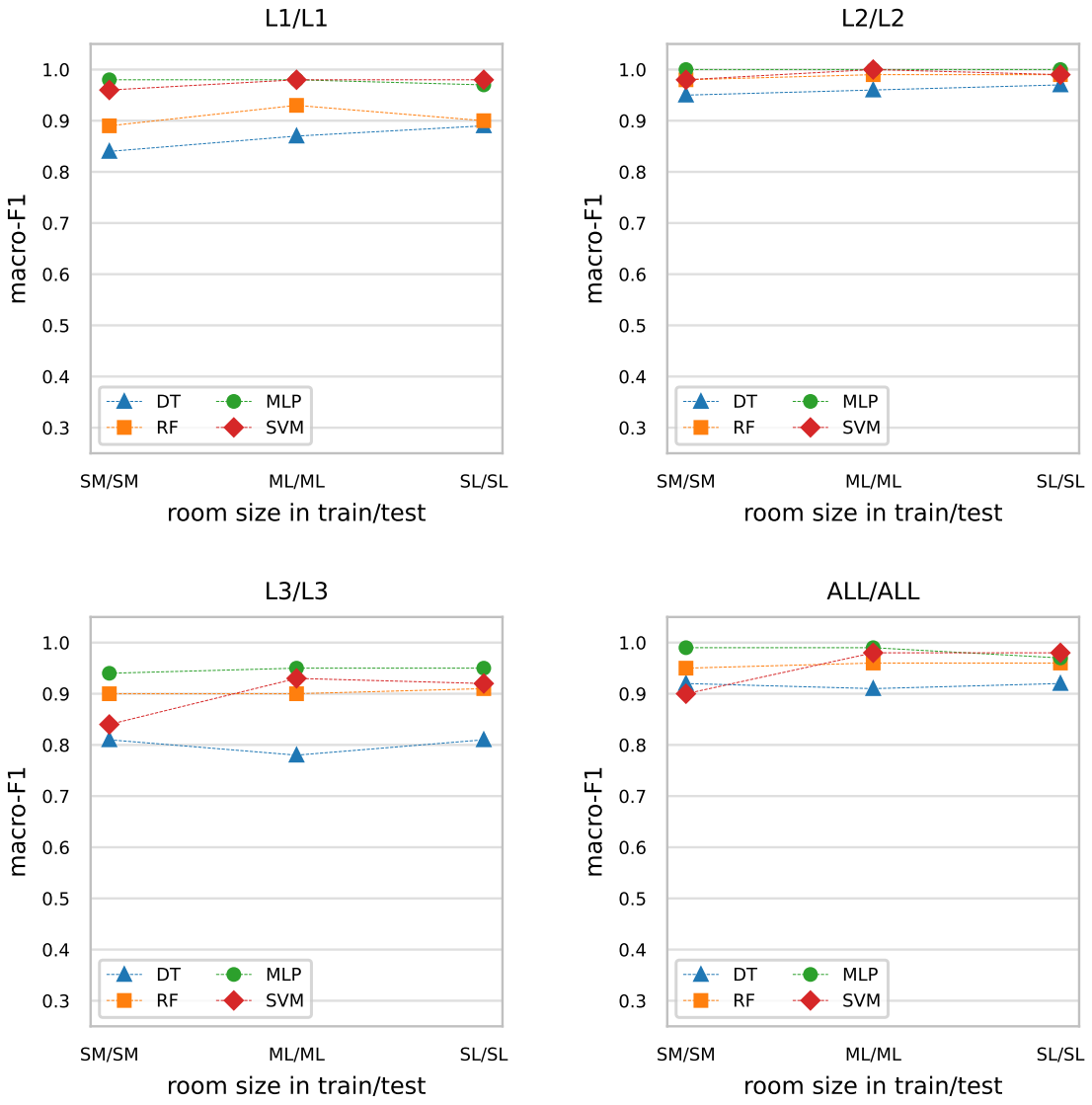

Figure 6.2: Predictive performance of models in *Baseline Scenario* trained/tested on two room sizes for different link layouts. The evaluation scheme is *EvalSch–Init*. The evaluation settings are *2RS–train/2RS–test* (SM/SM, ML/ML, and SL/SL) in terms of room size in train/test set, and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

6.1.3 Evaluation setting *3RS–train/3RS–test*

The results of models trained using data from three room sizes are presented; and the effect of joining several different sizes on the model's performance is analysed. The results in terms of macro-F1 are shown in Figure 6.3. Specifically, the figure shows the scores of models evaluated using the *EvalSch–Init* evaluation scheme; trained and tested using three room sizes, considering the *3RS–train/3RS–test* evaluation setting (SML/SML), and different link layouts, with evaluation settings *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3), and *mLyt–train/mLyt–test* (ALL/ALL).

The results indicate that performance scores above 0.9 are achieved by most of the

models; MLP, SVM, and RF models, regardless of the layout employed for training and testing, and DT models when L2 or ALL layouts are applied. However, exceptions are observed in the case of DT models when L1 or L3 are used, wherein scores of approximately 0.85 and 0.78 are attained, respectively. Similar scores are achieved by MLP, SVM, and RF models when L1 is utilized, with the highest scores being achieved by MLP models, while a lower score (approximately 0.85) is obtained by the DT model. When employing L3 layout or ALL layouts, almost identical scores are attained by MLP and SVM models, with the SVM model exhibiting slightly inferior performance. The DT models demonstrate the lowest scores among the other models when both L3 and ALL layouts are considered. When considering ALL layouts, the DT models achieve higher scores (approximately 0.9) compared to when L3 is employed, where they obtain lower scores (approximately 0.78).

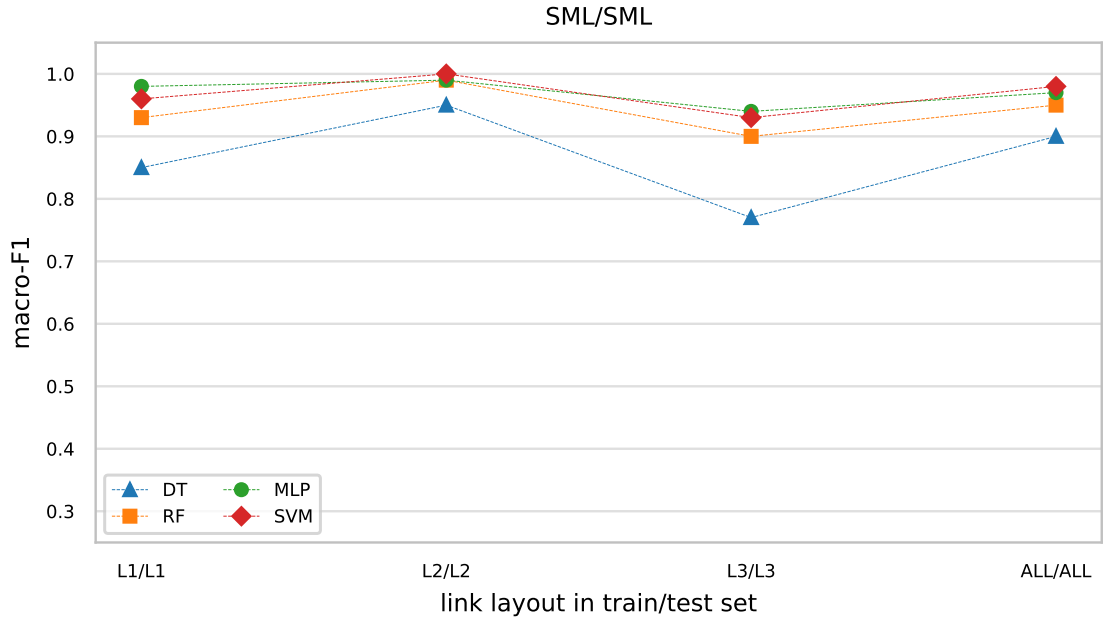

Figure 6.3: Predictive performance of models in *Baseline Scenario* trained/tested on three room sizes for different link layouts. The evaluation scheme is *EvalSch–Init*. The evaluation settings are *3RS–train/3RS–test* (SML/SML) in terms of room size in train/test set, and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

The results show that the general approach for training a model, previously being discussed in terms of two room sizes, can be extended to multiple room sizes. This approach is adequate for scenarios where the model is intended to predict the material in several room sizes, because one model is trained instead of separate models for each of the room sizes where the model will be applied.

In terms of the link layout used for data acquisition, variations of the performance of the models for different layouts is noticed. Similarly, as being concluded when the evaluation settings *1RS–train/1RS–test* and *2RS–train/2RS–test* are considered; the variations in the scores are more evident in the cases where L1, or L3 is used, compared to the cases where L2, or ALL layouts are used. Regarding the learning method, the results indicate that MLP, SVM, and RF models achieved similar scores for specific layouts, while the DT model yielded lower scores.

6.1.4 Summary

The results from the initial experimentation confirm (i) that the material of a single wall in basic indoor environments can be predicted from CIR using ML models, and (ii) that both specific and general approaches, in terms of the link layout used for CIR acquisition and the room sizes where the CIR is obtained, for learning a model can be considered. The feasibility of the proposed concept is confirmed with the results from the *Baseline Scenario*, while also foundational understanding regarding the utilization of ML approaches for addressing the problem of indoor material identification based on CIR.

Specifically, the results show that the material of a single wall in a room can be identified from CIR with ML models. With respect to the selection of room geometry (size) and the positions of the radio nodes for collecting CIR data to train the model, the results suggest that both specific and general approaches can be considered. The use case where the model is intended to make predictions, i.e., (1) the indoor environment characterization scenario in terms of the: (i) size of the rooms where the material is identified, and (ii) positions of the radio nodes used for CIR acquisition, and (2) the available resources in terms of data, computing power, and time, has to be considered when selecting a training setting.

In scenarios where the model must be optimized to identify materials in rooms of a particular size, the specific approach is deemed appropriate. Conversely, in cases where multiple room sizes are considered, a general approach is more suitable. This involves the construction of a single model that can address all room sizes, rather than creating separate models for each room size. When single CIR acquisition strategy is used for collecting CIR data for training the models and identifying the materials in new rooms, a specific approach, i.e., including data from one aquisition strategy in the train set is considered adequate. Conversely, when the model is intended to be applied on data obtained with various acquisition strategies, a general approach, i.e., representing several CIR acquisition strategies in the train set, is more suitable. Finally, in cases where the model is applied to a single room size, but to data from several acquisition strategies, including data from a single room size and multiple link layouts is appropriate. For those scenarios where the model is applied for material characterization of rooms with several sizes but the acquisition strategy is always the same, the train set should contain data from many room sizes obtained with the particular acquisition strategy.

According to the results obtained, the performance of the model is affected by the link layout selected for data collection. The layout comprises several factors, including the number of positions assigned to both the fixed and portable nodes, as well as the location of the fixed node with respect to the room geometry. The impact of these factors on the model's performance is pronounced for small rooms where the portable node is required to be moved around a smaller grid. Furthermore, the model's performance is affected when the fixed node is positioned in challenging locations, such as near the corners of the room, or when it is placed in a single position.

The findings from this initial experimentation set the basis for advanced experimentation. Specifically, the identification of the materials of all surfaces in a room using CIR is studied and the generalization capabilities of the models are assessed in more detail by extending the experiments. This investigation is conducted for rooms with various combinations of materials, and the generalization performance is evaluated through distinct evaluation schemes that are defined based on the room size(s) and the link layout used for data acquisition.

6.2 Results from *Comprehensive Scenario*

The results from the advanced experimentation, where the learning task is to predict the materials of all the surfaces in a room from CIR, are presented and discussed in this section. The generalization performance of the models in four evaluation schemes: (i) *EvalSch–Init*, (ii) *EvalSch–Lyt*, (iii) *EvalSch–RS*, and (iv) *EvalSch–RS–Lyt* is discussed in terms of the capabilities of the models to generalize to new CIR data when (i) the position of the portable node, (ii) both the position of the fixed and the portable node (i.e., the CIR-acquisition layout), (iii) the room sizes, and (iv) both the CIR-acquisition layout and the room sizes utilized for obtaining the test data are different than those for obtaining the train data, respectively.

6.2.1 Evaluation scheme *EvalSch–Init*

The materials identification in the case when the portable node for acquiring CIR data is moved in different positions over a grid that covers the room than the positions used for training the model is discussed. Thus, we analyse the generalization performance of the models when the test set is obtained with portable nodes in different positions than those used for obtaining the train set. The results from the following evaluation settings: *1RS–train/1RS–test*, *2RS–train/2RS–test*, and *3RS–train/3RS–test* in terms of the room sizes included in the train and test set, as well as *1Lyt–train/1Lyt–test*, and *mLyt–train/mLyt–test* in terms of the link layout included in the train and test set, are presented; and the impact of the room size, link layout as well as joining data of several room sizes and link layouts on the generalization performance is analysed.

The results in terms of macro-F1 of models trained with data obtained from a single room size are shown in Figure 6.4. In particular, the figure shows the scores of the models evaluated using *Eval–Init* evaluation scheme; trained and tested using a single room size, considering *1RS–train/1RS–test* evaluation setting (S/S, M/M, and L/L), and different link layouts, with evaluation settings *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL).

When L1 or L2 link layouts are employed, similar scores are obtained for M and L rooms, but S rooms exhibit slightly lower scores. This discrepancy is more pronounced in the case of the L1 link layout where the fixed node is in a single position, compared to the L2 layout where several fixed node positions are considered. The results indicate that in S rooms, where the portable node follows a smaller grid, considering several positions for the fixed node as in L2, as opposed to a single position in L1, can enhance the models' performance. Conversely, in rooms with a larger grid, such as M and L in comparison to S rooms, the performance of models trained with L1 is similar to that of models trained with L2. In the case of L3, the scores are similar for all room sizes, although with a slight decline in performance for RF, DT, and SVM models in L rooms compared to S and M rooms. This demonstrates that the performance in L rooms is marginally influenced by the distance between the fixed node positions in L3 and the walls. When considering all layouts, the scores of the MLP, DT, and SVM models are similar regardless of the room size. Small variation in the performance is discernible for RF models which achieve lower scores in S and L rooms compared to M rooms.

The scores of the models trained with different algorithms range from 0.7 to 1.0. Specifically, in all evaluation settings, the MLP models achieve scores exceeding 0.9, while the RF and DT models attain scores ranging from 0.8 to 0.9, exhibiting nearly identical performance for all room sizes when L2 is employed, and for S and L rooms when ALL layouts are employed. The SVM models demonstrate the lowest performance among the considered algorithms, with scores between 0.7 and 0.8, in all settings.

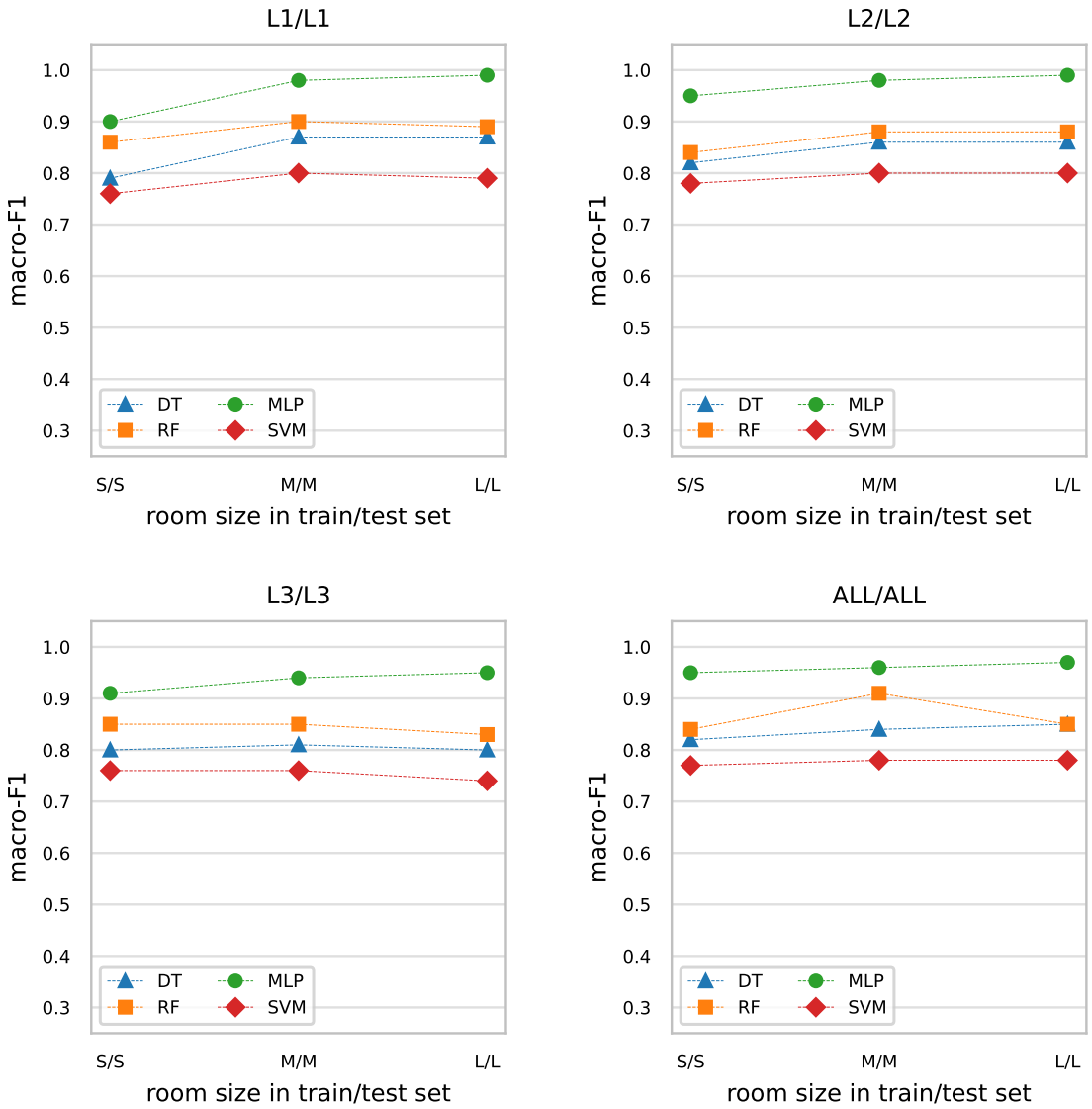

Figure 6.4: Generalization performance of models trained on a single room size to new locations of the portable node. The evaluation scheme is *Eval–Init*. The evaluation settings are *1RS–train/1RS–test* (S/S, M/M, and L/L) in terms of room size in train/test set, and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

The results confirm that the models trained with a single room size can generalize to new positions of the portable node when the room sizes and the positions of the fixed node are the same for training and testing. Therefore, a model that is trained for a single room size, using CIR data acquired with fixed nodes in layout-specific positions and portable nodes moved over a grid covering the room is able to predict the materials of all surfaces in new rooms with the same size as the size of the rooms used for training, based on CIR data obtained with fixed nodes in the same positions as for training and portable node moved in different positions over a grid. When the size of the room is small, the joint impact of the room size and the link layout on the performance of the model becomes evident. Specifically, the grid over which the portable node moves is small, thereby affecting the

model's performance. This effect is more pronounced when the fixed node is positioned in a single location. This can be noticed from the scores of the models trained with data from S rooms obtained using L1 link layout.

The impact of joining data from different pairs of rooms on the performance of the models is discussed based on the results shown in Figure 6.5. In particular, the figure shows the macro-F1 scores of the models evaluated using *Eval–Init* evaluation scheme; trained and tested using two room sizes, considering *2RS–train/2RS–test* evaluation setting (SM/SM, ML/ML, and SL/SL), and different link layouts, with evaluation settings *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL).

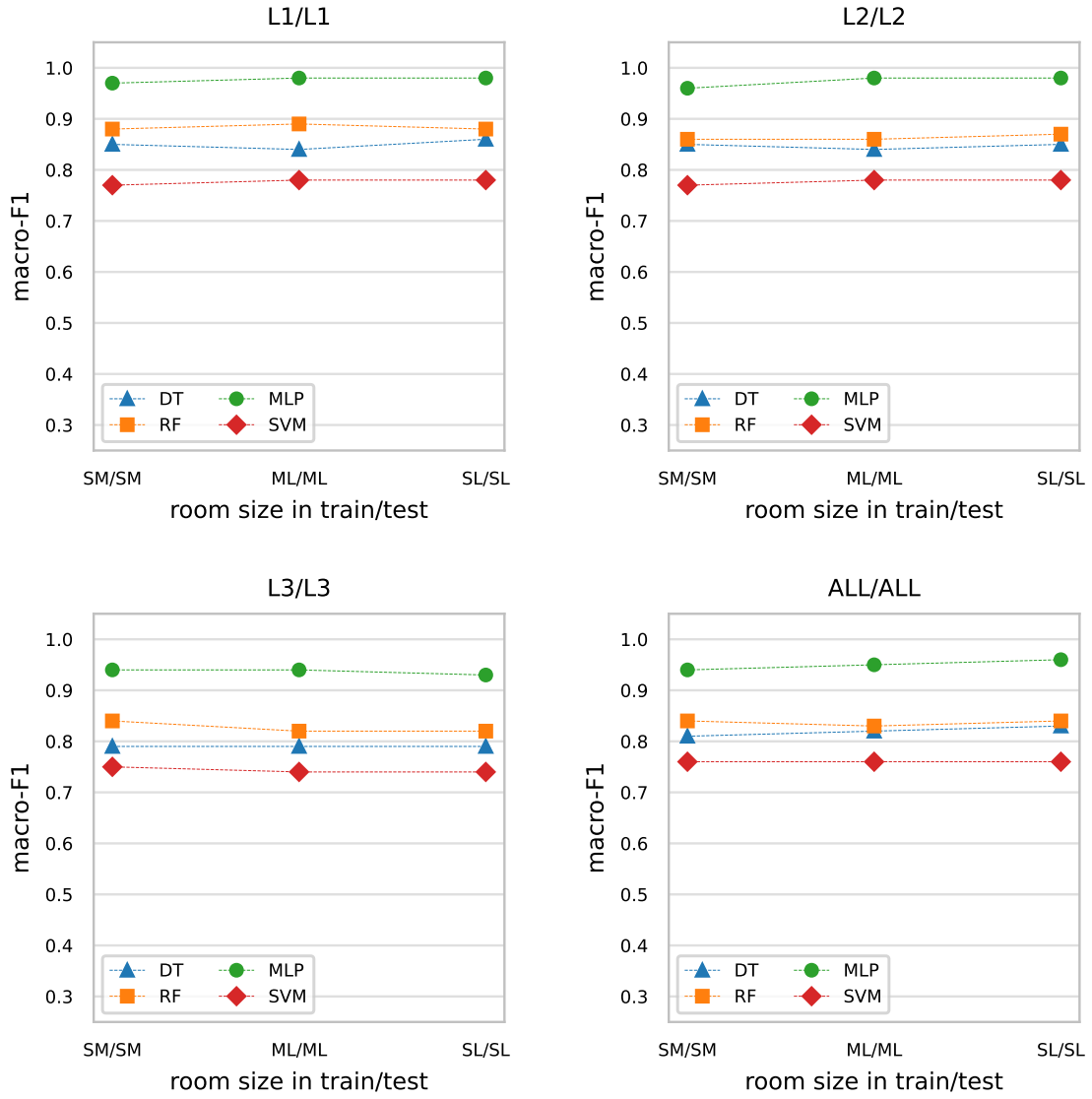

Figure 6.5: Generalization performance of models trained on two room sizes to new locations of the portable node. The evaluation scheme is *Eval–Init*. The evaluation settings are *2RS–train/2RS–test* (SM/SM, ML/ML, and SL/SL) in terms of room size in train/test set, and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

The results demonstrate that the models trained using data acquired with L1 exhibit

comparable performance to those trained using data obtained with L2, while the models trained with L3 data show slightly lower performance, regardless of the room size pairs utilized. It indicates that the challenging position of the fixed nodes in L3 (near the room corners), compared to central positions in L1 (single) and L2 (several on a circle), affects the performance of the models. The combined use of ALL link layouts yields models with comparable performance to those trained using individual link layouts. This finding confirms that the general approach for training the model can be considered. It can be inferred that the performances of the models trained on different pairs of room sizes did not exhibit significant differences due to the evaluation of the models on data from the same pair of room sizes used in the training phase. Noticeable variations in performance scores across different learning methods can be observed. Specifically, MLP models demonstrate the best performance with scores above 0.9. RF and DT models exhibit very similar performances (particularly when L2 or ALL link layouts are employed), with scores ranging between 0.8 and 0.9. On the other hand, SVM models show the lowest performance with scores ranging between 0.7 and 0.8, regardless of the link layout used.

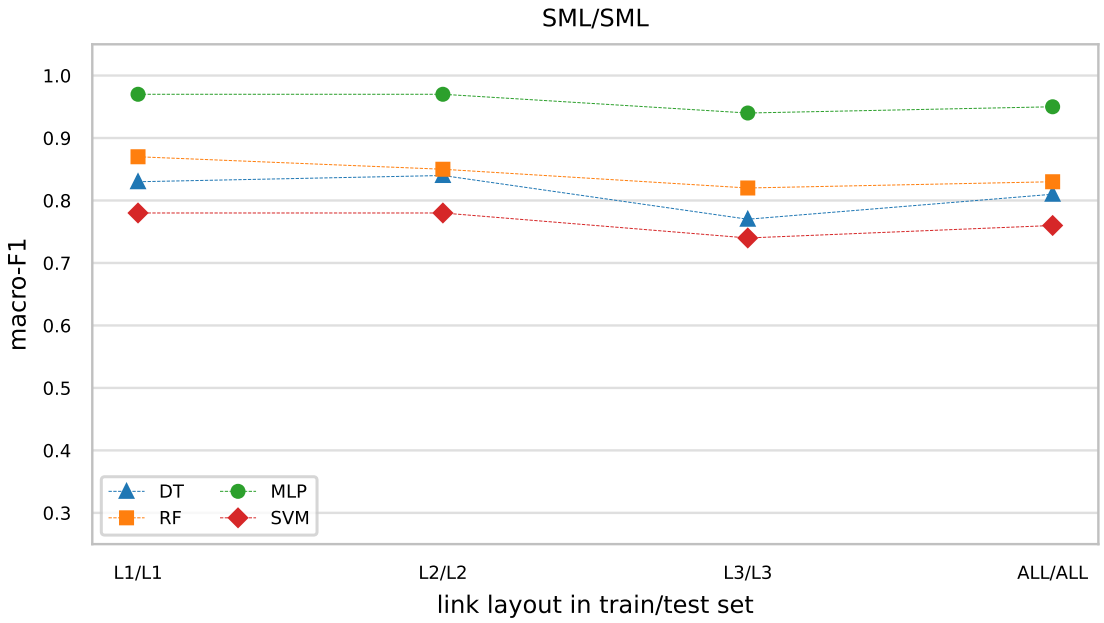

Figure 6.6: Generalization performance of models trained on three room sizes to new locations of the portable node. The evaluation scheme is *Eval–Init*. The evaluation settings are *3RS–train/3RS–test* (SML/SML) in terms of room size in train/test set, and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

The effect of joining multiple room sizes on the performance of the models is discussed based on the performances of the models trained with data obtained from three room sizes, presented in Figure 6.6. Specifically, the figure shows the scores of the models evaluated using *Eval–Init* evaluation scheme; trained and tested using three room sizes, considering *3RS–train/3RS–test* evaluation setting (SML/SML), and different link layouts, with evaluation settings *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3) and *mLyt–train/mLyt–test* (ALL/ALL). The conclusions previously inferred from the results when two room sizes are considered, i.e. evaluation setting *2RS–train/2RS–test*, are confirmed when three room sizes are utilized for training and testing the models.

Based on the evaluation results of the models in *Eval-Init*, it can be concluded that ML models can accurately predict the materials of all surfaces in rooms with sizes that were included in the training data when the fixed node positions are identical to those used during training. The ML models achieve a performance above 0.9 when the MLP algorithm is used, and slightly lower, between 0.8 and 0.9, when tree-based algorithms are employed (RF achieves slightly higher scores than DT, regardless of the room sizes and link layouts utilized). Additionally, it can be concluded that the materials can be accurately identified even when the portable node is moved to positions (over a grid that covers the room) that are not included in the training data. Depending on the use case, a model can be trained for each specific room size, or a single model can be trained for multiple room sizes. The variations in performances between models when different link layouts and room sizes are utilized are small.

6.2.2 Evaluation scheme *EvalSch-Lyt*

The capabilities of the models to generalize to new link layout, not included in the training process, are discussed for rooms with sizes that are considered in the training phase. The performances of the models from the following evaluation settings: *1RS–train/1RS–test*, *2RS–train/2RS–test*, and *3RS–train/3RS–test* in terms of the room sizes included in the train and test set, as well as *L1–train/diffLyt–test*, *L2–train/diffLyt–test*, *L3–train/diffLyt–test*, in terms of the link layout included in the train and test set, are presented; and the impact of the layout considered in the training process on the capability of the model to generalize to a new layout is discussed.

The results of the models trained with a single room size are presented in Figure 6.7. Particularly, the figure shows the macro-F1 scores of the models evaluated using *EvalSch-Lyt* evaluation scheme; trained and tested using single room size, considering *1RS–train/1RS–test* evaluation setting (S/S, M/M, and L/L), and different link layouts, with evaluation settings *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2). The results indicate that the majority of the models exhibit a high generalization capability, as evidenced by scores above 0.7 on new link layouts, regardless of the link layout considered for training/testing.

Evident variability in the scores is discernible for different link layouts in the train/test set. It indicates that the positions of the fixed node used in the training phase, relative to the room geometry and to the positions of the fixed node in testing affect the performance of the models. Models trained on data from L1 exhibit superior generalization performance when presented with data from L2 compared to data from L3, the difference increases as the size of the room increases. Models trained on data from L2 exhibit superior generalization performance when presented with data from L1 relative to data from L3, while models trained on data from L3 exhibit comparable generalization performance for data from both L1 and L2. In the case when L2/L1 or L3/L1, the scores as slightly higher compared to L1/L2 and L1/L3 in S rooms, respectively; demonstrating that when the grid over which the portable node is moved is small, the larger number of layout-specific positions in L2 and L3 compared to L1 affects the performance of the models. For larger rooms (M or L), this effect is less pronounced.

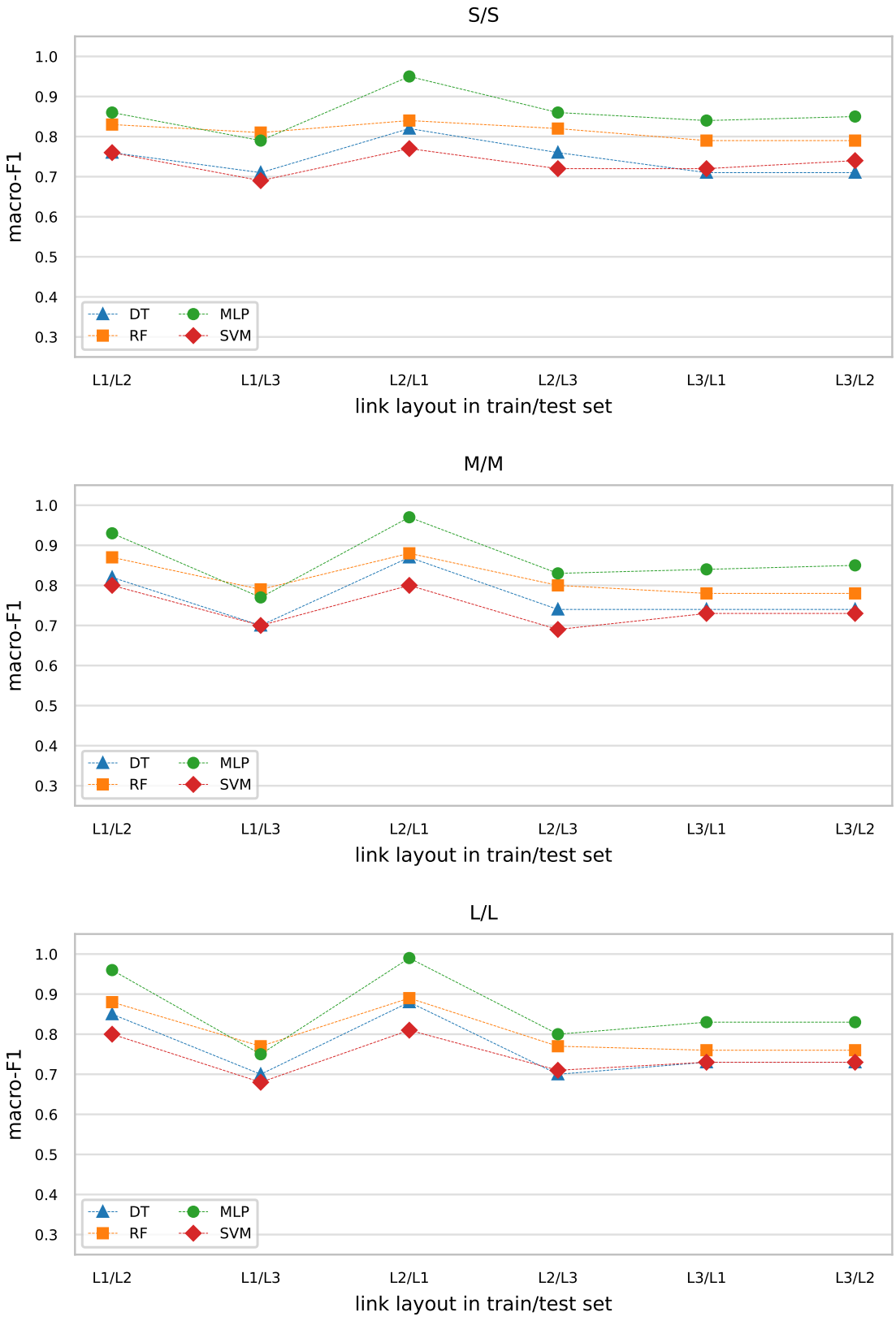

Figure 6.7: Generalization performance of models trained on a single room size to new link layouts. The evaluation scheme is *EvalSch–Lyt*. The evaluation settings are *1RS–train/1RS–test* (S/S, M/M, and L/L) in terms of room size in train/test set; and *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2), in terms of link layout in train/test set.

When L1 is used for training, in S rooms the MLP and RF models have similar performance (above 0.8), while DT and SVM models have comparable performance (above 0.7). In M and L rooms, when L2 is used for testing, a difference in the performance of MLP and RF models can be noticed. When L2 or L3 are considered for training, MLP models achieve the highest scores, above 0.9 in the L2/L1 case, and between 0.8 and 0.9 in other cases. RF models have lower performance, the scores are higher when tested with data from L1 (around 0.8 in S rooms and higher in M and L rooms). DT and SVM models achieve the lowest scores, above 0.7, in the cases L2/L3, L3/L1, and L3/L2, while in L2/L1, DT models have slightly higher scores comparable to RF models.

In the challenging cases, where the models trained on radio links with fixed nodes positioned in the center of the room or following a circular pattern centered in the room are expected to make predictions to CIR data from links with fixed nodes in the corners of the room (L1/L3 and L2/L3) slightly better performance is observed in the L2/L3 case compared to the L1/L3 case for all room sizes. In both cases, RF and MLP models have higher scores than DT and SVM models.

The generalization performance of models trained with L1 or L2 and tested on L3 slightly drops as the size of the room increases. This is expected because the distance from the center to the corners, where the fixed nodes for training and testing are placed, respectively, is larger in larger rooms. On the other hand, the results from the cases when models trained on data from L3 and tested on data from L1 or L2 (L3/L1 and L3/L2), there is no large variation in the scores for different room sizes because the distance between the corner nodes used in the train set and the walls is the same in rooms with different sizes. We can conclude that the distance between the layout-specific nodes and the walls affects the generalization capabilities of the models.

The results presented in Figure 6.8 show the macro-F1 scores of the models trained with two room sizes. Particularly, the figure shows the performance of the models when evaluated using *EvalSch-Lyt* evaluation scheme; trained and tested using two room sizes, considering *2RS–train/2RS–test* evaluation setting (SM/SM, ML/ML, and SL/SL), and different link layouts, with evaluation settings *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2). The combined impact of joining data from two room sizes and considering different link layouts in training and testing is analysed. Similar conclusions can be inferred about the generalization between different layouts (L1 and L2 to L3 and vice versa), and the impact of the positions of the fixed node relative to the room geometry, as when *1RS–train/1RS–test* evaluation setting is used. For most of the link evaluation settings in terms of the link layout used for training/testing, the MLP models achieve the highest scores. Exceptions are the L1/L3 and L3/L1 settings for SM/SM room size setting. This suggests that the algorithm is challenging to make predictions when L room is not considered.

To further discuss the impact of combining data from several room sizes, we analyse the results from models trained with data from three room sizes shown in Figure 6.9. The figure shows the performance of the models when evaluated using *EvalSch-Lyt* evaluation scheme; trained and tested using three room sizes, considering *3RS–train/3RS–test* evaluation setting (SML/SML), and different link layouts, with evaluation settings *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2). The conclusions made about the impact of the link layout used in the training and testing phase in *1RS–train/1RS–test*, and *2RS–train/2RS–test* evaluation settings are confirmed with these results. The MLP models have the highest performance in all evaluation settings in terms of the link layout except in the L1/L3 where RF model achieves higher scores.

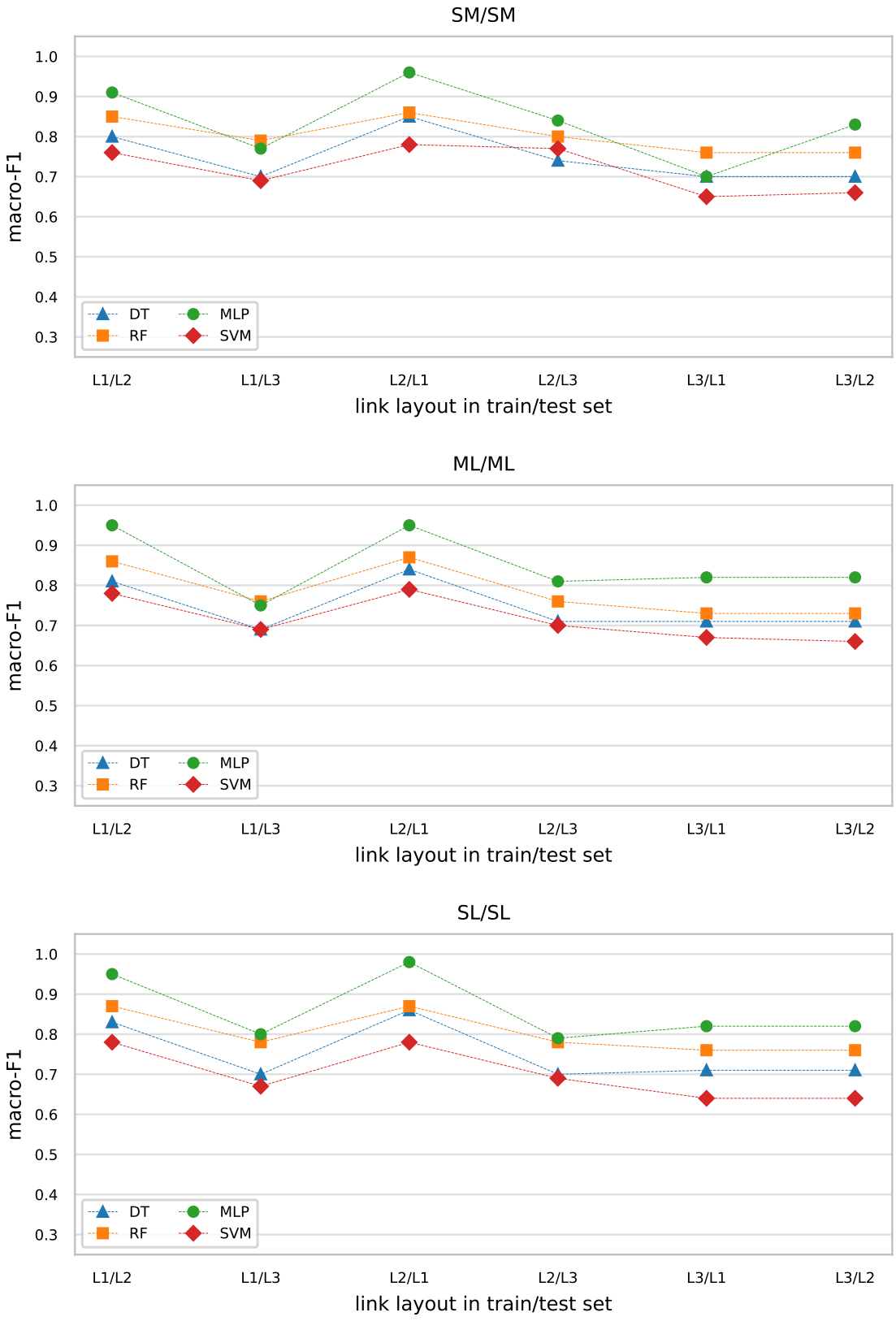

Figure 6.8: Generalization performance of models trained on two room sizes to new link layouts. The evaluation scheme is *EvalSch–Lyt*. The evaluation settings are *2RS–train/2RS–test* (SM/SM, ML/ML, and SL/SL) in terms of room size in train/test set; and *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2), in terms of link layout in train/test set.

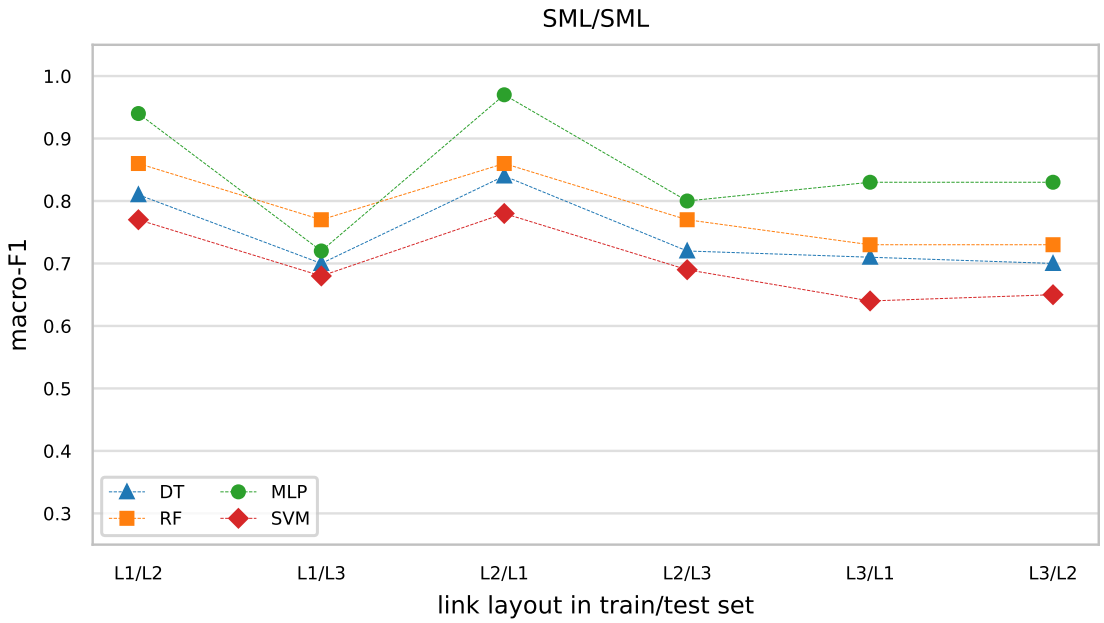

Figure 6.9: Generalization performance of models trained on three room size to new link layouts. The evaluation scheme is *EvalSch–Lyt*. The evaluation settings are *3RS–train/3RS–test* (SML/SML) in terms of room size in train/test set; and *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2), in terms of link layout in train/test set.

Based on results in *EvalSch-Lyt*, we conclude that the models trained with data from one layout can generalize to another layout in rooms with sizes that were included in the training process. This confirms that the material can be accurately identified when the CIR acquisition strategy, i.e., the positions of the fixed node, employed in the prediction phase is different than the strategy applied for collecting training data. However, it is important to notice that the position of the fixed nodes relative to the walls and the distance between the fixed node positions in the used strategies have an impact on the performance of the model.

6.2.3 Evaluation scheme *EvalSch-RS*

The capabilities of the models to generalize to new room sizes, not included in the training process, are discussed for the scenarios where the link layout used in the training and testing phase is the same. The performances of the models from the following evaluation settings: *1RS–train/2RS–test*, and *2RS–train/1RS–test* in terms of the room sizes included in the train and test set, as well as *1Lyt–train/1Lyt–test*, and *mLyt–train/mLyt–test* in terms of the link layouts included in the train and test set, are presented; and the impact of the room sizes on the generalization performance of the models is discussed.

The generalization performance of the models evaluated using the *1RS–train/2RS–test* evaluation setting is discussed based on the results presented in Figure 6.10. The figure shows the macro-F1 scores of the models evaluated using *EvalSch-RS* evaluation scheme; trained using one room size and tested using two room sizes, considering *1RS–train/2RS–test* (S/ML, M/SL, and L/SM), and different link layouts, with evaluation settings *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3), and *mLyt–train/mLyt–test* (ALL/ALL). Specifically, we discuss the generalization to larger rooms, smaller rooms and both larger and

smaller rooms. The scores are the lowest when models trained on S rooms are tested on data from larger rooms (S/ML), the scores slightly increase when models trained on M rooms are tested on smaller and larger rooms (M/SL), and the scores are the highest when models trained on L rooms are tested on smaller rooms (L/SM). This effect can be noticed for all models, with variations in scores' values for different learning methods. An exception are the MLP models when L1 or L2 are used, which achieve the highest scores in M/SL setting.

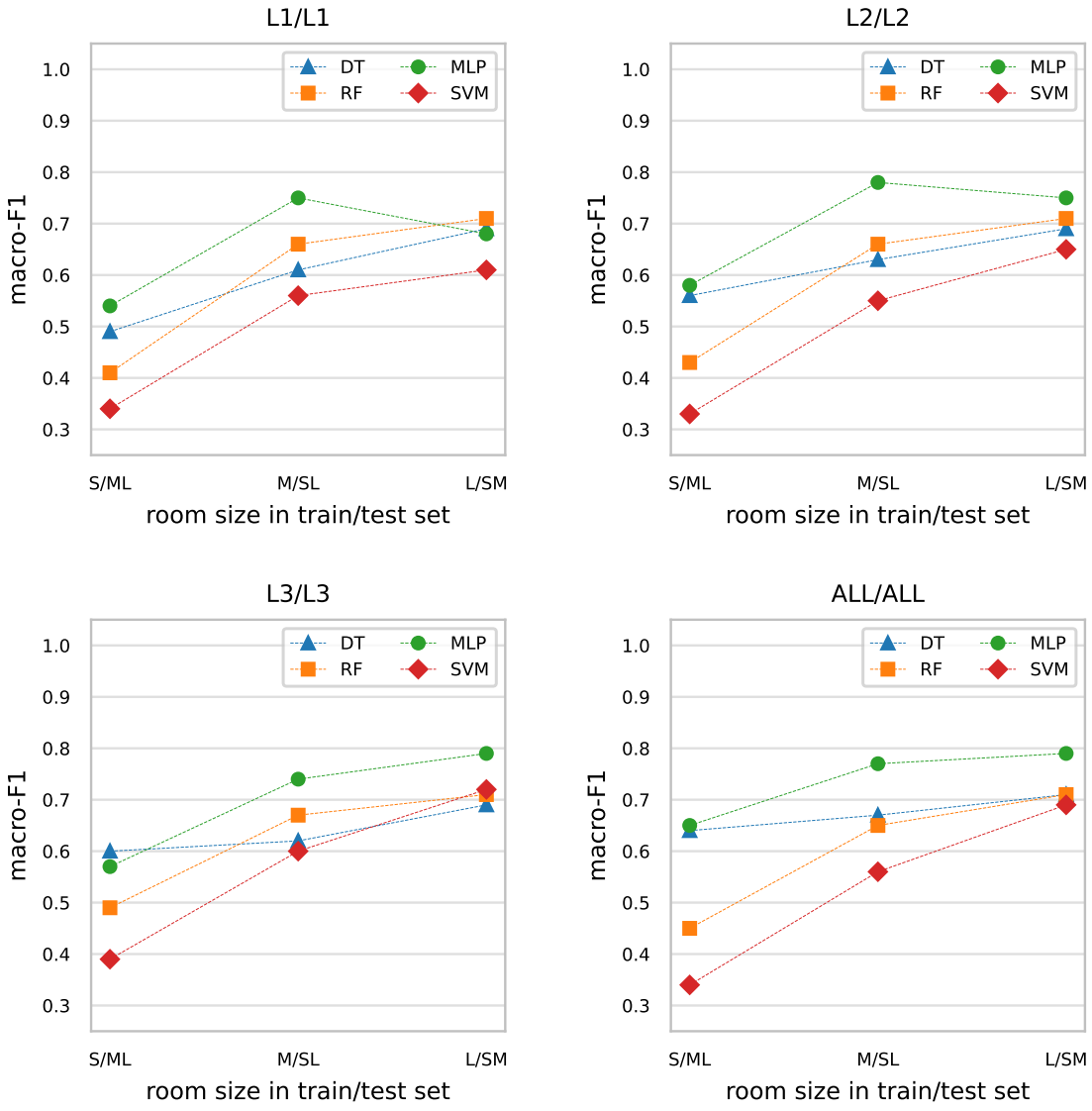

Figure 6.10: Generalization performance of models trained on a single room size to new room sizes. The evaluation scheme is *EvalSch–RS*. The evaluation settings are *1RS–train/2RS–test* (S/ML, M/SL, and L/SM) in terms of room size in train/test set; and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3), and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

In the most challenging evaluation setting S/ML, the performance of the algorithms is between 0.3 and 0.7, with the lowest scores when L1 is used and the highest scores when ALL layouts are used; suggesting that considering several layouts for obtaining the CIR

can improve the performance of the models. However, the prediction of materials in a room with a model trained for a smaller room is not adequate.

In evaluation setting M/SL and L/SM, the scores vary between 0.65 and 0.8 for different layouts, emphasizing the impact of the link layout on the model performance. Specifically, when M/SL is considered, MLP models have the highest scores, RF and DT have slightly lower, and SVM have the lowest scores in all link layout settings; while when L/SM is applied, the differences in the performance of the models are smaller, with MLP achieving highest scores in most of the link layout settings. Also, for the L/SM setting, it can be noticed that RF and DT have similar scores to SVM when L3 or ALL link layouts are used, but when L1 is used, the scores are similar to MLP.

The results confirm that models trained using data from a single room size are capable of generalizing to rooms with different sizes. However, it is important to note that the performance of such models in terms of generalization is affected by the room size included in the train and test set. Thus, for identification of the materials in a room using a model trained for a different room size, it is adequate to use a model trained for a larger room size. Furthermore, the results indicate that there is a slight variation in the performance of models trained and tested on different room layouts (L1/L1, L2/L2, L3/L3) due to differences in the number of radio links present.

The generalization performance of the models evaluated using the *2RS–train/1RS–test* evaluation setting is discussed based on the results given in Figure 6.11. The figure presents the macro-F1 scores of the models evaluated using *EvalSch-RS* evaluation scheme; trained using two room sizes and tested using one room size, considering *2RS–train/1RS–test* (SM/L, ML/S, and SL/M), and different link layouts, with evaluation settings *1Lyt–train/ 1Lyt–test* (L1/L1, L2/L2, and L3/L3), and *mLyt–train/mLyt–test* (ALL/ALL). The discussion focuses on the impact of joining data from different room sizes in the train set on the capability of the model to generalize new room sizes. Specifically, we analyse the capability of the models trained with two room sizes to generalize to large room size, small room size, and room size in between the room sizes included in the train set.

Variations in the performances of the models are evident when models are trained using data from different pairs of room sizes. The scores are the lowest when the models are tested on larger rooms compared to the rooms that are represented in the training process (SM/L), the scores are slightly higher when models are tested on smaller rooms (ML/S) and the scores are the highest when the models are tested on rooms with size in between the sizes included in the train set. Similar conclusions in terms of generalization to different room sizes can be made for the models trained/tested on the single layouts (L1/L1, L2/L2, L3/L3) as for the models trained on ALL layouts. From the learning algorithm perspective, a small variation in the scores can be noticed between models trained with different algorithms due to the distinct number of links in different layouts and room sizes. The MLP models achieve the highest performance in all evaluation settings in terms of room size (SM/L, ML/S, SL/M) and link layout (L1/L1, L2/L2, L3/L3). Across all link layout settings, the scores achieved are the highest for the SL/M room size setting, with values ranging between 0.8 and 0.9. In contrast, the scores obtained for the other room size settings are slightly lower.

The results obtained using the *EvalSch–RS* evaluation scheme demonstrate that the room sizes used during the training process have a significant impact on the ability of models to make accurate predictions in new rooms with different sizes. Specifically, models trained using data from a single room size exhibit better generalization performance in smaller rooms than in larger rooms; and models trained on data from two room sizes tend to generalize better to room sizes between the sizes of the rooms included in the training set.

The results show that for characterization of rooms with different sizes, models can be trained using data from one or multiple room sizes. A model trained on one room size generalizes better to smaller rooms than larger ones. Including data from various room sizes results in more robust models with better generalization performance, compared to models optimized to a single room size. When CIR data from all possible room sizes that the model may encounter is not available and its acquisition is expensive, or the time and computational resources for training the model are limited; the models should be trained with data from a representative subset of room sizes. The subset should include rooms that are smaller and larger than the ones where the model will be applied, to ensure the model captures a wide range of variability in room sizes.

For characterization of rooms with different sizes, models trained with data of one or several room sizes can be used. A model that is trained for one room size is able to generalize better to smaller rooms than to larger rooms. Considering data from various room sizes results in robust models in terms of generalization performance compared to models tailored to a single room size. In cases where the available resources in terms of time and computational constraints as well as limited data availability do not allow including all the room sizes in the train set, it is appropriate to include a selection of smaller and larger rooms compared to the rooms that will be characterized.

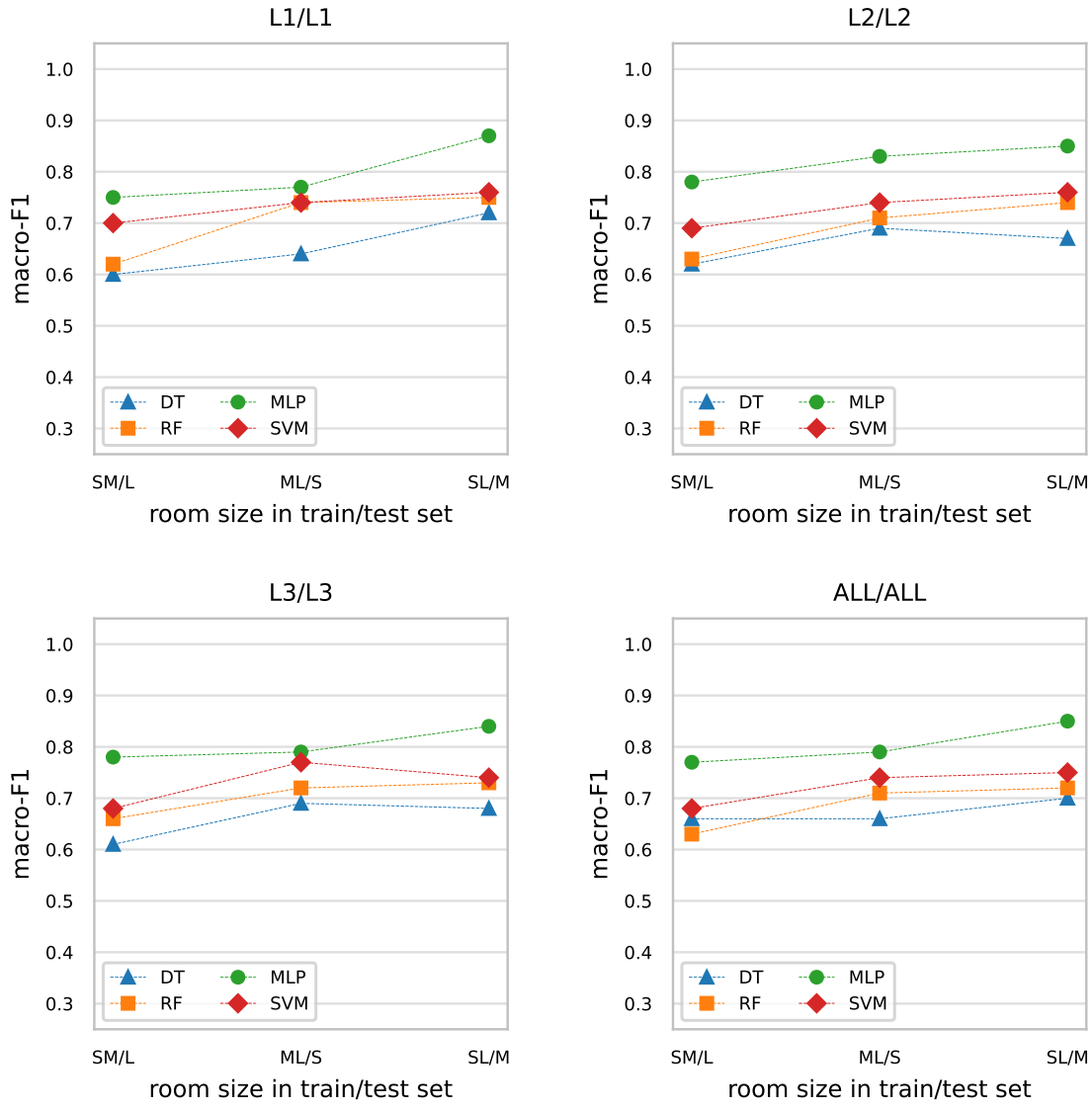

Figure 6.11: Generalization performance of models trained on two room sizes to a new room size. The evaluation scheme is *EvalSch–RS*. The evaluation settings are *2RS–train/1RS–test* (SM/L, ML/S, and SL/M) in terms of room size in train/test set; and *1Lyt–train/1Lyt–test* (L1/L1, L2/L2, and L3/L3), and *mLyt–train/mLyt–test* (ALL/ALL) in terms of link layout in train/test set.

6.2.4 Evaluation scheme *EvalSch–RS–Lyt*

The capabilities of the models to generalize to room sizes and link layouts that are not included in the train set are discussed. The performances of the models in the following evaluation settings: *1RS–train/2RS–test*, and *2RS–train/1RS–test* in terms of the room sizes included in the train and test set, as well as *L1–train/diffLyt–test*, *L2–train/diffLyt–test*, and *L3–train/diffLyt–test* in terms of the link layouts included in the train and test set, are presented; and the impact of the room size and the link layout on the generalization performance of the models is discussed.

The generalization performance of the models evaluated using the *1RS–train/2RS–test* evaluation setting is discussed according to the results given in Figure 6.12. The figure presents the macro-F1 scores of the models evaluated using *EvalSch-RS-Lyt* evaluation scheme; trained using one room size and tested using two room sizes, considering *1RS–train/2RS–test* (S/ML, M/SL, and L/SM), and different link layouts, with evaluation settings *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2).

Models tested on data from larger rooms (S/ML) have the lowest performance (below 0.6) regardless of the layout that is considered in the training/testing phase. When the models are tested on both smaller and larger rooms (M/SL), the scores are higher compared to S/ML. Exceptions are the DT models which have similar scores in both settings for L1/L3 and L2/L3 layouts in the train/test set. Specifically, in M/SL, the MLP models achieve scores between 0.68 and 0.8, RF and DT have comparable scores in the range between 0.6 and 0.7, and SVM models have the lowest performance between 0.5 and 0.6. Variation in the scores can be noticed for all models for different layouts used in the train/test set. The results confirm the conclusions drawn from *EvalSch–Lyt*.

Most of the models have larger scores in the L/SM setting, exception are the MLP models trained on L1 or L2, which have higher scores in the M/SL setting. The performance of the models learned with different algorithms varies, depending on the link layout included in the train set. Specifically, the three base methods exhibit superior performance compared to other methods when data from L1 is utilized for training. When training MLP, RF, and DT models using data from L2, they exhibit comparable performance. Finally, when data from L3 is used for training, MLP models demonstrate superior performance in comparison to other models, while the alternative models exhibit comparable performance. SVM models, in particular, exhibit inferior scores across all training settings.

The impact of joining data from different room sizes in the train set on the generalization performance is discussed from the results presented in Figure 6.13. The figure shows the macro-F1 scores of the models evaluated using *EvalSch-RS-Lyt* evaluation scheme trained using two room sizes and tested using one room size, considering *2RS–train/1RS–test* (SM/L, ML/S, and SL/M), and different link layouts, with evaluation settings *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2).

In the SM/L setting, it is observed that MLP models exhibit the highest scores across all link layout evaluation settings (above 0.7). The RF and DT models demonstrate comparable performance levels across the different link layouts used in the train/test set, with scores ranging between 0.55 and 0.65. However, when L3 data is utilized for training, RF models demonstrate slightly better performance. The performance of SVM models is significantly impacted by the link layout in train/test. Specifically, in scenarios involving L1/L2, L1/L3, and L2/L1, SVM models outperform tree-based models. However, in other settings, their performance is slightly lower than that of tree-based methods.

When the ML/S evaluation setting is applied, MLP models exhibit the highest scores when data from L1 and L2 are utilized for training, while their performance is comparable

to that of RF models when data from L3 is employed. RF and SVM models demonstrate almost identical scores when data from L1 and L2 are utilized for training, while SVM models exhibit slightly better performance when data from L3 is used for training. On the other hand, DT models demonstrate the lowest performance levels across all link layout evaluation settings. However, when data from L2 is utilized for training, the performance of DT models is similar to that of RF and SVM models.

In the SL/M, both MLP and RF models achieve high scores above 0.7 across all settings. Specifically, MLP models exhibit superior performance in L1/L2 and L2/L1 (above 0.8), while for L1/L3 and L2/L3, the performance scores are comparable between MLP and RF models. Compared to MLP and RF models, DT models show slightly lower performance in all link layout evaluation settings. However, the difference in scores is small in L1/L2, L3/L1, or L3/L2 evaluation settings when compared to RF models. SVM models' performance is sensitive to the link layout setting and shows variability across different settings. In L1/L3, L2/L3, L3/L1, and L3/L2 evaluation settings, SVM models' performance is similar to DT models, while in L1/L2 and L2/L1, the performance is closer to that of RF models.

The results from evaluation scheme *EvalSch–RS–Lyt* confirm that the room size and the link layout used for training the model affect the generalization capabilities of the model. It can be concluded that the models trained with one room size can generalize best to smaller room sizes and joining data from different room sizes improves the generalization capabilities of the models. To enhance the robustness and generalization capability of the model, it is suggested that various room sizes are included in the train set. In particular, it is important that the train set includes smaller and larger rooms in comparison to the size of the rooms where the model will be applied.

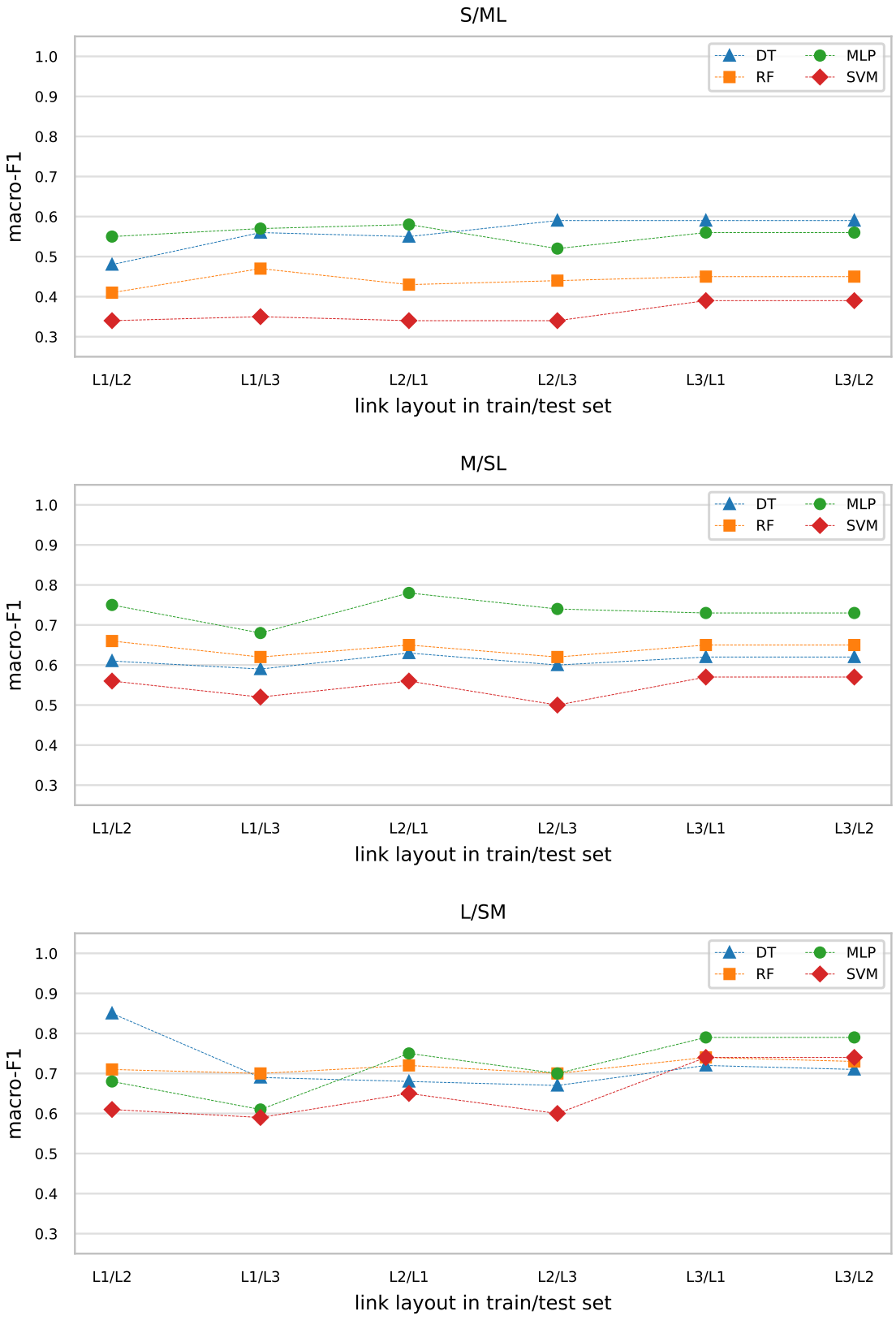

Figure 6.12: Generalization performance of models trained on a single room size to new link layouts and room sizes. The evaluation scheme is *EvalSch–RS–Lyt*. The evaluation settings are *1RS–train/2RS–test* (S/ML, M/SL, and L/SM) in terms of room size in train/test set; and *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2) in terms of link layout in train/test set.

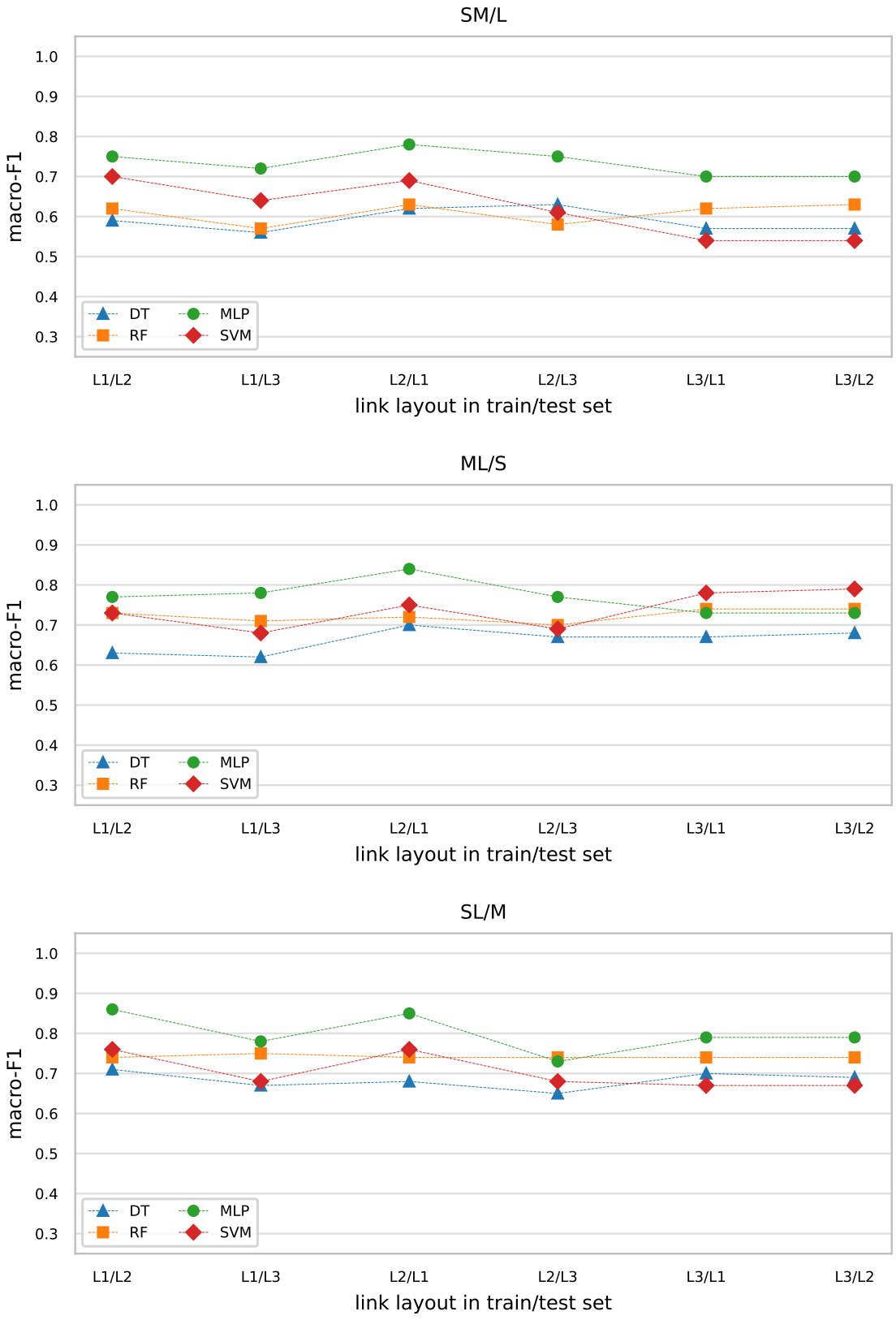

Figure 6.13: Generalization performance of models trained on two room sizes to new link layouts and room sizes. The evaluation scheme is *EvalSch–RS–Lyt*. The evaluation settings are *2RS–train/1RS–test* (SM/L, ML/S, and SL/M) in terms of room size in train/test set; and *L1–train/diffLyt–test* (L1/L2, and L1/L3), *L2–train/diffLyt–test* (L2/L1, and L2/L3), and *L3–train/diffLyt–test* (L3/L1, and L3/L2), in terms of link layout in train/test set.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

Accurate characterization of the properties of indoor RE is essential for future wireless communications, which are envisioned to be environmentally aware and dynamically adapt to the propagation environment. Current environment characterization methodologies need to be revised to meet the emerging spatial information requirements because they have limited ability to describe the environment and depend on specialized equipment, resulting in inflexibility, the need for manual intervention, maintenance, and high cost. Therefore, novel methodologies for seamless indoor characterization that provide more detailed and accurate information about the environment while overcoming the limitations of available approaches are needed.

The large number of connected devices with sensing capabilities in next-generation wireless networks offers unprecedented opportunities for intelligent indoor characterization. The ability of the emerging radio technologies to estimate CSI for each transmit-receive antenna pair resulting in an availability of a vast amount of indoor CSI data, which contains information about the propagation environment, opens the way for radio-based indoor propagation environment characterization. ML approaches are foreseen as an appropriate tool for modeling the relationship between the environmental information contained in CSI and the properties of the environment. However, their use must be substantiated by domain knowledge of indoor radio channels. By using state-of-the-art radio technologies and ML approaches, the strengths of both can be combined to capture the RE signature and extract knowledge from the data for seamless characterization of the indoor environment.

This dissertation proposes and studies the identification of indoor RE environment properties based on the environmental information embedded in CSI using ML approaches. The proposed approach is novel and, to our knowledge, has not been studied, published, or patented before. It is based on the assumptions that (i) the received radio signal is distorted due to interaction with the surrounding objects and therefore contains a signature of RE, and (ii) the UWB signal conveys information about the environment it interacts with and is, therefore, suitable for environment characterization.

A methodology for intelligent indoor RE characterization RE based on CIR of spatially distributed radio links is formalized and evaluated. The proposed methodology is evaluated for characterizing surfaces' materials in the microwave frequency band using UWB radio technology and ray-based techniques. A large amount of CIR data, representative and diverse in terms of radio node positions and EM characteristics of the environment, is collected and annotated with environmental properties.

The procedure for constructing a RE identification model based on RE signature is streamlined in a framework. Domain knowledge is crucial in all procedure steps to ensure

propagation features with rich RE signatures are used, and the RE identification model addresses the indoor RE characterization problem. The framework specifies the (i) RE signature, (ii) RE acquisition, (iii) feature selection from CSI, (iv) CSI processing and storing, (v) ML task, and (vi) ML workflow.

The proposed methodology is evaluated for identifying the material of (i) a single surface in basic environments and (ii) all surfaces in plain environments based on CIR. Common building materials used in residential and office buildings, including brick, concrete, glass, plaster, and wood, are studied. Material identification is formalized as a classification task, and representative learning methods of tree-based learning, ensemble learning, kernel-based learning, and neural networks are considered. The identification of the material of a single surface is formalized as a multi-class classification task, and the identification of the material of all surfaces is formalized as a multi-label classification task.

The experiments are defined to prove the stated hypotheses. Specifically, initial experimentation is conducted for (i) proving the feasibility of the concept for basic indoor environments, (ii) investigating different approaches for constructing a model in terms of the CIR acquisition strategy and the size of the rooms where CIR is obtained, (iii) assessing the capability of ML models to extract knowledge from CIR. More detailed experimentation is conducted for (i) proving the feasibility of the concept for environments with varying complexity in terms of the materials used, and (ii) studying the capabilities of the models to make accurate predictions on CIRs from acquisition strategies and room sizes not considered in the training process.

The results confirm that the proposed methodology can be applied for the identification of the material of a single surface as well as for the identification of the material of all surfaces in plain indoor environments based on CIR. The material can be predicted using models learned with all the proposed methods, however, the performances vary depending on the room size and the radio links included in the training set. Across the various evaluation schemes, the MLP models show superior performance. In most cases, three-based models exhibit slightly lower performance. Ensembles of trees tend to achieve higher scores than single-tree models, reflecting the advantages of combining multiple trees. SVM models show the lowest performance and sensitivity to CIR acquisition strategy CIR and room size, highlighting the importance of parameter selection to optimize performance. The selection of an optimal learning method for a given environment characterization scenario depends on the amount of available CIR data, time constraints, and available computational resources.

Material identification can be approached in a (i) specific manner—with models optimized for single room size and CIR acquisition strategy, or (ii) general manner—with models trained using CIR data acquired with several strategies from rooms of different sizes. The choice of approach depends on the use case where the material identification model is used. For basic indoor characterization scenarios, where the model is applied in rooms of a certain size and the CIR acquisition strategy used in the training process is known, the specific approach is adequate; for advanced scenarios, where rooms with different sizes are characterized without knowledge of the CIR acquisition strategy used in the training process, the general approach is more appropriate. However, the results show that models trained for a single room size can generalize to smaller rooms. The results suggest that when a general approach is considered for training a model, CIR data from different room sizes that are smaller and larger compared to the rooms where the model is applied to identify the materials have to be included in the train set. The training data has to be acquired using fixed nodes in several locations and portable nodes moved over a uniform grid covering the room.

The dissertation makes an original contribution to the research in seamless and in-

telligent indoor environment characterization by proposing, formalizing, and evaluating a novel methodology for identifying RE properties based on CSI using ML approaches. The proposed methodology provides an environment description that can be used to enhance the digital twin of buildings and contributes to the development of pioneering methods for indoor environment-aware communications in the next-generation communication era. The experimental evaluation presented in the dissertation is only an initial example of environment characterization using the proposed methodology. The extensibility of the methodology to other CSI properties, CSI estimation methods, indoor environments, and indoor characterization tasks is one of its most important values. The importance of the methodology is highlighted in the next-generation intelligence-enabled wireless communications, where a detailed description of the indoor environment is a prerequisite for achieving state-of-the-art performance.

7.2 Future Work

The current study has demonstrated the feasibility of indoor RE characterization based on RE signatures using ML tools, ray tracing, and UWB technology in the microwave frequency band. The presented work can be extended to comprehensively explore the potential of the proposed approach. To build upon the findings presented in the dissertation, future work will focus on (1) exploiting the potentials of state-of-the-art approaches from wireless communications, such as (i) frequencies in the terahertz frequency band offering a much larger bandwidth, leading to better time resolution of CIR, and (ii) multiple antenna systems offering high accuracy and resolution in the angle of arrival estimation, for accurate estimation of the RE signature in complex indoor environments; and (2) leveraging the cutting-edge ML approaches, to model the relation between the RE signature and the indoor RE properties. Specifically, we identify the following directions for future work:

- *Extending the proposed approach for UWB measurements in versatile indoor environments.* The bandwidth of the measurement equipment is a crucial factor in the temporal resolution of the CIR, which affects the estimation of the size and shape of the room. Existing UWB communication systems have a bandwidth in the range of GHz, resulting in a time resolution of the CIR in the range of tens of centimetres. The CIR will be obtained using UWB transceivers with radio chips capable of communicating according to IEEE 802.15.4-2011 UWB standard with centre frequencies in microwave frequency range and different bandwidths. The indoor environments will vary in geometry, materials, and interacting objects. In that context, we have already built a portable radio-equipment and have started extensive CIR measurement campaigns in office environments.

- *Extending the RE signature with the angle of arrival information.* The utilization of multiple antenna systems for state-of-the-art angle of arrival estimation has the potential to enable the comprehensive characterization of the RE. A massive MIMO approach at both the transmitter and receiver, or alternatively, distributed multiple transmitter-receiver links throughout the room have to be considered. In this case, the RE signature is the angle of arrival-dependent power delay profile.

- *Extending the proposed approach for characterization of surfaces with multiple materials.* The characterization of complex rooms where the surfaces have multiple material is foreseen possible with the use of MIMO antennas.

- *Investigation of the impact of label dependencies in the output space on the performance of the models.* In scenarios where the propagating wave interacts with multiple

surfaces, exploiting label dependencies can add additional information to the learning method, leading to improved performance.

- *Extending the proposed approach for characterization of the materials of specific surfaces in the room.* The utilization of RE signatures that include the angle of arrival information opens the possibility for precise identification of the material used for constructing a particular surface.

References

[1] R. Otero, S. Laguela, I. Garrido, and P. Arias, "Mobile indoor mapping technologies: A review," *Automation in Construction*, vol. 120, p. 103 399, 2020.

[2] H. Hashemi, "The indoor radio propagation channel," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 943–968, 1993.

[3] S. R. Saunders and A. Aragón-Zavala, *Antennas and propagation for wireless communication systems*. Chichester, UK: John Wiley & Sons, 2007.

[4] Y. Liu, X. Liu, X. Mu, *et al.*, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1546–1577, 2021.

[5] L. Barclay, *Propagation of radiowaves*, 2nd ed., ser. Electromagnetic Waves. London, UK: The Institution of Engineering and Technology, 2003.

[6] C. A. Balanis, *Advanced Engineering Electromagnetics*, 2nd ed. New Jersey, USA: John Wiley & Sons, 2012.

[7] T. Rappaport, S. Seidel, and K. Takamizawa, "Statistical channel impulse response models for factory and open plan building radio communicate system design," *IEEE Transactions on Communications*, vol. 39, no. 5, pp. 794–807, 1991.

[8] J. Andersen, T. Rappaport, and S. Yoshida, "Propagation measurements and models for wireless communications channels," *IEEE Communications Magazine*, vol. 33, no. 1, pp. 42–49, 1995.

[9] D. Molkdar, "Review on radio propagation into and within buildings," in *IEE Proceedings H-Microwaves, Antennas and Propagation*, vol. 138, IET, 1991, pp. 61–73.

[10] P. Stenumgaard, J. Chilo, J. Ferrer-Coll, and P. Angskog, "Challenges and conditions for wireless machine-to-machine communications in industrial environments," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 187–192, 2013.

[11] I. Rodriguez, H. C. Nguyen, N. T. Jorgensen, T. B. Sorensen, and P. Mogensen, "Radio propagation into modern buildings: Attenuation measurements in the range from 800 MHz to 18 GHz," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Vancouver, BC, Canada: IEEE, 2014, pp. 1–5.

[12] M. E. Diago-Mosquera, A. Aragón-Zavala, and G. Castañón, "Bringing it indoors: A review of narrowband radio propagation modeling for enclosed spaces," *IEEE Access*, vol. 8, pp. 103 875–103 899, 2020.

[13] R. sector of International Telecommunication Union (ITU-R), "Effects of building materials and structures on radiowave propagation above about 100 MHz," International Telecommunication Union, ITU-R Recommendation P.2040-2, 2021.

[14] A. Hrovat, K. Guam, T. Kocevska, and T. Javornik, "3D indoor environment characterization based on radio scanning: Initial idea and methodolgy," in *2019 23rd International Conference on Applied Electromagnetics and Communications (ICECOM2019)*, Dubrovnik, Croatia, Sep. 2019, pp. 1–6.

[15] T. Sarkar, Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma, "A survey of various propagation models for mobile communication," *IEEE Antennas and Propagation Magazine*, vol. 45, no. 3, pp. 51–82, 2003.

[16] Z. Yun and M. F. Iskander, "Ray tracing for radio propagation modeling: Principles and applications," *IEEE Access*, vol. 3, pp. 1089–1100, 2015.

[17] F. Ikegami, T. Takeuchi, and S. Yoshida, "Theoretical prediction of mean field strength for urban mobile radio," *IEEE Transactions on Antennas and Propagation*, vol. 39, no. 3, pp. 299–302, 1991.

[18] J. McKown and R. Hamilton, "Ray tracing as a design tool for radio networks," *IEEE Network*, vol. 5, no. 6, pp. 27–30, 1991.

[19] M. Katz, *Introduction to geometrical optics*. Singapore: World Scientific Publishing, 2002.

[20] A. Taflove and C. Hagness Susan, *Computational Electrodynamics: The Finite-Difference Time-Domain Method*, 3rd. Norwood, MA: Artech House, 2005.

[21] W. C. Gibson, *The Method of Moments in Electromagnetics*, 3rd. New York, USA: Chapman and Hall/CRC, Taylor & Francis Group, 2008.

[22] J.-M. Jin, *The Finite Element Method in Electromagnetics*, 3rd. New York, USA: John Wiley & Sons, 2014.

[23] J. B. Keller, "Geometrical theory of diffraction," *Journal of the Optical Society of America*, vol. 52, no. 2, pp. 116–130, 1962.

[24] P. H. Pathak, G. Carluccio, and M. Albani, "The uniform geometrical theory of diffraction and some of its applications," *IEEE Antennas and Propagation Magazine*, vol. 55, no. 4, pp. 41–69, 2013.

[25] M. Born, E. Wolf, A. B. Bhatia, *et al.*, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 7th ed. Cambridge, UK: Cambridge University Press, 1999.

[26] W. Honcharenko, H. Bertoni, J. Dailing, J. Qian, and H. Yee, "Mechanisms governing UHF propagation on single floors in modern office buildings," *IEEE Transactions on Vehicular Technology*, vol. 41, no. 4, pp. 496–504, 1992.

[27] S. Seidel and T. Rappaport, "Site-specific propagation prediction for wireless in-building personal communication system design," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 4, pp. 879–891, 1994.

[28] N. Noori, A. Shishegar, and E. Jedari, "A new double counting cancellation technique for three-dimensional ray launching method," in *2006 IEEE Antennas and Propagation Society International Symposium*, Albuquerque, NM, USA, 2006, pp. 2185–2188.

[29] Z. Chen, H. Bertoni, and A. Delis, "Progressive and approximate techniques in ray-tracing-based radio wave propagation prediction models," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 1, pp. 240–251, 2004.

[30] Y. Tao, H. Lin, and H. Bao, "GPU-based shooting and bouncing ray method for fast RCS prediction," *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 2, pp. 494–502, 2010.

[31] Remcom. "Wireless In Site propagation software v. 3.3.3." (2019), [Online]. Available: `https://www.remcom.com/wireless-insite-em-propagation-software` (visited on 03/15/2023).

[32] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the Dartmouth summer research project on artificial intelligence," *AI magazine*, vol. 27, no. 4, pp. 12–12, 1955.

[33] P. H. Winston, *Artificial intelligence*. Addison-Wesley Publishing Company, 1992.

[34] M. A. Bramer, *Principles of Data Mining*, 3rd, ser. Undergraduate Topics in Computer Science. London: Springer, 2007.

[35] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[36] T. Mitchel, *Machine Learning*. New York, USA: McGraw-Hill, 1997.

[37] R. S. Sutton and A. G. Barto, *Reinforcement learning*, ser. Adaptive computation and machine learning series. Cambridge, MA, USA: MIT Press, 2018.

[38] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, pp. 37–66, 1991.

[39] G. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, vol. 38, p. 100 285, 2020.

[40] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *Advances in neural information processing systems*, vol. 11, 1998.

[41] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Cham: Springer, 2016.

[42] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Tree ensembles for predicting structured outputs," *Pattern Recognition*, vol. 46, no. 3, pp. 817–833, 2013.

[43] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] L. Buitinck, G. Louppe, M. Blondel, *et al.*, "API design for machine learning software: Experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[45] A. Géron, "Hands-on machine learning with scikit-learn and tensorflow: Concepts," *Tools, and Techniques to build intelligent systems*, 2017.

[46] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*, 1st ed. Bocca Raton, FL, USA: Taylor & Francis Group, 1984.

[47] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[48] ——, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann Publishers, Inc., 1993.

[49] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, vol. 41, pp. 77–93, 2004.

[50] M. U. Fayyad and B. K. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning*, vol. 8, pp. 87–102, 1992.

[51] H. Friedman Jerome, "A recursive partitioning decision rule for nonparametric classification," *IEEE Transactions on Computers*, vol. 26, no. 4, pp. 404–408, 1977.

[52] P. Taylor and B. W. Silverman, "Block diagrams and splitting criteria for classification trees," *Statistics and Computing*, vol. 3, no. 4, pp. 147–161, 1993.

[53] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[54] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[55] *Intelligent data analysis*. Springer, 2003, vol. 2.

[56] V. Vapnik, I. Guyon, and T. Hastie, "Support vector machines," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[57] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: A review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 803–855, 2019.

[58] D.-X. Zhou and K. Jetter, "Approximation with polynomial kernels and svm classifiers," *Advances in Computational Mathematics*, vol. 25, no. 1-3, pp. 323–344, 2006.

[59] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, 2003.

[60] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," *Neural Computing*, vol. 3, no. 1-32, p. 16, 2003.

[61] J. Platt, "Sequential minimal optimization: A fast algorithm for training Support Vector Machines," Microsoft, Tech. Rep. MSR-TR-98-14, Apr. 1998. [Online]. Available: `https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/`.

[62] S. I. Gallant *et al.*, "Perceptron-based learning algorithms," *IEEE Transactions on neural networks*, vol. 1, no. 2, pp. 179–191, 1990.

[63] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *International Journal of Engineering Applied Sciences and Technology*, vol. 4, no. 2455-2143, pp. 310–316, 2017.

[64] M. L. Minsky and S. A. Papert, *Perceptrons: An introduction to computational geometry*. USA: MIT press, 1969.

[65] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and intelligent laboratory systems*, vol. 39, no. 1, pp. 43–62, 1997.

[66] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[67] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, "Comprehensive comparative study of multi-label classification methods," *Expert Systems with Applications*, vol. 203, p. 117 215, 2022.

[68] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *Principles of Data Mining and Knowledge Discovery: 5th European Conference, (PKDD2001)*, Springer, Freiburg, Germany, 2001, pp. 42–53.

[69] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *International Conference on Machine Learning*, 1999.

[70] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern recognition*, vol. 45, no. 9, pp. 3084–3104, 2012.

[71] "MLTSVM: A novel twin support vector machine to multi-label learning," *Pattern Recognition*, vol. 52, pp. 61–74, 2016.

[72] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *2005 IEEE International Conference on Granular Computing*, vol. 2, Beijing, China: IEEE, 2005, pp. 718–721.

[73] ——, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[74] M.-L. Zhang, "ML-RBF: Rbf neural networks for multi-label learning," *Neural Processing Letters*, vol. 29, pp. 61–74, 2009.

[75] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[76] F. Liu, Y. Cui, C. Masouros, *et al.*, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[77] Y. Chen, J. Zhang, W. Feng, and M.-S. Alouini, "Radio sensing using 5G signals: Concepts, state of the art, and challenges," *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 1037–1052, 2022.

[78] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 5–36, 2022.

[79] H. Wymeersch, D. Shrestha, C. M. de Lima, *et al.*, "Integration of communication and sensing in 6G: A joint industrial and academic perspective," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMR2021)*, Helsinki, Finland: IEEE, 2021, pp. 1–7.

[80] D. Ma, N. Shlezinger, T. Huang, Y. Liu, and Y. C. Eldar, "Joint radar-communication strategies for autonomous vehicles: Combining two key automotive technologies," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 85–97, 2020.

[81] D. K. Pin Tan, J. He, Y. Li, *et al.*, "Integrated sensing and communication in 6G: Motivations, use cases, requirements, challenges and future directions," in *2021 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S)*, Dresden, Germany: IEEE, 2021, pp. 1–6.

[82] J. A. Zhang, M. L. Rahman, K. Wu, *et al.*, "Enabling joint communication and radar sensing in mobile networks—A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 306–345, 2022.

[83] M. L. Rahman, J. A. Zhang, X. Huang, Y. J. Guo, and R. W. Heath, "Framework for a perceptive mobile network using joint communication and radar sensing," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 3, pp. 1926–1941, 2020.

[84] Q. Huang, H. Chen, and Q. Zhang, "Joint design of sensing and communication systems for smart homes," *IEEE Network*, vol. 34, no. 6, pp. 191–197, 2020.

[85] Y. Chen, P. Zhu, G. He, X. Yan, H. Baligh, and J. Wu, "From connected people, connected things, to connected intelligence," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*, Levi, Finland: IEEE, 2020, pp. 1–7.

[86] H. Rahman, *Fundamental Principles of Radar*. Florida, USA: CRC Press, 2019.

[87] C. Xu, Z. Ren, J. Wei, and C. Lee, "Reconfigurable terahertz metamaterials: From fundamental principles to advanced 6G applications," *iScience*, vol. 25, no. 2, 2022.

[88] C. A. Balanis and P. I. Ioannides, "Introduction to smart antennas," *Synthesis Lectures on Antennas*, vol. 2, no. 1, pp. 1–175, 2007.

[89] N. Shlezinger, G. C. Alexandropoulos, M. F. Imani, Y. C. Eldar, and D. R. Smith, "Dynamic metasurface antennas for 6g extreme massive mimo communications," *IEEE Wireless Communications*, vol. 28, no. 2, pp. 106–113, 2021.

[90] S. Basharat, S. A. Hassan, H. Pervaiz, A. Mahmood, Z. Ding, and M. Gidlund, "Reconfigurable intelligent surfaces: Potentials, applications, and challenges for 6G wireless networks," *IEEE Wireless Communications*, vol. 28, no. 6, pp. 184–191, 2021.

[91] S. Tripathi, N. V. Sabu, A. K. Gupta, and H. S. Dhillon, "Millimeter-wave and terahertz spectrum for 6G wireless," in *6G Mobile Wireless Networks*, Y. Wu, S. Singh, T. Taleb, *et al.*, Eds., Cham: Springer International Publishing, 2021, pp. 83–121.

[92] C. Sturm and W. Wiesbeck, "Waveform design and signal processing aspects for fusion of wireless communications and radar sensing," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1236–1259, 2011.

[93] J. A. Zhang, F. Liu, C. Masouros, *et al.*, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 6, pp. 1295–1315, 2021.

[94] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3834–3862, 2020.

[95] Z. Feng, Z. Fang, Z. Wei, X. Chen, Z. Quan, and D. Ji, "Joint radar and communication: A survey," *China Communications*, vol. 17, no. 1, pp. 1–27, 2020.

[96] X. Wang, A. Hassanien, and M. G. Amin, "Dual-function MIMO radar communications system design via sparse array optimization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 3, pp. 1213–1226, 2019.

[97] F. Liu, L. Zhou, C. Masouros, A. Lit, W. Luo, and A. Petropulu, "Dual-functional cellular and radar transmission: Beyond coexistence," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC2018)*, Kalamata, Greece: IEEE, 2018, pp. 1–5.

[98] C. De Lima, D. Belot, R. Berkvens, *et al.*, "Convergent communication, sensing and localization in 6G systems: An overview of technologies, opportunities and challenges," *IEEE Access*, vol. 9, pp. 26 902–26 925, 2021.

[99] A. R. Chiriyath, B. Paul, and D. W. Bliss, "Radar-communications convergence: Coexistence, cooperation, and co-design," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 1, pp. 1–12, 2017.

[100] T. Wild, V. Braun, and H. Viswanathan, "Joint design of communication and sensing for beyond 5G and 6G systems," *IEEE Access*, vol. 9, pp. 30 845–30 857, 2021.

[101] L. Zheng, M. Lops, Y. C. Eldar, and X. Wang, "Radar and communication coexistence: An overview: A review of recent methods," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 85–99, 2019.

[102] A. Abdelhadi and T. C. Clancy, "Network MIMO with partial cooperation between radar and cellular systems," in *2016 International Conference on Computing, Networking and Communications (ICNC2016)*, Kauai, HI, USA, 2016, pp. 1–5.

[103] M. Bică and V. Koivunen, "Radar waveform optimization for target parameter estimation in cooperative radar-communications systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 5, pp. 2314–2326, 2019.

[104] A. Khawar, A. Abdelhadi, and T. C. Clancy, "Coexistence analysis between radar and cellular system in LoS channel," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 972–975, 2016.

[105] K. Singh, S. Biswas, O. Taghizadeh, and T. Ratnarajah, "Beamforming design for coexistence of full-duplex multi-cell MU-MIMO cellular network and MIMO radar," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019)*, Brighton, UK, 2019, pp. 7775–7779.

[106] S. Kumar, K. V. Mishra, S. Gautam, B. S. M. R., and B. Ottersten, "Interference mitigation methods for coexistence of radar and communication," in *2021 15th European Conference on Antennas and Propagation (EuCAP2021)*, Dusseldorf, Germany: IEEE, 2021, pp. 1–4.

[107] P. Stinco, M. Greco, F. Gini, and B. Himed, "Channel parameters estimation for cognitive radar systems," in *2014 4th International Workshop on Cognitive Information Processing (CIP2014)*, Copenhagen, Denmark: IEEE, 2014, pp. 1–6.

[108] S. D. Blunt, M. R. Cook, and J. Stiles, "Embedding information into radar emissions via waveform implementation," in *2010 International Waveform Diversity and Design Conference*, Niagara Falls, ON, Canada: IEEE, 2010, pp. 000 195–000 199.

[109] C. Sahin, J. Jakabosky, P. M. McCormick, J. G. Metcalf, and S. D. Blunt, "A novel approach for embedding communication symbols into physical radar waveforms," in *2017 IEEE Radar Conference (RadarConf)*, Seattle,WA, USA: IEEE, 2017, pp. 1498–1503.

[110] P. M. McCormick, S. D. Blunt, and J. G. Metcalf, "Simultaneous radar and communications emissions from a common aperture, part i: Theory," in *2017 IEEE Radar Conference (RadarConf)*, 2017, pp. 1685–1690.

[111] A. Hassanien, M. G. Amin, Y. D. Zhang, and F. Ahmad, "Dual-function radar-communications: Information embedding using sidelobe control and waveform diversity," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 2168–2181, 2016.

[112] R. Thomä, T. Dallmann, S. Jovanoska, P. Knott, and A. Schmeink, "Joint communication and radar sensing: An overview," in *2021 15th European Conference on Antennas and Propagation (EuCAP2021)*, Dusseldorf, Germany: IEEE, 2021, pp. 1–5.

[113] A. Aubry, A. De Maio, Y. Huang, M. Piezzo, and A. Farina, "A new radar waveform design algorithm with improved feasibility for spectral coexistence," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 2, pp. 1029–1038, 2015.

[114] C. W. Rossler, E. Ertin, and R. L. Moses, "A software defined radar system for joint communication and sensing," in *2011 IEEE RadarCon (RADAR)*, Kansas City, MO, USA, 2011, pp. 1050–1055.

[115] O. Ma, A. R. Chiriyath, A. Herschfelt, and D. W. Bliss, "Cooperative radar and communications coexistence using reinforcement learning," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA: IEEE, 2018, pp. 947–951.

[116] N. Nartasilpa, D. Tuninetti, N. Devroye, and D. Erricolo, "Let's share CommRad: Effect of radar interference on an uncoded data communication system," in *2016 IEEE Radar Conference (RadarConf2016)*, Philadelphia, PA, USA: IEEE, 2016, pp. 1–5.

[117] M. Ghorbanzadeh, E. Visotsky, P. Moorut, W. Yang, and C. Clancy, "Radar in-band and out-of-band interference into LTE macro and small cell uplinks in the 3.5 GHz band," in *2015 IEEE Wireless Communications and Networking Conference (WCNC2015)*, New Orleans, LA, USA, 2015, pp. 1829–1834.

[118] J. H. Reed, A. W. Clegg, A. V. Padaki, *et al.*, "On the co-existence of TD-LTE and radar over 3.5 GHz band: An experimental study," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 368–371, 2016.

[119] B. D. Cordill, S. A. Seguin, and L. Cohen, "Electromagnetic interference to radar receivers due to in-band OFDM communications systems," in *2013 IEEE International Symposium on Electromagnetic Compatibility*, Denver, CO, USA, 2013, pp. 72–75.

[120] W. Liu, J. Fang, H. Tan, B. Huang, and W. Wang, "Coexistence studies for TD-LTE with radar system in the band 2300–2400 MHz," in *2010 International Conference on Communications, Circuits and Systems (ICCCAS2010)*, Chengdu, China: IEEE, 2010, pp. 49–53.

[121] A. R. Chiriyath, B. Paul, G. M. Jacyna, and D. W. Bliss, "Inner bounds on performance of radar and communications co-existence," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 464–474, 2016.

[122] J. A. Mahal, A. Khawar, A. Abdelhadi, and T. C. Clancy, "Spectral coexistence of MIMO radar and MIMO cellular system," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 2, pp. 655–668, 2017.

[123] F. Liu, A. Garcia-Rodriguez, C. Masouros, and G. Geraci, "Interfering channel estimation in radar-cellular coexistence: How much information do we need?" *IEEE Transactions on Wireless Communications*, vol. 18, no. 9, pp. 4238–4253, 2019.

[124] B. K. Chalise, M. G. Amin, and B. Himed, "Performance tradeoff in a unified passive radar and communications system," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1275–1279, 2017.

[125] S. Y. Nusenu, W.-Q. Wang, and A. Basit, "Time-modulated fd-mimo array for integrated radar and communication systems," *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 6, pp. 1015–1019, 2018.

[126] T. W. Tedesso and R. Romero, "Code shift keying based joint radar and communications for EMCON applications," *Digital Signal Processing*, vol. 80, pp. 48–56, 2018.

[127] A. Ahmed, Y. D. Zhang, and Y. Gu, "Dual-function radar-communications using QAM-based sidelobe modulation," *Digital Signal Processing*, vol. 82, pp. 166–174, 2018.

[128] Z. Geng, R. Xu, H. Deng, and B. Himed, "Fusion of radar sensing and wireless communications by embedding communication signals into the radar transmit waveform," *IET Radar, Sonar & Navigation*, vol. 12, no. 6, pp. 632–640, 2018.

[129] A. Hassanien, B. Himed, and M. G. Amin, "Transmit/receive beamforming design for joint radar and communication systems," in *2018 IEEE Radar Conference (RadarConf18)*, Oklahoma City, OK, USA: IEEE, 2018, pp. 1481–1486.

[130] P. Kumari, J. Choi, N. González-Prelcic, and R. W. Heath, "Ieee 802.11ad-based radar: An approach to joint vehicular communication-radar system," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3012–3027, 2018.

[131] K. V. Mishra, M. Bhavani Shankar, V. Koivunen, B. Ottersten, and S. A. Vorobyov, "Toward millimeter-wave joint radar communications: A signal processing perspective," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 100–114, 2019.

[132] M. Alloulah and H. Huang, "Future millimeter-wave indoor systems: A blueprint for joint communication and sensing," *Computer*, vol. 52, no. 7, pp. 16–24, 2019.

[133] Z. Lin and Z. Wang, "Interleaved OFDM signals for MIMO radar," *IEEE Sensors Journal*, vol. 15, no. 11, pp. 6294–6305, 2015.

[134] K. Wu, J. A. Zhang, X. Huang, Y. J. Guo, and R. W. Heath, "Waveform design and accurate channel estimation for frequency-hopping MIMO radar-based communications," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1244–1258, 2021.

[135] D. Ma, N. Shlezinger, T. Huang, Y. Liu, and Y. C. Eldar, "FRaC: FMCW-based joint radar-communications system via index modulation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 6, pp. 1348–1364, 2021.

[136] S. D. Liyanaarachchi, C. B. Barneto, T. Riihonen, and M. Valkama, "Experimenting joint vehicular communications and sensing with optimized 5G NR waveform," in *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*, Helsinki, Finland: IEEE, 2021, pp. 1–5.

[137] H. Jiang, C. Cai, X. Ma, Y. Yang, and J. Liu, "Smart home based on WiFi sensing: A survey," *IEEE Access*, vol. 6, pp. 13 317–13 325, 2018.

[138] C. B. Barneto, S. D. Liyanaarachchi, M. Heino, T. Riihonen, and M. Valkama, "Full duplex radio/radar technology: The enabler for advanced joint communication and sensing," *IEEE Wireless Communications*, vol. 28, no. 1, pp. 82–88, 2021.

[139] M. Weib, "Synchronisation of bistatic radar systems," in *2004 IEEE International Geoscience and Remote Sensing Symposium (IGARSS2004)*, vol. 3, 2004, pp. 1750–1753.

[140] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, P. Popovski, and M. Debbah, "Seven defining features of terahertz (THz) wireless systems: A fellowship of communication and sensing," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 967–993, 2022.

[141] F. Liu and C. Masouros, "Hybrid beamforming with sub-arrayed MIMO radar: Enabling joint sensing and communication at mmWave band," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019)*, 2019, pp. 7770–7774.

[142] T. Mao, J. Chen, Q. Wang, C. Han, Z. Wang, and G. K. Karagiannidis, "Waveform design for joint sensing and communications in millimeter-wave and low terahertz bands," *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 7023–7039, 2022.

[143] A. Liu, Z. Huang, M. Li, *et al.*, "A survey on fundamental limits of integrated sensing and communication," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 994–1034, 2022.

[144] K. Khoshelham and S. Zlatanova, "Sensors for indoor mapping and navigation," *Sensors*, vol. 16, no. 5, p. 655, 2016.

[145] A. Masiero, F. Fissore, A. Guarnieri, F. Pirotti, D. Visintini, and A. Vettore, "Performance evaluation of two indoor mapping systems: Low-cost UWB-aided photogrammetry and backpack laser scanning," *Applied Sciences*, vol. 8, no. 3, p. 416, 2018.

[146] Y. Cui, Q. Li, B. Yang, W. Xiao, C. Chen, and Z. Dong, "Automatic 3-D reconstruction of indoor environment with mobile laser scanning point clouds," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3117–3130, 2019.

[147] A. Morar, A. Moldoveanu, I. Mocanu, *et al.*, "A comprehensive survey of indoor localization methods based on computer vision," *Sensors*, vol. 20, no. 9, p. 2641, 2020.

[148] C. Kolhatkar and K. Wagle, "Review of SLAM algorithms for indoor mobile robot with lidar and RGB-D camera technology," in *Innovations in Electrical and Electronic Engineering*, M. N. Favorskaya, S. Mekhilef, R. K. Pandey, and N. Singh, Eds., Singapore": Springer Singapore, 2021, pp. 397–409.

[149] F. Guidi, A. Guerra, and D. Dardari, "Personal mobile radars with millimeter-wave massive arrays for indoor mapping," *IEEE Transactions on Mobile Computing*, vol. 15, no. 6, pp. 1471–1484, 2016.

[150] A. Guerra, F. Guidi, A. Clemente, R. D'Errico, L. Dussopt, and D. Dardari, "Application of transmitarray antennas for indoor mapping at millimeter-waves," in *2015 European Conference on Networks and Communications (EuCNC2015)*, Paris, France: IEEE, 2015, pp. 77–81.

[151] S. Dogru and L. Marques, "Grid based indoor mapping using radar," in *2019 Third IEEE International Conference on Robotic Computing (IRC2019)*, Naples, Italy: IEEE, 2019, pp. 451–452.

[152] J. Chen and K. C. Clarke, "Indoor cartography," *Cartography and Geographic Information Science*, vol. 47, no. 2, pp. 95–109, 2020.

[153] A. S. Nossum, "Developing a framework for describing and comparing indoor maps," *The Cartographic Journal*, vol. 50, no. 3, pp. 218–224, 2013.

[154] S. Hong, J. Jung, S. Kim, H. Cho, J. Lee, and J. Heo, "Semi-automated approach to indoor mapping for 3D as-built building information modeling," *Computers, Environment and Urban Systems*, vol. 51, pp. 34–46, 2015.

[155] V. V. Lehtola, H. Kaartinen, A. Nüchter, *et al.*, "Comparison of the selected state-of-the-art 3D indoor scanning and point cloud generation methods," *Remote sensing*, vol. 9, no. 8, p. 796, 2017.

[156] C. Thomson and J. Boehm, "Automatic geometry generation from point clouds for BIM," *Remote Sensing*, vol. 7, no. 9, pp. 11 753–11 775, 2015.

[157] P. Hübner, M. Weinmann, S. Wursthorn, and S. Hinz, "Automatic voxel-based 3D indoor reconstruction and room partitioning from triangle meshes," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 254–278, 2021.

[158] M. Li and L. Nan, "Feature-preserving 3D mesh simplification for urban buildings," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 173, pp. 135–150, 2021.

[159] K. Wang, Y.-A. Lin, B. Weissmann, M. Savva, A. X. Chang, and D. Ritchie, "Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–15, 2019.

[160] J. Chen and K. C. Clarke, "Indoor cartography," *Cartography and Geographic Information Science*, vol. 47, no. 2, pp. 95–109, 2020.

[161] K.-J. Li, S. Zlatanova, J. Torres-Sospedra, A. Perez-Navarro, C. Laoudias, and A. Moreira, "Survey on indoor map standards and formats," in *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN2019)*, 2019, pp. 1–8.

[162] J. Chen and K. C. Clarke, "Modeling standards and file formats for indoor mapping.," in *3rd International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM2017)*, Portugal, 2017, pp. 268–275.

[163] A. Guerra, F. Guidi, D. Dardari, A. Clemente, and R. D'Errico, "A millimeter-wave indoor backscattering channel model for environment mapping," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 9, pp. 4935–4940, 2017.

[164] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[165] C. Cadena, L. Carlone, H. Carrillo, *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[166] B. Wang, Q. Xu, C. Chen, F. Zhang, and K. R. Liu, "The promise of radio analytics: A future paradigm of wireless positioning, tracking, and sensing," *IEEE Signal Processing Magazine*, vol. 35, no. 3, pp. 59–80, 2018.

[167] S. Tan, Y. Ren, J. Yang, and Y. Chen, "Commodity WiFi sensing in ten years: Status, challenges, and opportunities," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 832–17 843, 2022.

[168] S. Zlatanova, G. Sithole, M. Nakagawa, and Q. Zhu, "Problems in indoor mapping and modelling," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-4/W4, pp. 63–68, 2013.

[169] K. Kensek, *Building Information Modeling*. New York, USA: John Wiley & Sons, 2014.

[170] F. Yang, G. Zhou, F. Su, *et al.*, "Automatic indoor reconstruction from point clouds in multi-room environments with curved walls," *Sensors*, vol. 19, no. 17, p. 3798, 2019.

[171] Z. Kang, J. Yang, Z. Yang, and S. Cheng, "A review of techniques for 3D reconstruction of indoor environments," *ISPRS International Journal of Geo-Information*, vol. 9, no. 5, p. 330, 2020.

[172] W. Shi, Y. Gong, C. Ding, Z. M. Tao, and N. Zheng, "Transductive semi-supervised deep learning using min-max features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 299–315.

[173] U. M. Khan, R. H. Venkatnarayan, and M. Shahzad, "RFMap: Generating indoor maps using RF signals," in *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN2020)*, Sydney, NSW, Australia: IEEE, 2020, pp. 133–144.

[174] P. Fritsche, B. Zeise, P. Hemme, and B. Wagner, "Fusion of radar, LiDAR and thermal information for hazard detection in low visibility environments," in *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, Shanghai, China: IEEE, 2017, pp. 96–101.

[175] T. Raj, F. H. Hashim, A. B. Huddin, M. F. Ibrahim, and A. Hussain, "A survey on LiDAR scanning mechanisms," *Electronics*, vol. 9, no. 5, 2020.

[176] S. De Geyter, J. Vermandere, H. De Winter, M. Bassier, and M. Vergauwen, "Point cloud validation: On the impact of laser scanning technologies on the semantic segmentation for BIM modeling and evaluation," *Remote Sensing*, vol. 14, no. 3, 2022.

[177] Y. Li, W. Li, S. Tang, W. Darwish, Y. Hu, and W. Chen, "Automatic indoor as-built building information models generation by using low-cost RGB-D sensors," *Sensors*, vol. 20, no. 1, 2020.

[178] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft Kinect sensor: A review," *IEEE Transactions on Cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[179] M. Filipenko and I. Afanasyev, "Comparison of various SLAM systems for mobile robot in an indoor environment," in *2018 International Conference on Intelligent Systems (IS)*, Funchal, Portugal: IEEE, 2018, pp. 400–407.

[180] Y. Chen, J. Tang, C. Jiang, *et al.*, "The accuracy comparison of three simultaneous localization and mapping (SLAM)-based indoor mapping technologies," *Sensors*, vol. 18, no. 10, p. 3228, 2018.

[181] C. Baquero Barneto, E. Rastorgueva-Foi, M. F. Keskin, *et al.*, "Millimeter-wave mobile sensing and environment mapping: Models, algorithms and validation," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 4, pp. 3900–3916, 2022.

[182] H. Sarieddeen, N. Saeed, T. Y. Al-Naffouri, and M.-S. Alouini, "Next generation terahertz communications: A rendezvous of sensing, imaging, and localization," *IEEE Communications Magazine*, vol. 58, no. 5, pp. 69–75, 2020.

[183] A. Yassin, Y. Nasser, A. Y. Al-Dubai, and M. Awad, "MOSAIC: Simultaneous localization and environment mapping using mmWave without a-priori knowledge," *IEEE Access*, vol. 6, pp. 68 932–68 947, 2018.

[184] Z. Chen, C. Han, Y. Wu, *et al.*, "Terahertz wireless communications for 2030 and beyond: A cutting-edge frontier," *IEEE Communications Magazine*, vol. 59, no. 11, pp. 66–72, 2021.

[185] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Computing Surveys*, vol. 52, no. 3, 2020.

[186] S. Zhang and D. Zhu, "Towards artificial intelligence enabled 6G: State of the art, challenges, and opportunities," *Computer Networks*, vol. 183, p. 107 556, 2020.

[187] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao, and K. Wu, "Artificial-intelligence-enabled intelligent 6G networks," *IEEE Network*, vol. 34, no. 6, pp. 272–280, 2020.

[188] J. W. Ma, T. Czerniawski, and F. Leite, "An application of metadata-based image retrieval system for facility management," *Advanced Engineering Informatics*, vol. 50, p. 101 417, 2021.

[189] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2020.

[190] T. Xin, B. Guo, Z. Wang, *et al.*, "FreeSense: A robust approach for indoor human detection using Wi-Fi signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–23, 2018.

[191] A. Jamali, A. Abdul Rahman, P. Boguslawski, P. Kumar, and C. M. Gold, "An automated 3D modeling of topological indoor navigation network," *GeoJournal*, vol. 82, pp. 157–170, 2017.

[192] F. Poux, R. Neuville, G.-A. Nys, and R. Billen, "3D point cloud semantic modelling: Integrated framework for indoor spaces and furniture," *Remote Sensing*, vol. 10, no. 9, p. 1412, 2018.

[193] L. Deren, Y. Wenbo, and S. Zhenfeng, "Smart city based on digital twins," *Computational Urban Science*, vol. 1, no. 1, pp. 1–11, 2021.

[194] M. I. AlHajri, N. T. Ali, and R. M. Shubair, "A machine learning approach for the classification of indoor environments using RF signatures," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP2018)*, Anaheim, CA, USA: IEEE, 2018, pp. 1060–1062.

[195] Z. Chen, M. I. AlHajri, M. Wu, N. T. Ali, and R. M. Shubair, "A novel real-time deep learning approach for indoor localization based on rf environment identification," *IEEE Sensors Letters*, vol. 4, no. 6, pp. 1–4, 2020.

[196] M. I. AlHajri, N. T. Ali, and R. M. Shubair, "Classification of indoor environments for IoT applications: A machine learning approach," *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 12, pp. 2164–2168, 2018.

[197] B. C. Csáji, A. Browet, V. Traag, *et al.*, "Exploring the mobility of mobile phone users," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[198] D. Ferreira, I. Cuiñas, R. F. Caldeirinha, and T. R. Fernandes, "A review on the electromagnetic characterisation of building materials at micro- and millimetre wave frequencies," in *The 8th European Conference on Antennas and Propagation (EuCAP2014)*, The Hague, Netherlands, 2014, pp. 145–149.

[199] U. Kaatze, "Measuring the dielectric properties of materials. Ninety-year development from low-frequency techniques to broadband spectroscopy and high-frequency imaging," *Measurement Science and Technology*, vol. 24, no. 1, p. 012 005, 2012.

[200] U. T. Virk, S. L. H. Nguyen, K. Haneda, and J.-F. Wagen, "On-site permittivity estimation at 60 GHz through reflecting surface identification in the point cloud," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 7, pp. 3599–3609, 2018.

[201] I. Vilovic, N. Burum, and R. Nadj, "Estimation of dielectric constant of composite materials in buildings using reflected fields and PSO algorithm," in *Proceedings of the Fourth European Conference on Antennas and Propagation*, Barcelona, Spain: IEEE, 2010, pp. 1–5.

[202] O. Landron, M. Feuerstein, and T. Rappaport, "In situ microwave reflection coefficient measurements for smooth and rough exterior wall surfaces," in *IEEE 43rd Vehicular Technology Conference*, Secaucus, NJ, USA: IEEE, 1993, pp. 77–80.

[203] J. Ahmadi-Shokouh, S. Noghanian, and H. Keshavarz, "Reflection coefficient measurement for north american house flooring at 57–64 GHz," *IEEE Antennas and Wireless Propagation Letters*, vol. 10, pp. 1321–1324, 2011.

[204] J. Lu, D. Steinbach, P. Cabrol, P. Pietraski, and R. V. Pragada, "Propagation characterization of an office building in the 60 GHz band," in *The 8th European Conference on Antennas and Propagation (EuCAP2014)*, The Hague, Netherlands: IEEE, 2014, pp. 809–813.

[205] T. Kocevska, A. Hrovat, and T. Javornik, *Indoor UWB CIR data set for material prediction*, version v.1, Zenodo, Apr. 2023. [Online]. Available: `https://doi.org/10.5281/zenodo.7840706`.

[206] A. F. Molisch, *Wireless Communications*, 2nd ed. Chichester, West Sussex, U.K: Wiley: IEEE, 2011.

[207] IEEE, "Standard for local and metropolitan area networks–Part 15.4: Low-rate wireless personal area networks (LR-WPANs)," IEEE, Standard IEEE 802.15.4-2011, 2011.

[208] R. Turner, *Python Programming: 3 Books in 1 - Ultimate Beginner's, Intermediate & Advanced Guide to Learn Python Step by Step*. Nelly B.L. International Consulting, Ltd., 2020.

[209] E. Gündüzalp, G. Yildirim, and Y. Tatar, "Radio propagation prediction using SBR in a tunnel with narrow cross-section and obstacles," in *2019 International Conference on Applied Automation and Industrial Diagnostics (ICAAID2019)*, vol. 1, Elazig, Turkey, 2019, pp. 1–6.

[210] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement," *SIGKDD Explor. Newsl.*, vol. 12, no. 1, pp. 49–57, Nov. 2010.

Bibliography

Publications Related to the Thesis

Journal Articles

T. Kocevska, T. Javornik, A. Švigelj, and A. Hrovat, "Framework for the machine learning based wireless sensing of the electromagnetic properties of indoor materials," *Electronics*, vol. 10, no. 22, p. 2843, 2021.

T. Kocevska, T. Javornik, A. Švigelj, A. Rashkovska, and A. Hrovat, "Identification of indoor radio environment properties from channel impulse response with machine learning models," *Electronics*, 2023, (under review).

Conference Paper

T. Kocevska, T. Javornik, A. Švigelj, K. Guan, A. Rashkovska, and A. Hrovat, "Impact of room size on machine learning-based material prediction using channel impulse response," in *30th International Conference on Systems, Signals and Image Processing (IWSSIP2023)*, (accepted), Ohrid, North Macedonia, Jun. 2023.

——, "Comparison of machine learning models for predicting indoor materials from channel impulse response," in *2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM2022)*, Split, Croatia, Sep. 2022, pp. 1–6.

T. Javornik, A. Hrovat, and T. Kocevska, "Ali lahko impulzni odziv radijskega kanala uporabimo za oceno lastnosti notranjih prostorov = Estimation of the indoor environment by measured channel impulse response," in *24 seminar radijske komunikacije = 24th Sem. Radio Comm. (SRK2020)*, Ljubljana, Slovenia, Feb. 2020, pp. 5–7.

A. Hrovat, K. Guam, T. Kocevska, and T. Javornik, "3D indoor environment characterization based on radio scanning: Initial idea and methodolgy," in *2019 23rd International Conference on Applied Electromagnetics and Communications (ICECOM2019)*, Dubrovnik, Croatia, Sep. 2019, pp. 1–6.

Technical Documents

T. Kocevska, T. Javornik, A. Švigelj, K. Guan, A. Rashkovska, and A. Hrovat, "Generalization of machine learning models for CIR-based materials identification in indoor radio environment," in *COST: 4th Technical Meeting*, Dubrovnik, Croatia, 2023.

T. Kocevska, T. Javornik, A. Švigelj, and A. Hrovat, "ML algorithms comparison and hyperparameters tuning for wall material prediction based on CIR data," in *COST: 2nd Technical Meeting*, Lyon, France, 2022.

——, "Wireless sensing of the electromagnetic properties of wall materials: A ML approach based on indoor radio-environment signature," in *COST: 2nd Post-IRACON Meeting*, Bologna, Italy, 2021.

Data Sets

T. Kocevska, A. Hrovat, and T. Javornik, *Indoor UWB CIR data set for material prediction*, version v.1, Zenodo, Apr. 2023. [Online]. Available: `https://doi.org/10.5281/zenodo.7840706`.

Other Publications

Journal Articles

T. Kocevska, P. Latkoski, M. Porjazoski, and B. Popovski, "Analysis of latency, blocking probability and network utilization for a specific routing and spectrum assignment algorithm, in elastic optical networks," *Acta Polytechnica Hungarica*, 2020.

Conference Papers

T. Kocevska, A. Hrovat, and T. Javornik, "Study of cooperative MIMO approach in hilly environment," in *2020 8th International Conference on Wireless Networks and Mobile Communications (WINCOM2020)*, 2020, pp. 1–5.

Biography

Teodora Kocevska was born on 5 May 1995 in Tetovo, Macedonia. She finished primary education in Tearce and secondary education at the Gymnasium Kiril Pejcinovikj in Tetovo. She received her bachelor's and master's degrees from the Faculty of Electrical Engineering and Information Technologies, University Ss. Cyril and Methodius in Skopje, Macedonia. In 2013, she enrolled in the four-year bachelor's degree program "Telecommunications and Information Engineering" and defended her bachelor's thesis in September 2017. In October 2017, she enrolled in the one-year master's program "Communication and Information Technologies" and defended her master's thesis in October 2018. In October 2019, she enrolled in the three-year doctoral program "Information and Communication Technologies" at the Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, under the supervision of Asst. Prof. Dr. Andrej Hrovat.

During her bachelor studies, she held a state scholarship awarded by the Ministry of Education and Science of Macedonia. Her doctoral studies were funded by an Ad-futura scholarship awarded by the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia.

After completing her school education, she was named Best Student of the Generation. After completing her bachelor's studies, she received the Dean's recognition for her overall success in university studies. She received the award "Best graduate engineer in the field of telecommunications in 2017/2018" from the Society for Electronics, Telecommunications, Automatics, and Informatics in Macedonia.

During her bachelor studies, she collaborated, as an intern, with the "National Agency of Electronic Communications" in Macedonia. From 2017 to 2018, she collaborated as a student researcher in the Optical Communications Laboratory at the Institute of Telecommunications, Faculty of Electrical Engineering in Skopje, Macedonia. During this time, she conducted research in the field of optical communications. In the spring of 2019, she was a visiting researcher at the Laboratory of Communication Technology, Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia. During her stay, she focused on indoor environment characterization with radio-based methods.

Her main research interest is in telecommunications, specifically wireless communications. Her current research is focused on indoor radio channels, integrated communications and sensing, environmentally aware communications, and the application of machine learning in telecommunications. Her latest research is in characterizing the indoor radio environment using wireless technology and machine learning techniques.