The Use of Machine Learning Techniques for Optimal Multicasting in 5G NR Systems

Nadezhda Chukhno, Olga Chukhno, Dmitri Moltchanov, Anna Gaydamaka, Andrey Samuylov, Antonella Molinaro, Yevgeni Koucheryavy, Antonio Iera, and Giuseppe Araniti

***Abstract*—Multicasting is a key feature of cellular systems, which provides an efficient way to simultaneously disseminate a large amount of traffic to multiple subscribers. However, the efficient use of multicast services in fifth-generation (5G) New Radio (NR) is complicated by several factors, including inherent base station (BS) antenna directivity as well as the exploitation of antenna arrays capable of creating multiple beams concurrently. In this work, we first demonstrate that the problem of efficient multicasting in 5G NR systems can be formalized as a special case of multi-period variable cost and size bin packing problem (BPP). However, the problem is known to be NP-hard, and the solution time is practically unacceptable for large multicast group sizes. To this aim, we further develop and test several machine learning alternatives to address this issue. The numerical analysis shows that there is a trade-off between accuracy and computational complexity for multicast grouping when using decision tree-based algorithms. A higher number of splits offers better performance at the cost of an increased computational time. We also show that the nature of the cell coverage brings three possible solutions to the multicast grouping problem: (i) small-range radii are characterized by a single multicast subgroup with wide beamwidth, (ii) middle-range deployments have to be solved by employing the proposed algorithms, and (iii) BS at long-range radii sweeps narrow unicast beams to serve multicast users.**

***Index Terms*—5G, machine learning, millimeter Wave, multicast, multi-beam antennas, New Radio, optimization.**

I. Introduction

By issuing Release 15 and Release 16 [1], [2], 3GPP has completed most of the efforts towards New Radio (NR) technology standardization. Operating in both microwave (μWave) and millimeter wave (mmWave) bands, the standardized systems promise to deliver extraordinary rates to the air interface [3], [4]. Hence, the current focus of both 3GPP and the research community is shifting toward delivering value-added services on top of this new radio access technology, with multicast capabilities being on the list of tasks for coming Releases [5], [6].

N. Chukhno, O. Chukhno, A. Molinaro, G. Araniti are with Mediterranea University of Reggio Calabria, Reggio Calabria, Italy and CNIT, Italy. Email: {nadezda.chukhno, olga.chukhno, araniti, antonella.molinaro}@unirc.it

O. Chukhno, D. Moltchanov, A. Gaydamaka, A. Samuylov, Y. Koucheryavy are with Tampere University, Tampere, Finland. Email: {dmitri.moltchanov, anna.gaydamaka, andrey.samuylov, evgeny.kucheryavy}@tuni.fi

N. Chukhno is also with Universitat Jaume I, Spain.

A. Molinaro is also with Université Paris-Saclay, Gif-sur-Yvette, France.

A. Iera is with University of Calabria, Italy and CNIT, Italy. Email: antonio.iera@unical.it

Several 5G use cases naturally benefit from multicasting capabilities of cellular systems, including content dissemination for enhanced mobile broadband (eMBB), Internet of Things (IoT), Vehicle-to-everything (V2X) communication, and public safety applications, among others [7]–[9]. For these applications, multicast service provides an effective way to deliver information to multiple recipients by minimizing the amount of resources utilized at the air interface and enhancing user experience [10], [11]. Furthermore, the utilization of the multicast mode can offer benefits in terms of link bandwidth saving and application throughput enhancement and is expected to minimize the power consumption of user equipment (UE) [12].

5G NR technology comes with several unique solutions [13], [14]. To compensate for high propagation losses and to efficiently suppress inter-cell interference, NR relies upon large antenna arrays forming directional radiation patterns [15], especially at the NR base station (BS) side [16]. The latter induces the inherent trade-off for multicast service [17], i.e., the use of smaller half-power beamwidths (HPBW) allows for expanding the coverage of NR BSs due to higher gain but decreases the number of UEs that can be served in a single transmission [18]. Furthermore, the use of advanced antenna designs, allowing to form multiple beams at the same time [19], [20], adds a further dimension to the already complex problem.

This work analyzes multicast operation in 5G NR systems with directional multi-beam antennas and offers a computationally efficient solution for optimal multicast grouping. We first formalize the problem mathematically as a subclass of multi-period variable cost and size bin packing problem (BPP) and provide the exact algorithm for computing the optimal solution based on the branch-and-bound technique. We then propose machine learning (ML) algorithms, including decision trees, random forests, and several types of neural networks for multicast grouping. Here, the exact solution, which is only feasible for a limited number of UEs in the multicast group, is utilized to obtain a training dataset for ML algorithms. The performance of the algorithms is finally compared based on the minimum amount of used resources as a metric of interest.

The main contributions of this work are as follows:

- we characterize multi-beam mmWave BS operation under realistic assumptions at the system level by modeling it as multi-period variable cost and size BPP minimizing the amount of resources required to serve UEs and offering the optimal solution;

This is the post-print of the following article: Chukhno, N., Chukhno, O., Moltchanov, D., Gaydamaka, A., Samuylov, A., Molinaro, A., Koucheryavy, Ye., Iera, A., Araniti, A., 2022. The Use of Machine Learning Techniques for Optimal Multicasting in 5G NR Systems. Article has been published in final form at: https://ieeexplore.ieee.org/document/9904964. DOI: 10.1109/TBC.2022.3206595 Copyright © 2022 IEEE.

- formulating the task of optimal multicasting in 5G NR systems with directional multi-beam antennas as a classification problem, we implement and compare the set of ML algorithms (logistic regression, decision trees and forests, and neural networks) and test their efficacy;
- we provide selected numerical results showing that the use of trees with a high number of splits provides the best trade-off between the accuracy of multicast grouping and computational complexity;
- we offer practical considerations for the optimal number of beams (subgroups) that needs to be utilized, showing that there is only a small range of distances where comprehensive multicasting grouping algorithms need to be utilized; specifically, for the considered system and environmental parameters, one subgroup is selected for BS coverage up to 225 m, unicast transmissions to each user are used when BS coverage is higher than 275 m, while for the range 225-275 m the optimal number of subgroups need to be calculated by using the proposed method.

The rest of the text is organized as follows. First, in Section II, we review the related studies. Then, in Section III, we introduce the system model. In Section IV, we formalize the problem mathematically, characterize its complexity and provide the solution algorithm. In Section V, we introduce the ML methods to solve the problem. Numerical results are elaborated in Section VI. Finally, conclusions are drawn in the last section.

II. Related Work

The 5G multicast/broadcast requirements can be classified into two different operation modes: *stand-alone* deployment of dedicated broadcast networks and the *mixed* unicast/multicast mode. Requirements for mixed-mode (MM) multicast aim to incorporate point-to-point (PTP) transmissions in the Radio Access Network (RAN) as a built-in network delivery optimization feature. This requires seamless switching between PTP and point-to-multipoint (PMP) transmissions [21]. Particularly, 5G NR MM is proposed in the 3GPP Release 17 (Rel-17) to enable the use of multicast and provide a flexible, dynamic, and seamless switching between unicast and multicast or broadcast transmissions and traffic multiplexing under the same radio structures. In [22], the authors propose the RAN multicast area (RMA) mechanism to seamlessly switch between multicast and unicast modes, which takes into account UE activity, the number of devices, and their geographical distribution.

A. Single-Beam Antennas

Several works have been performed on multicast grouping (stand-alone deployment) and subsequent resource utilization problems in directional systems with a single antenna lobe. In [23], a heuristic algorithm that aims to construct the optimal multicast subgroups in IEEE 802.11ad networks is presented to achieve high throughput for scattered multicast devices. Specifically, multicast beamforming is performed through the process of an association beamforming training interval and is divided into two steps. First, the devices are grouped based on the distances between them. That is, those UEs, whose distances are smaller than a reference value, belong to the same set. Then, by utilizing the law of cosine with respect to the coordinates of two edge UEs in the set, the beamwidth and optimal data rate (according to the modulation and coding scheme (MCS) table) are obtained for each multicast subgroup.

An alternative heuristic solution for multicast grouping with an adaptive beamwidth is proposed in [18], wherein the beamwidth that maximizes the sum rate of devices is chosen incrementally. More precisely, the beamwidth is adaptively determined based on the UEs locations and the requested data rates. According to the presented simulation results, the proposed scheme can improve the overall throughput by 28% to 79% compared to the existing multicast schemes that consider only a fixed data rate to cover all sectors or one sector.

Unlike the studies mentioned above, those in [24] and [25] investigate both optimal and sub-optimal multicast schemes for mmWave communications in NR with directional beams. The solution aims to reach an optimal trade-off between achieving a high signal-to-noise ratio value by utilizing narrow beams and serving many UEs simultaneously, thereby reducing the channel usage time. The optimal solution requires solving a Markov Decision Process (MDP) with a large state and action spaces, thus resulting in super-exponential complexity in the number of UEs. To this end, the authors reduced the complexity of the problem by introducing a practical hierarchical solution.

B. Multi-Beam Antennas

Although many previous works on mmWave communications focused on optimizing multicast data transmission using single-lobe antennas, the problem of multicast grouping and associated optimal resource allocation for multi-beam systems has received limited attention so far. In [26], *switched* beamforming-multicasting[1] trade-off is investigated to minimize the total time required for 100% guaranteed packet delivery to all multicast UEs. The authors consider both continuous (Shannon capacity) and discrete rate functions under two power allocation models, where the power is either equally split (EQP) or asymmetrically split (ASP) among the lobes. Both optimal and heuristic solutions are designed for the continuous rate function, while greedy solutions are provided for the discrete rate case. It is also shown that in the continuous rate case, the greedy solution (GREP) provides near-optimal performance, almost coinciding with the optimal one.

Since neither optimal nor approximate solutions have been proposed in [26] for the ASP model, the authors in [27] present several key results. Specifically, for the EQP model, the authors provide a dynamic-programming-based optimal solution with low complexity of $O(B^2)$, compared to the $O(B^7)$ complexity of the optimal solution in [26], where B is the total number of non-overlapping single-lobe beams, for both continuous and discrete rate functions. Even though the problem

[1]Switched beamforming consists of pre-determined beams that cover the entire azimuth of 360°.

formulated in [27] is characterized by the polynomial-time optimal solutions under the EQP model, the same problem under the ASP model is NP-hard. Applying generalized-bin-packing algorithms for the discreet rate function allows for obtaining asymptotic polynomial-time approximation schemes. The solution enables drastic improvement over GRASP2 in [26], which handles ASP for the discrete rate case.

Unlike [26], [27], *adaptive* beamforming-based multicast system with multi-lobe antenna pattern is designed in [28] with the same goal of minimizing the time of the data dissemination to the multicast UEs. In case of adaptive beamforming, the formulated problem is stated to be non-convex and NP-hard for both discrete and continuous versions. Hence, the authors present efficient algorithms implemented in an adaptive beamforming system for multicasting (ADAM) and suitable for a practical system design with a complexity of $O(BK^2K!)$, where K is a number of multicast UEs in the multicast group. In [29], an optimal multicast grouping and resource allocation problem is formulated based on a variable-sized bin packing problem, which is known to be NP-hard. To this end, several heuristics with different complexities and approximation accuracy are developed to provide practical algorithms with reduced computational requirements.

C. Machine Learning Solutions

Unlike the above-mentioned studies, several researchers focus on the beam direction, beam weights, power predictions, and transmission mode selection using ML methods. In [30], a D2D-assisted multicast grouping that leverages an unsupervised ML algorithm is designed. In [31], the authors utilize reinforcement learning for the smart mode selection (macrocell broadcasting, mmWave small cell unicasting, and D2D multicasting) to find the optimal transmission strategy in every time slot. However, neither multicast grouping nor multi-beam antennas are considered. Alternatively, in [32], the vehicle situational awareness is used as an input for ML algorithms to predict the received power of each beam in the codebook with low or almost zero feedback overhead. The authors compare the prediction results among linear regression, support vector regression, random forest regression, and gradient boosting regression. It is shown that the random forest is a good fit for the specific dataset since it is able to implicitly select the features and generalizes well by ensembles. In [33], ML-assisted beam training is performed for mmWave cellular systems using realistic beamforming data and GPS coordinates of UEs by leveraging Random Forest Classifiers and Multilayer Perceptrons.

In this work, we focus on supervised learning algorithms for classification of users to the multicast subgroups [34]. The following classes of classification algorithms are considered:

- *Decision Trees* are supervised algorithms used both for classification and regression. Their main advantage is the building of an interpretable model. Thus, they are also known as white-box algorithms [34]. We provide more details on decision trees in Section V-C1.
- *Logistic Regression* is used for the classification problems to assign observations to a discrete set of classes. This technique transforms the output by using the logistic sigmoid function to return a probability value class mapping.
- *Naive Bayes* is a simple but powerful classification algorithm, "probabilistic classifier", based on Bayes' theorem with the assumption of conditional independence among considered features of objects.
- *Support Vector Machine* (SVM) is an algorithm that can distinguish between two or more classes by defining a hyperplane that separates those classes. The support vectors are the closest points to the hyperplane. A change in the support vectors results in a modification of the hyperplane [35]. SVM can be used for solving classification and regression problems.
- *K-Nearest Neighbors* (KNN) is a well-known algorithm used both for solving classification and regression problems. The output of the algorithm is obtained by comparing the input with known data [35].
- *Neural network* (NN) classifiers are used for multiclass classifications. These models typically outpace other algorithms in prediction accuracy. However, the flexibility of NN models increases with the number and size of connected layers. We offer more details on NN models in Section V-C3.

D. Summary

To summarize, optimal multicasting is a complex problem for both single- and multi-beam mmWave systems that can be solved exactly for a limited set of UEs only. Specifically, in [24], authors demonstrate the results considering only 8 UEs in the system, whereas only 12 UEs are considered in [29] due to the solution time complexity as well as memory challenges.

In this work, to efficiently solve the multicast problem in NR systems for a meaningful number of UEs, we advocate the use of ML techniques. No accurate and computationally efficient solutions have been proposed so far. In contrast to existing works, we also consider adaptive multi-lobe beamforming antennas, which further complicates the problem. To this aim, in order to determine a computationally efficient and accurate solution of the multicasting problem, in what follows, we first provide the exact solution [29] and then utilize it to train candidate ML alternatives.

III. System model and Assumptions

This section introduces our system model by specifying deployment, traffic, resource, propagation, blockage, and antenna models. We also present our metrics of interest. The notation used throughout this paper is offered in Table I.

A. Deployment, Traffic, and Resource Models

We assume a tri-sector cellular deployment option illustrated in Fig. 1 for our NR system by concentrating on a randomly chosen sector of a typical "cell". K UEs are uniformly distributed in the cell sector. These UEs are assumed to request a single multicast session in the downlink direction with the bitrate of C Mbps. The height of UEs, NR BS, and blockers are assumed to be constant and given by h_U, h_A, and h_B,

TABLE I
MAIN NOTATION.

Parameter	Definition
K	Number of multicast UEs in multicast group
C	Bitrate of multicast session, Mbps
h_U	Height of UEs, m
h_A	Height of NR BS, m
h_B	Height of blockers, m
W	Available bandwidth, MHz
f_c	Carrier frequency, GHz
μ	5G NR numerology
w_{PRB}	Size of PRB, MHz
y	Three-dimensional distance between UE and NR BS, m
$S(y)$	Distance-dependent SINR, W
P_A	NR BS transmit power, W
G_A, G_U	Antenna array gains at NR BS and UE ends, dBi
M_I	Interference margin, dB
N_0	Power spectral density of noise, dBi/Hz
$L(y)$	Path loss in linear scale
$L_{dB}(y)$	Path loss in decibel scale
A, ζ	Propagation coefficients
$p_B(y)$	Distance-dependent blockage probability
r_B	Radius of blockers, m
λ_B	Blockers density, bl./m^2
α	HPBW of a linear antenna array, rad
$\theta^{\pm}_{3db}$	Upper and lower 3-dB points of antenna array, $^\circ$
θ_m	Location of array maximum, $^\circ$
β	Antenna array orientation, $^\circ$
N	Number of planar antenna array elements
L	Number of beams in the system
$P_{\max}$	NR BS total available power, W
M	Number of time slots in 1 ms subframe
R_b	Number of available primary resource blocks, PRBs
S_{th}	SINR threshold, dB
s_j	Spectral efficiency of the worst UE in subgroup j, bit/s/Hz
R	Service (cell) area radius, m

respectively. We utilize W MHz bandwidth available for a sector antenna, consider the carrier frequency of $f_c = 28$ GHz, and the corresponding NR numerology $\mu = 3$ with the physical resource block (PRB) size of $w_{\text{PRB}} = 1.44$ MHz. We assume the orthogonal frequency division multiple access (OFDMA) scheme as specified in NR [36]. Depending on UE locations, the number of PRBs required to serve multicast UEs might differ and can be computed using NR MCS according to the procedure outlined in, e.g., [37].

By following [38], we use the term term *subgroup* to denote the subset of UEs belonging to the multicast group served by the same beam, whereas a *multicast group* contains all UEs interested in receiving a *multicast session* (i.e., data flow/content). With the term *suit* we imply a configuration of multicast subgroups that covers all UEs (i.e., a multicast group) without repetitions.

B. Propagation and Blockage Models

The Signal-to-interference-plus-noise-ratio (SINR) at the distance y between the receiver and the NR BS is given by

$$S(y) = \frac{P_A G_A G_U}{(N_0 W + M_I) L(y)}, \tag{1}$$

where P_A is the transmit power, G_A and G_U are the antenna gains at the NR BS and the UE, N_0 is the thermal noise power, W is the bandwidth, M_I is the interference margin, whereas $L(y)$ is the path loss in linear scale. The path loss measured in dB is determined according to [39] as

$$L_{dB}(y) = 32.4 + 21 \log_{10} y + 20 \log_{10} f_c, \tag{2}$$

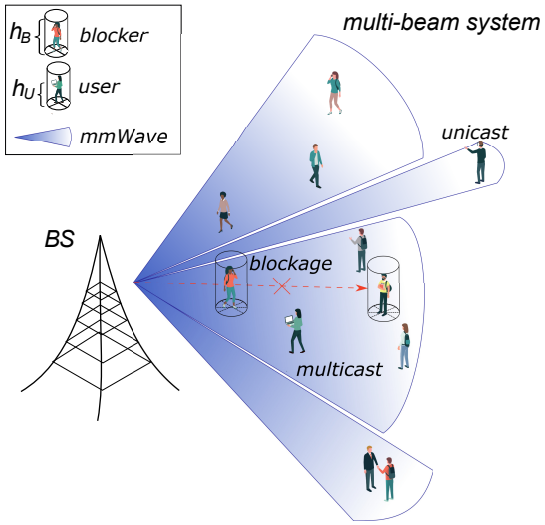


Fig. 1. The considered deployment scenario.

where f_c is the carrier frequency in GHz and y is the three-dimensional (3D) distance between the NR BS and the UE.

The path loss in the form of (2) can be represented in the linear scale by utilizing the model in the form of Ay^{ζ}, where A and ζ are the propagation coefficients. By introducing the coefficients (A_1, ζ) and (A_2, ζ) that correspond to LoS non-blocked and blocked conditions, we have

$$A_1 = 10^{2 \log_{10} f_c + 3.24}, A_2 = 10^{2 \log_{10} f_c + 4.74}, \zeta = 2.1. \tag{3}$$

We assume the blockage attenuation by the human body is 15 dB [40] and model blockers as cylinders with the height and the radius of h_B and r_B. The blockers are assumed to follow Poisson distribution in $\Re^2$ with density λ_B. Hence, the probability of blockage at distance y is calculated by [41] as

$$p_B(y) = 1 - \exp^{-2\lambda_B r_B \left[\sqrt{y^2 - (h_A - h_U)^2} \frac{h_B - h_U}{h_A - h_U} + r_B\right]}. \tag{4}$$

C. Antenna Model

We assume a cone antenna model, where the radiation pattern is represented as a conical zone with an angle α coinciding with the HPBW of the antenna array. The HPBW, α, is determined by [42] as

$$\alpha = 2|\theta_m - \theta_{3db}|, \tag{5}$$

where θ_{3db} is the angle at which the value of the radiated power is 3dB below the maximum, whereas θ_m is the location of the array maximum and is given by $\theta_m = \arccos(-\beta/\pi)$, where β is the phase excitation difference affecting the physical orientation of the array.

The antenna gain is determined by [42] as

$$G = \frac{1}{\theta^{+}_{3db} - \theta^{-}_{3db}} \int_{\theta^{-}_{3db}}^{\theta^{+}_{3db}} \frac{\sin(N\pi \cos(\theta)/2)}{\sin(\pi \cos(\theta)/2)} d\theta, \tag{6}$$

where the upper and the lower 3-dB points are

$$\theta^{\pm}_{3db} = \arccos[-\beta \pm 2.782/(N\pi)], \tag{7}$$

and N is the number of antenna elements.

We also assume that several beams, $L \geq 1$, can be formed simultaneously on the BS NR and be steered in different directions splitting the total power, $P_{\max}$. The HPBW of these beams depends on the number of antenna elements and is limited by a lower bound (5).

D. Metrics of Interest

To optimize the mmWave NR resource utilization with multi-beam directional antennas, we consider the ratio of occupied resources to the overall set of available resources, ρ, as the optimization criterion. By solving the optimization problem, we also determine the inter-site distance D and, thus, η – the minimum NR BS deployment density required for multicast service provisioning [29].

We emphasize that both multicast and unicast traffic is simultaneously served at BS in real-life scenarios. However, the coexistence of these traffic types is a complex problem that is outside the scope of this paper. For example, network operators may inherently prioritize the multicast traffic as more UEs will be served with a single transmission, increasing overall network utility. Alternatively, the priority might be given to unicast traffic. Thus, in what follows, without loss of generality, we neglect unicast traffic.

IV. Problem Formalization

This section introduces our framework for optimal resource allocation of UEs in a multi-beam environment [29]. We first formulate the problem as a bin packing formalism. We then discuss the complexity of the proposal.

We consider a multicast group formed by K UEs deployed in the 5G NR BS sector coverage served by a multi-beam system, $L \geq 1$. We assume the OFDMA scheme with M time slots in the time horizon $t \in \mathcal{T}$, $\mathcal{T} = \{1, \ldots, M\}$. The maximum number of PRBs available in the system is MLR_b, where R_b is the available number of resource blocks for the beam in the time slot t. The potential maximum number of subgroups served during the time horizon is limited to ML.

The set of K multicast UEs of the multicast group is denoted by $\mathcal{K} = \{1, \ldots, K\}$. There are $2^K - 1$ options for assigning K UEs to multicast subgroups [24], that is, $\mathcal{K}_j$ is a set of UEs forming a subgroup j, $j \in \mathcal{J}, \mathcal{J} = \{1, \ldots, 2^K - 1\}$, whereas $|\mathcal{K}_j|$ is the number of UEs in subgroup j. We define $\mathcal{G}_k$ as a collection of subgroup's indices from set $\mathcal{J}$ such that the corresponding subgroups cover all the UEs from set $\mathcal{K}$ without repetition. Let Ω be the set of all such collections of subgroup's indicies. We note that each multicast UE belongs to one and only one subgroup; therefore, $\mathcal{G}_k$ is restricted by $\bigcup_{j \in \mathcal{G}_k} \mathcal{K}_j = \mathcal{K}$ and $\mathcal{K}_{j_1} \bigcap \mathcal{K}_{j_2} = \emptyset, j_1 \neq j_2, \ \forall j_1, j_2 \in \mathcal{G}_k$. We also define the so-called "suits" $\mathcal{G}_k^l$ as the subset of indices from $\mathcal{G}_k$, which is planned for beam l by the scheduler, $\mathcal{G}_k^l \subseteq \mathcal{G}_k$, $l = 1, 2, ..., L$. Therefore, suits $\mathcal{G}_k^l$ satisfy $\mathcal{G}_k = \bigcup_{l=1}^{L} \mathcal{G}_k^l$ and $\mathcal{G}_k^{l_1} \bigcap \mathcal{G}_k^{l_2} = \emptyset, l_1 \neq l_2, \ \forall l_1, l_2 \in \{1, 2, ..., L\}$.

We model the optimization problem by introducing a binary indicator, $g_j^t \in \{0, 1\}$, which denotes the subgroup assignment decision variable. Let $g_j^t = 1$ if subgroup j is served at time slot t, and $g_j^t = 0$ otherwise. Then, we have a vector-indicator, $\mathbf{g}^t = (g_1^t, \ldots, g_{|\mathcal{J}|}^t)$, of subgroups that are served at time slot t.

We assume the constraint on the maximum number of subgroups to be served (or beams to be swept) in the system at each time slot t, that is,

$$\sum_{j \in \mathcal{G}_k} g_j^t \leq L, \forall t \in \mathcal{T}. \tag{8}$$

A suit service time should not exceed the time horizon, i.e.,

$$\sum_{j \in \mathcal{G}_k^l} \sum_{t \in \mathcal{T}} g_j^t \leq M, \forall l = 1, ..., L, \forall k = 1, ..., |\Omega|. \tag{9}$$

We also have to ensure the following constraint to be held on the transmit power budget per antenna that serves subgroup j,

$$\sum_{j \in \mathcal{G}_k} g_j^t P_j \leq P_{\max}, \forall t \in \mathcal{T}, \tag{10}$$

where P_j is the transmit power of beam that serves subgroup j and is calculated as

$$P_j = \frac{A_1 A_2 S_{th} (N_0 W + M_I)}{G_{\mathrm{A}} G_{\mathrm{U}} d_j^{\zeta} \left[A_2 (1 - p_B(d_j)) + A_1 p_B(d_j) \right]}, \tag{11}$$

where S_{th} is the SINR threshold corresponding to a chosen NR MCS, where A_1, A_2, and ζ are the propagation coefficients, d_j is the NR BS-worst multicast UE distance.

We assume that the session requires a constant bit rate C. Then, cost a_j is represented in terms of the number of PRBs for the assigned beam for subgroup j and is given by

$$a_j = \frac{C}{s_j w_{\mathrm{PRB}}}, \tag{12}$$

where s_j is a spectral efficiency in bit/s/Hz of the worst UE in subgroup j, w_{PRB} is a PRB size.

Note that the scheduler's time slot assignment is reflected in vector $\mathbf{g}_j = (g_j^1, \ldots, g_j^M)$ with

$$\sum_{t \in \mathcal{T}} g_j^t = \left\lceil \frac{a_j}{R_b} \right\rceil, \forall j \in \mathcal{J}. \tag{13}$$

We assume that the scheduler assigns a beam to the subgroup such that the following holds true

$$a_j \leq MR_b, \forall j \in \mathcal{J}. \tag{14}$$

Finally, in constraints (9) and (14), the following condition for the maximum available resources in the system should be satisfied

$$\sum_{j \in \mathcal{G}_k} a_j \leq MLR_b, \forall j \in \mathcal{J}, k = 1, ..., |\Omega|. \tag{15}$$

Recall that the goal is to determine the optimal grouping of multicast UEs, which minimizes the total cost of service in terms of the ratio of occupied PRBs to the total available number of PRBs for the entire time horizon. We formalize the multi-beam operation optimization problem as a special class of BPP. Thus, the optimization problem takes the form of

$$\min_{k \in 1, \ldots, |\Omega|} \sum_{j \in \mathcal{G}_k} \frac{a_j}{MLR_b}, \tag{16}$$

$$\text{s.t. } (8), (9), (10), (14), (15).$$

Algorithm 1: Optimal Solution [29]

Input: Deployment of UEs, $i \in \mathcal{K}$
Output: Optimal solution for multicast grouping
Create $2^K - 1$ multicast subgroups
for *each subgroup* $\mathcal{K}_j$ **do**
  find the farthest UE i and the distance from BS to this UE: $y \leftarrow \max_{i \in \mathcal{K}_j} y_i$;
  find HPBW needed to cover the subgroup $\mathcal{K}_j$;
  find P_j from (11) using $d_j = y$;
  find the cost a_j from (12);
end
Solve the problem (16) with exhaustive search.

Here the objective function $\sum_{j \in \mathcal{G}_k} \frac{a_j}{MLR_b}$ is the ratio of the amount of utilized resources for k-th multicast grouping to all the resources in the system. We refer to such ratio as ρ and use it as the key metric in our work.

The Algorithm 1 describes the globally optimal solution. The complexity of the Algorithm 1 is determined by the underlying BPP. Thus, Algorithm 1 is NP-hard, while the associated complexity is exponential. We note that the proposed solution cannot solve the problem in a reasonable time when the number of UEs in a multicast group is higher than 12. Table II offers execution times in minutes for different UE density scenarios.

V. Machine Learning Solution Algorithms

In this section, we propose the set of ML techniques for resource optimization for multicasting. As the problem of interest is the classification of users into multicast subgroups, we consider three classes of algorithms with varying complexity suitable for this task, including (i) logistic regression model, (ii) decision trees and forests, and (iii) neural networks.

A. Type of ML Problem: Classification

In this work, the problem at hand is a classification of users into multicast subgroups. Furthermore, due to availability of limited training data, supervised ML-based classification algorithms can be utilized. The exploited ML algorithm has to be as simple as possible to run in real-time, on the BS side, when a new UE joins the multicast group or leaves it. Therefore, the execution time and the training phase (preferably) have to be small. Since we aim at a practical implementation, low complexity ML tools receive the priority in what follows. That is, we first consider simple supervised classification algorithms based on logistic regression and decision trees. To evaluate whether advanced supervised classification techniques may provide more accurate results, we further consider random forests and neural networks. Note that the decision tree is computationally faster in comparison to the random forest because of the ease in generating rules. Several factors need to be considered in a random forest classifier to interpret the patterns among the data points.

[2]60 minutes is an execution time limit. When reaching 60 minutes, the algorithm provides the current solution that can be different from optimal.

TABLE II
Execution Time in Minutes for Algorithm 1

Time/K	2	5	7	10	12	15
Optimal [29]	0.008	0.01	0.06	10.03	54.35	60[2]

B. Data Preprocessing, Features, and Metrics

We aim to use the obtained data from the direct solution for a limited number of UEs in a multicast group to design an algorithm capable of solving it for more UEs in a group, thereby utilizing supervised algorithms. In a supervised learning model, the algorithm *learns* on a labeled dataset (e.g., data from the direct solution) and provides the results that the algorithm can evaluate in terms of accuracy based on the training set. The algorithm can be implemented as an offline learning tool within the recently standardized ML framework for 5G systems, see [43]–[45].

We consider the following model's features as parameters that form the dataset for the training of supervised algorithms: *(i)* UE's coordinates, *(ii)* number of UEs K, *(iii)* service area radius R, *(iv)* bandwidth W, *(v)* session data rate C, and *(vi)* number of clusters (subgroups). Model's features form predictor's set $\mathcal{P}$. We choose these parameters since they all may affect the results of the classification. Later, in the numerical result section, we explore which of those parameters are more important, i.e., have the higher importance level. The algorithms learn from the dataset (provided by the optimization presented in Section IV) by predicting the data and adjusting it for the correct answer of multicast subgroup formation 5G NR systems.

To evaluate the accuracy of ML algorithms, we introduce two types of similarity metrics. The first metric is based on the number of clusters and UEs in each cluster. More precisely, the following criterion is used: $\frac{\text{number of correctly classified data}}{\text{number of test data}} \cdot 100\%$. The second metric measures the actual resource usage and is thus considered the ultimate metric in our work.

C. Utilized ML Algorithms

1) Decision Trees and Forests: The idea of a decision tree is to form queries with which the algorithm accesses data. When using the classification and regression trees algorithm (CART), questions, known as node separations, are defined to reduce the Gini impurity index

$$\text{Gini}_i = 1 - \sum_{k=1}^{J} p_{i,k}^2, \tag{17}$$

where $\mathcal{J}$ is a set of classes, $|\mathcal{J}| = J$ is a number of classes, $p_{i,k}$ is the probability that an observation i belongs to class k, $k \in \mathcal{J}$.

In this work, we consider two decision tree algorithms, namely, Fine and Coarse trees, that differ in terms of accuracy and model flexibility. A fine tree contains many leaves (which helps to make many fine distinctions between classes) and is usually highly accurate on the training dataset. However, such

a leafy tree tends to overtrain, and its validation accuracy is often far lower than the training one. Differently, a coarse tree with a lower number of leaves does not attain high training accuracy but can be more robust because its training accuracy can approach that of a representative testing dataset. Hence, we aim to compare those two borderline tree algorithms in terms of accuracy and practical applicability.

Among ensemble classifiers, we select Random Forest (bagged trees) since Boosted trees can usually perform better but might require searching many parameter values, which is time-consuming. Random Forest classifier is a model consisting of a set of decision trees, see Algorithm 2. Instead of averaging different trees' predictions (this concept is called a "forest"), this model uses two key concepts that make the forest random. First, samples are randomly chosen by bootstrapped training samples $\mathbf{Z}^*$ from a dataset when building trees (line 2). Second, when splitting nodes, random sets of v parameters (i.e., variables) are selected (line 4). An unpruned tree is grown from each bootstrap sample so that at each node, v predictors are randomly selected as a subset of predictor variables out of p predictors from predictors' set $\mathcal{P}$ (usually $v = \sqrt{p}$), and the best separation from among these variables is selected (lines 4-5). We note that the selected number of prediction variables that ensure sufficiently low correlation with adequate predictive power is critical. The final forecast is made by averaging the forecasts from all the trees (lines 8-9). Further, the variables' importance can be calculated by exploiting the out-of-bag (OOB) data [46]. Here, each variable is randomly permuted, and the permuted cases of the OOB are sent down the tree again. Subtracting the number of correctly classified cases using permuted data from the number of correctly classified cases using non-permuted data reveals the importance value of the variable. For each tree, these values are different. Averaging values across all the trees in the forest provides an importance score for each variable.

Moreover, suppose that we fit the algorithm with features (or, equally, variables) that are not useful. In that case, the algorithm simply will not use them to split into the data and get the lowest importance level. We expect that UEs' coordinates, BS coverage radius, and the number of clusters will receive the highest importance in our task. We emphasize that random forests forces each split to consider only a random subset of v predictors, making the resulting trees' average less variable and, hence, more reliable.

2) Logistic Regression, Naive Bayes, SVM, and KNN: Logistic regression is one of the most common classification algorithms, especially if there are only two classes to be categorized. In statistics, the binary logistic regression model is a statistical model that models the probability of one event (out of two alternatives) taking place by having the logarithm of the odds for the event be a linear combination of one or more independent variables (features). Later, in Section VI, we demonstrate that Logistic regression works well until the number of classes is higher than two.

Naive Bayes classifiers are useful for multiclass classification. The naive Bayes algorithm leverages the Bayes theorem and assumes that predictors are conditionally independent, given the class. Kernel Naive Bayes, more flexible than Gaussian ones, can work with our dataset. Support Vector Machine (SVM) and Nearest Neighbor (KNN) Classifiers are also popular ML algorithms. KNN classifiers typically have good predictive accuracy in low dimensions but might not when dimensions are large. We found that Cubic SVM and Weighted KNN Classifiers give the best result within the classifier type.

Algorithm 2: Random Forest Classification

1 **Input** training set, including predictors set $\mathcal{P}$ of size p; number of selected predictors v; number of trees B; **for** *b=1 to B* **do**
2 Draw a bootstrap sample $\mathbf{Z}^*$ from the training dataset;
3 Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each node of the tree until the minimum node size $n_{\min}$ is reached;
4 (i) Select v variables at random from p variables;
5 (ii) Pick the best variable/split-point among v;
6 (iii) Split the node into two daughter nodes;
7 **end**
8 **Output** the ensemble of trees $\{T_b\}_1^B$.
9 To make a prediction at a new observation x let $J_b(x)$ be the class prediction of the b-th random forest tree. Then, $J_{\text{rf}}^B(x) =$ majority vote$\{J_b(x)\}_1^B$.

3) Neural Networks: Neural Network is an ML system that emulates brain behavior. The basic component is the *perceptron*, which is composed of linear and non-linear parts. The former is a weighted sum, whereas the latter is an activation function. Examples of activation functions are rectified linear unit (ReLU), exponential linear unit (ELU), and leaky ReLU. An NN is composed by a set of perceptrons divided into layers of three types: input layer, hidden layer, and output layer. Two particular types of NNs are Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). CNN involves the use of matrix multiplications and convolutional filters to reduce the input size. RNNs are capable of analyzing and predicting time series and are characterized by loops between layers [35].

NN classifiers represent fully connected neural networks for the classification tasks. The first neural network layer has a connection to the network input (dataset created with the help of optimization of multicast grouping), and each subsequent layer connects with the previous one. Each layer multiplies the input data by a weights matrix and adds a bias vector. The activation function tracks each layer (the activation function for the last fully connected layer is always softmax). The last layer and the subsequent softmax activation function deliver the output data of the network (multicast grouping).

We note that there are different NN types, such as Bilayered, Narrow, Medium, Wide, and Trilayered neural networks. In our case, Narrow NN shows the best accuracy on our dataset. In Table III, we summarize the ML algorithms used in this paper and the types the algorithms belong to. Generally, we use the best algorithm from a given class. However, we test

TABLE III
UTILIZED ML ALGORITHMS.

Algorithm Type	Algorithm
Decision Trees	Fine Tree
Decision Trees	Coarce Tree
Ensemble Classifiers	Random Forest (bagged trees)
Logistic Regression Classifiers	Logistic Regression
Naive Bayes Classifiers	Kernel Naive Bayes
Support Vector Machine	Cubic SVM
Nearest Neighbor Classifiers	Weighted KNN
Neural Network Classifiers	Narrow NN

TABLE IV
DEFAULT PARAMETERS FOR NUMERICAL ASSESSMENT.

Parameter	Value
Operating frequency, f_c	28 GHz
Bandwidth, W	50 MHz
PRB size, w_{PRB}	1.44 MHz
Subcarrier spacing, Δ	0.12 MHz
Height of BS, h_A	10 m
Height of blocker, h_B	1.7 m
Height of UE, h_U	1.5 m
Interference margin, M_I	3 dB
SINR threshold, S_{th}	-9.47 dB
Transmit power, P_A	33 dBm
Power spectral density of noise, N_0	-174 dBm/Hz
UE planar antenna elements, N	4 el
UE receive gain, G_U	5.57 dBi
Session data rate, C	25 Mbps
BS antenna array	32×4
BS transmit gain, G_A	14.58 dBi
Service area radius, R	250 m
Number of UEs, K	2-30
Subframe duration	1 ms
Slot duration	125 μs
5G NR numerology, μ	3
Number of time slots in 1 ms subframe, M	8
Number of available resource blocks, R_b	32
Number of beams available in the system, L	1,3,5

more trees since their computational complexity is low.

VI. NUMERICAL RESULTS

In this section, we provide our assessment campaign. Our simulations are performed in the MATLAB environment. Note that the algorithms are implemented using the reference MATLAB implementations by adapting them to the specifics of the multicast deployment described in Section III. To evaluate the performance of the considered ML algorithms for optimal multicasting in mmWave systems and identify the best candidates, we utilize the following approach. First, we consider using ML algorithms for 10 UEs and compare the performance with the exact solution obtained by the proposed optimization algorithm. At this stage, we also identify the training set size sufficient for the best performing ML candidates. Then, in the next step, we proceed with an assessment of the extrapolation capabilities of the best ML candidates identified in the first step. To this aim, we train ML algorithms based on 10 UEs in the cell coverage and apply them to evaluate the case of 13 UEs. By comparing these results to that of the optimization framework, we further discriminate the considered candidates identifying those providing the best extrapolation capabilities. Finally, we report the performance evaluation results by utilizing the specified algorithms for standard 3GPP scenarios with 30 and 60 UEs. The parameters used in this Section are provided in Table IV.

Recall that the identified metric of interest in the optimization framework is the ratio of the amount of utilized resources to all the resources in the system, ρ, as per (16). The first metric of interest that we utilize for accuracy assessment, σ, is based on the exact matching of the number of multicast subgroups and UEs assigned to these subgroups. We consider the data to be classified correctly if the ML classification resulted in the same number of multicast subgroups and the same composition of UEs for each subgroup as the result of the optimization framework. Observe that if the match between ML and optimization results is perfect, the considered metrics ρ_{opt} and ρ_{ML} coincide. However, due to the discrete nature of resource allocation and mapping between MCSs and spectral efficiency, the considered metrics ρ_{opt} and ρ_{ML} might be close even when the different number of subgroups and UEs assignment to these subgroups is observed. Since resource utilization is the main metric of interest, in addition to perfect matching between UEs assignments, we also consider the metric $\gamma = (\rho_{ML}/\rho_{opt})100\%$. This metric measures the closeness of resource allocation produced by the considered ML algorithm and optimization framework.

We utilize 30 and 60 UEs in the numerical results as these values are generally recommended by 3GPP [47] for performance assessment of various LTE and 5G NR functionalities. Note that due to the presumed usage type of the algorithm, we do not consider the operation of the algorithm in dynamics, i.e., when UEs enter and leave the multicast session. UE arrival and departure from the multicast session may introduce rather large changes in multicast subgroup formations. To this aim, we assume that whenever UE leaves the multicast session or new UE joins it, the algorithm has to be rerun producing new subgroups.

A. Accuracy Assessment of ML Algorithms

We start with the accuracy assessment of the ML candidates for optimal multicasting by presenting the results for 10 UEs uniformly distributed in the cell coverage area. Here, we train all the algorithms by preparing the training sets of length H_1 for 10 UEs in the system, apply them to the testing sets of H_2 in length, and, finally, compare the results to those of the optimization framework with 10 UEs.

First, we take into account the parameter σ to explore how closely the ML algorithms classify UEs to the multicast subgroups as a function of the cell radius R as illustrated in Fig. 2 and Table V for $H_1 = 5000$, $H_2 = 5000$. Here, we see that up to the cell coverage of $R = 225$ m there is a perfect match between the optimal solution and all the considered algorithms. However, the rationale is that up until this distance, only one multicast subgroup is utilized for the optimal solution, which is correctly predicted by the algorithm. Starting from $R = 250$ m, we see that all the algorithms begin to deviate from the optimal solution, with the best providing 70% of matching accuracy at most. It is also interesting to note that most algorithms perform the same way, with only simple

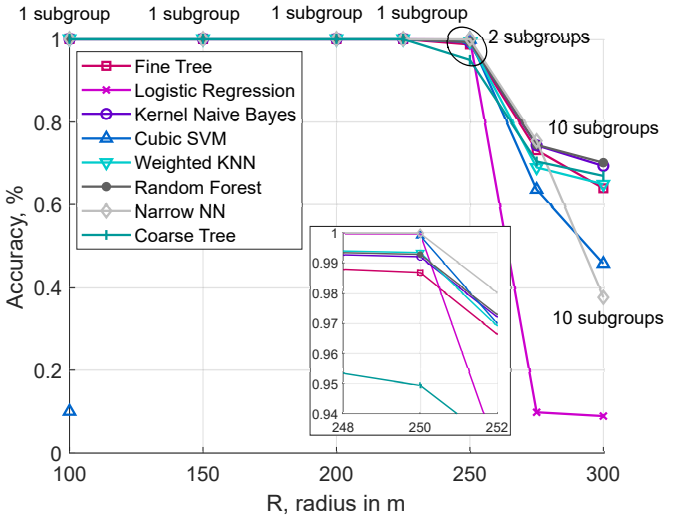


Fig. 2. Group assignment accuracy, σ, for $H_1 = H_2 = 5000$, $K = 10$.

logistic regression providing significantly worse results than the rest. The explanation is the simplicity of the model that cannot capture complex dependencies in the training data.

Note that the perfect matching of UE allocations to subgroups is sufficient for optimal resource usage but *(i)* may not be a unique solution optimizing resource allocation and *(ii)* does not imply that other UE allocations to subgroups result in significantly worse resource utilization. To this aim, Table V also shows the accuracy of resource matching, γ, for the considered ML algorithms with respect to optimal allocations obtained from the exact algorithm. As one may observe, even those algorithms characterized by drastic deviation from the optimal solution in terms of UE allocations to multicast subgroups, σ, show very good matching in terms of resources, γ. Specifically, only one algorithm, Weighted KNN, provides resource matching accuracy of less than 95% for $R = 300\,\text{m}$. Four algorithms, the Fine Tree, Coarse Tree, Random Forest (Bagged Tree), and Kernel Naive Bayes, provide excellent agreement in terms of UE allocations, σ, and resource matching, γ, with the latter being higher than 98% for all the considered distances.

TABLE V
GROUP AND RESOURCE MATCHING ACCURACY, $H_1 = 5000$, $H_2 = 5000$, $K = 10$.

Radius	**100m**	**150-225m**	**250m**	**275m**	**300m**
Fine Tree					
UE assignment, σ	100%	100%	98.68%	73.14%	63.89%
Resources, γ	100%	100%	100%	96.07	98.08%
Logistic Regression					
UE assignment, σ	100%	100%	99.96%	9.80%	8.88%
Resources, γ	100%	100%	100%	98.03%	98.03%
Kernel Naive Bayes					
UE assignment, σ	100%	100%	99.20%	74.30%	69.31%
Resources, γ	100%	100%	100%	98.07%	98.07%
Cubic SVM					
UE assignment, σ	99.98%	NaN	99.90%	63.58%	45.64%
Resources, γ	100%	NaN	100%	93.73%	98.07%
Weighted KNN					
UE assignment, σ	100%	100%	99.34%	68.94%	64.89%
Resources, γ	100%	100%	100%	96.14%	91.81%
Random Forest					
UE assignment, σ	100%	100%	99.28%	74.40%	70.11%
Resources, γ	100%	100%	100%	98.03%	95.99%
Narrow NN					
UE assignment, σ	100%	100%	99.98%	75.42%	37.62%
Resources, γ	100%	100%	100%	98%	96.07%
Coarse Tree					
UE assignment, σ	100%	100%	94.94%	70.40%	66.90%
Resources, γ	100%	100%	100%	97.27%	100%

TABLE VI
TRAINING TIME IN SECONDS, $R = 300$, $H_1 = 5000$, $K = 10$.

Algorithm	**Time, s**
Fine Tree	5.60
Logistic Regression	17.64
Kernel Naive Bayes	38.82
Cubic SVM	51321
Weighted KNN	0.81
Random Forest	5.66
Narrow NN	43.02
Coarse Tree	1.03

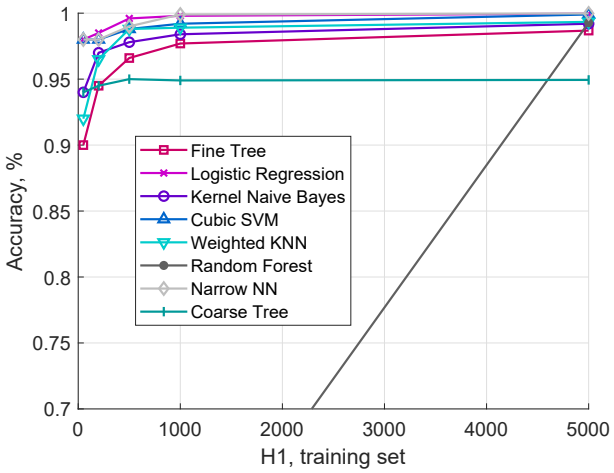


Fig. 3. Group assignment accuracy, σ, for $H_2 = 5000$, $R = 250\,\text{m}$, $K = 10$.

Note that different considered algorithms are characterized by different training complexity that may affect implementation in functioning 5G NR systems [43], [45]. To this aim, Table VI shows the training time in seconds for cell radius $R = 300\,\text{m}$ and training set of length $H_1 = 5000$, wherein each training sample includes input and desired output. Here, we see that the most computationally demanding algorithms are Cubic SVM, Narrow NN, and Kernel Naive Bayes. Recall that the latter one showed excellent performance in terms of matching both UE allocations to multicast subgroups, σ, and resource matching, γ. Observe that although tree algorithms are much simpler, the computational time is also non-negligible. One of the ways to decrease the computational time is to reduce the training set size. However, this may lead to worse performance. We now proceed with evaluating the minimum training set size providing a sufficient level of accuracy.

The accuracy of UE allocations to multicast subgroups and resource matching is shown in Fig. 3 and Table VII. Here, we see that UE allocation to multicast subgroups accuracy,

σ, increases with H_1 as expected. However, starting from approximately $H_1 = 1000$, the accuracy plateaus and does not increase any further. At the same time, note that perfect resource matching with an optimal solution approach is observed for this considered distance even for very small values of H_1.

Summarizing the results of this section, we state that tree-based algorithms, including Fine Tree, Coarse Tree, and Random Forest demonstrate excellent performance in terms of UE allocations to multicast subgroups and resource matching. Furthermore, the accuracy of all the considered algorithms (except for Random Forest) remains virtually unchanged when increasing the training sample size from $H_1 = 1000$ to higher values. This allows considering the latter as the lower bound on the training set size in practical implementations.

B. Extrapolation Capabilities of ML Algorithms

We now proceed with analyzing the extrapolation capabilities of the ML algorithms. To this aim, we train these algorithms by utilizing the training sample of length $H_1 = 1000$ for 10 UEs and then applying the trained algorithms to the system with 13 UEs. The accuracy metrics are calculated for the system with 13 UEs solved by applying the optimal solution.

Fig. 4 shows the accuracy of multicast subgroup formation for $H_1 = H_2 = 5000$ and $K = 13$ UEs. As one may observe, the match is perfect up to approximately $R = 250$ m and then drops abruptly for $R = 275$ m and beyond. The rationale is that the considered metric accounts for specific UEs classified into subgroups. Up to $R = 275$ m only one subgroup is utilized, explaining the perfect match between solutions. We also note that for service area radius higher than $R = 300$ m, UEs are served by using unicast transmissions.

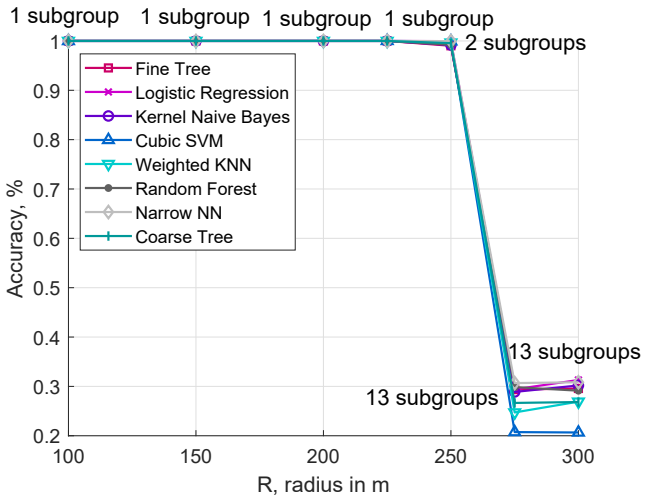


Fig. 4. Group assignment accuracy, σ, for $H_1 = H_2 = 5000$, $K = 13$.

Being incapable of learning specific UEs allocations to individual multicast subgroups shown by σ does not indicate that the considered ML algorithms cannot learn other specifics of UEs classification. To demonstrate it, we provide resource matching accuracy, γ, for various BS service area distances, R, in Table VIII. As one may observe, several algorithms show excellent performance. Specifically, of interest are tree algorithms showing excellent extrapolation capabilities as well. As one may notice, Random Forest and Fine Trees provide almost 100% accuracy in terms of resource utilization, γ, over all the considered distances. Surprisingly, simple logistic regression with minimal computational complexity also shows

TABLE VII
GROUP AND RESOURCE MATCHING ACCURACY FOR H_1, $H_2 = 5000$, $R = 250$ M, $K = 10$.

Sample size H_1	**50**	**200**	**500**	**1000**	**5000**
Fine Tree					
UE assignment, σ	90.00%	94.50%	96.60%	97.70%	98.68%
Resources, γ	100%	100%	100%	100%	100%
Logistic Regression					
UE assignment, σ	98.00%	98.50%	99.60%	99.80%	99.96%
Resources, γ	100%	100%	100%	100%	100%
Kernel Naive Bayes					
UE assignment, σ	94.00%	97.00%	97.80%	98.40%	99.20%
Resources, γ	100%	100%	100%	100%	100%
Cubic SVM					
UE assignment, σ	98.00%	98.00%	98.80%	99.20%	99.90%
Resources, γ	100%	100%	100%	100%	100%
Weighted KNN					
UE assignment, σ	92.00%	96.50%	98.80%	98.90	99.34
Resources, γ	100%	100%	100%	100%	100%
Random Forest					
UE assignment, σ	47.00%	47.00%	47.60%	56.10%	99.28%
Resources, γ	NaN	NaN	NaN	100%	100%
Narrow NN					
UE assignment, σ	98.00%	98.00%	99.00%	99.90%	99.98
Resources, γ	100%	100%	100%	100%	100%
Coarse Tree					
UE assignment, σ	94.00%	94.50%	95.00%	94.90%	94.94
Resources, γ	100%	100%	100%	100%	100%

TABLE VIII
GROUP AND RESOURCE MATCHING ACCURACY, $H_1 = 5000$, $H_2 = 5000$, $K = 13$.

Radius	**100m**	**150-225m**	**250m**	**275m**	**300m**
Fine Tree					
UE assignment, σ	100%	100%	99.02%	29.35%	29.58%
Resources, γ	100%	100%	100%	98.51	96.97%
Logistic Regression					
UE assignment, σ	100%	100%	99.96%	29.41%	31.30%
Resources, γ	100%	100%	100%	100%	98.53%
Kernel Naive Bayes					
UE assignment, σ	100%	100%	99.17%	28.88%	30.19%
Resources, γ	100%	100%	100%	98.44%	95.39%
Cubic SVM**					
UE assignment, σ	99.98%	NaN/100%	99.92%	20.74%	20.66%
Resources, γ	100%	NaN/100%	100%	85.00%	96.88%
Weighted KNN					
UE assignment, σ	100%	100%	99.67%	24.72%	26.91%
Resources, γ	100%	100%	100%	96.92%	98.53%
Random Forest					
UE assignment, σ	100%	100%	99.21%	29.86%	29.13%
Resources, γ	100%	100%	100%	96.92%	100%
Narrow NN					
UE assignment, σ	100%	100%	99.96%	30.67%	30.84%
Resources, γ	100%	100%	100%	98.53%	100%
Coarse Tree					
UE assignment, σ	100%	100%	99.55%	26.65%	26.83%
Resources, γ	100%	100%	100%	59.42%*	90.2%*

*the algorithm defines 5 clusters (on average) instead of 13
** no solution for 150, 200 m, accuracy is 100% is for 225 m

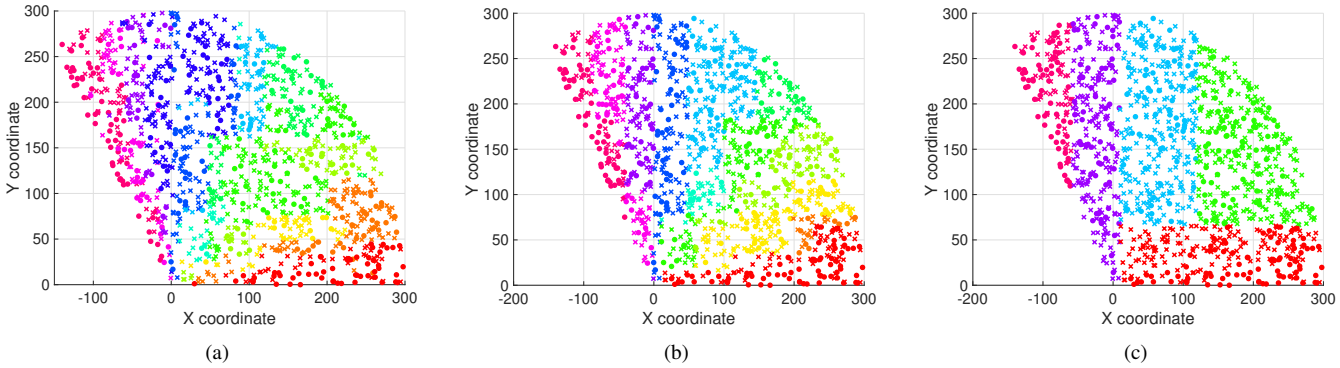


Fig. 5. Tree models predictions for: (a) Fine Tree (13 subgroups), (b) Random Forest (13 subgroups), and (c) Coarse Tree (5 subgroups). Here, crosses denote incorrect predictions, circles correspond to the correct ones, whereas colors represent different subgroups.

excellent performance. By recalling that trees are characterized by rather small computational efforts, we choose them as the best candidates for multicast subgroup formation.

C. Trees' Characteristics and Comparison

Having identified trees as the algorithms capable of providing high accuracy for the considered multicast problem, we now proceed by providing deeper insights on decision tree learners considered in this works, such as random forest, fine, and coarse trees, as summarized in Table IX. First, we recall that according to Table VI, coarse tree outperforms random forest and fine trees in terms of training time. However, one should investigate the other parameters responsible for the accuracy and the quality of predictions, e.g., model flexibility.

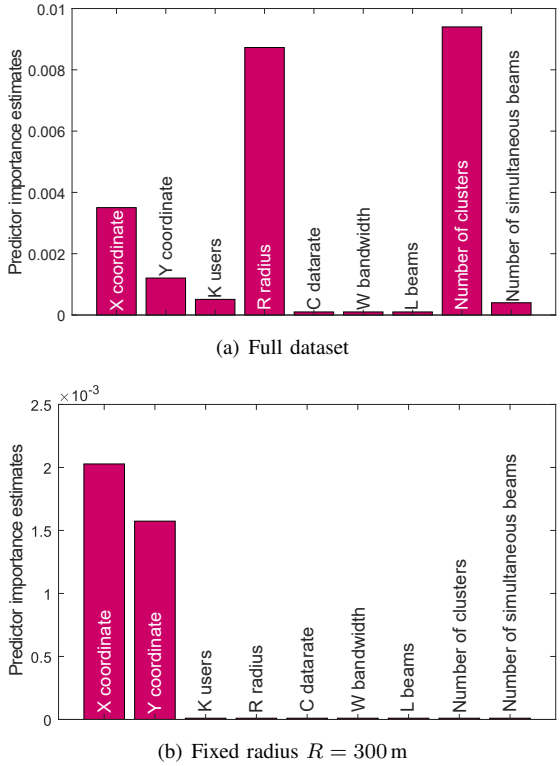
The flexibility of the tree-based model generally increases with the maximum number of splits. Here, a fine tree has high flexibility and operates with many leaves to make many fine distinctions between classes (the maximum number of splits in our case is set to 100). Thus, a fine tree with many leaves is usually highly accurate on the training datasets. Random Forest has medium to high flexibility, which increases with the number of learners or the maximum number of splits. They can usually do better than bagged trees but might require parameter tuning and more learners. Differently, a coarse tree is characterized by low model flexibility. It ensures few leaves only to make coarse distinctions between classes (the maximum number of splits is 4, hence, the maximum number of classes is 5). Hence, the model specifics can explain the unsatisfactory performance of coarse tree, especially in terms of UEs assignment, which we can see in Table VIII, for the area radius of 275 m and 300 m, where according to the exact optimal solution the model should have 13 unicast subgroups, whereas it provides in average only 5 clusters for 13 UEs.

TABLE IX
COMPARISON OF DECISION TREE ALGORITHMS.

Characteristics	**Fine Tree**	**Random Forest**	**Coarse Tree**
Prediction speed, obs/s	∼99000	∼24000	∼330000
Max. no. of splits	100	20 (30 learners)	4
Flexibility	High	Medium-high	Low

In Fig. 5, we show the impact of the model flexibility, or equally, the maximum number of splits in the model for random forest, fine, and coarse trees. Here, x- and y-coordinates are the UE coordinates and are two dataset features. As one can observe, a coarse tree indeed distinguishes fewer classes than the fine tree and random forest. Therefore, in the next section, we analyze the performance of random forest and fine trees only.

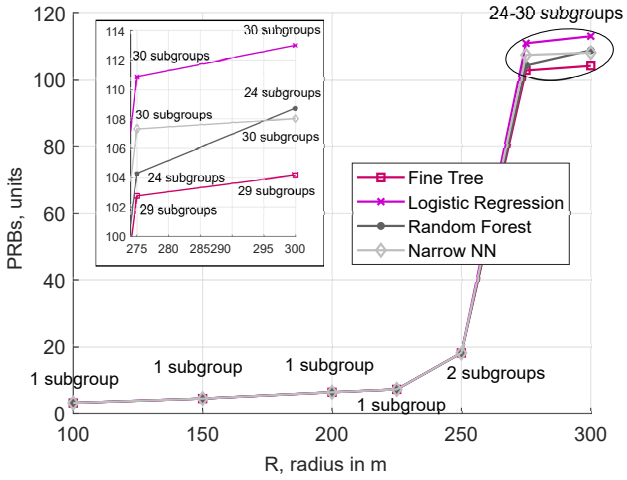
1) Predictors' Importance Estimation: Recall that to construct a dataset, we selected many variables of interest. However, these variables may or may not be utilized by the algorithms for classifications. Note that all algorithms are trained on the same dataset, while each algorithm may select a particular set of input predictor variables to build the final



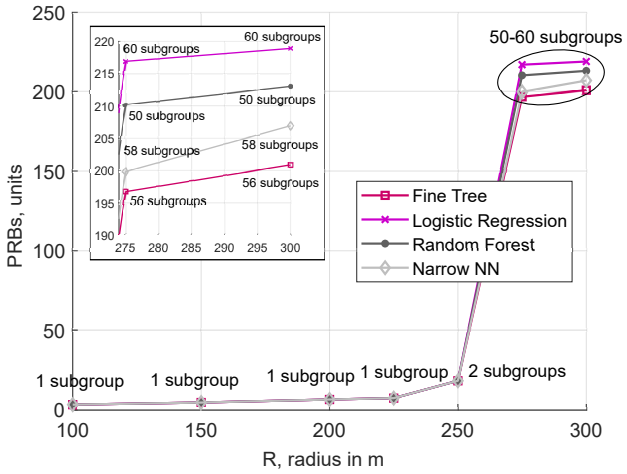
(a) Full dataset

(b) Fixed radius $R = 300\,\text{m}$

Fig. 6. Variables' importance estimates.



(a) 30 UEs



(b) 60 UEs

Fig. 7. Utilized PRBs as a function of service area radius.

classification. In practical work, an intelligent feature selection is essential to boost the predictive power of ML algorithms. For this purpose, we now proceed by exploring the question of which variables mainly affect the performance of the algorithms. To this aim, in Fig. 6, we provide predictor importance for the classification ensemble of decision trees. We use the Matlab function *predictorImportance* that computes estimates of predictor importance for the dataset by summing these estimates over all weak learners in the ensemble. Note that a high value indicates that this predictor is important for the model.

First, Fig. 6(a) illustrates the importance of the complete dataset, where we analyze the model's behavior as a function of the service area radius, R. Hence, it is predictable that R variable has a high importance estimate. However, we expected UEs coordinates to be the most important model's features. In contrast, the number of clusters obtained from the exact optimization problem solution and the cell radius are two key predictors that affect the learning process, followed by UEs coordinates.

Further, by studying Fig. 6(b), one can deduce that the importance of the predictors is dataset-specific. Here, fixing the radius R leads to the UEs coordinates domination. This behavior can be explained by the fact that in directional multicast systems, the radius of the service area impacts the type of transmission utilized for service (i.e., multicast for multiple UEs or unicast for all multicast UEs). Our numerical results confirm that the solution mainly depends on the cell radius. For example, we can observe that a single multicast subgroup is selected for the radius range $100 - 225$ m when we vary the number of UEs in the system. Then, for the range 275 m and further, unicast transmissions are exclusively utilized to serve multicast UEs, whereas the proposed ML solution can be utilized for the radius around 250 m.

D. Performance Assessment of Optimal Multicasting

We finalize our numerical exposition with examples of the ML solution for a realistic number of UEs in the multicast group. Following the 3GPP guidelines for evaluating 5G NR system performance, the number of UEs in the system is 30 or 60 [47]. To this aim, Fig. 7 shows the amount of occupied resources for multicast service with 30 UEs and 60 UEs in the system provided by the best-identified algorithms, including Fine and Bagged Trees (Random Forest), Logistic Regression and Narrow Neural Network.

By analyzing the data presented in Fig. 7 we observe that the trends observed for both numbers of UEs in the system are self-consistent. Specifically, both figures demonstrate the same results in terms of the number of utilized resources up to 250 m of the radius of interest as a single multicast subgroup can be utilized for serving all the UEs. Further, there is a sharp jump when the system changes its operational regime from multicast to unicast. Here, one can notice an almost doubled number of PRBs for the case of 60 UEs compared to 30 UEs, which is typical behavior for the case of unicast transmissions.

VII. Conclusions

In this paper, motivated by the need to support multicast services in 5G mmWave NR systems with directional antennas, we evaluated the suitable ML approaches for optimal multicasting. To this aim, we first developed an exact optimization framework. However, as the optimization test belongs to the class of mixed-integer programming problems and is thus characterized by exponential complexity, we then proceed by evaluating several ML techniques for optimal multicasting.

By applying the discrimination procedure via comparing modeling and exact optimal solution to the considered set of ML approaches, we revealed that tree algorithms show the best performance for the multicast problem. The number of splits of the trees also matters as Fine and Bagged Trees outperform the Coarse Tree, which has a much smaller amount of splits. The factors mainly responsible for the accuracy of ML approximations are the cell service area and UE coordinates, in addition to "external" knowledge of the number of multicast subgroups provided during the training process. We also discovered a narrow range of the cell area radius R where one has to solve multicasting problems in 5G NR systems with directional systems. Specifically, multicasting with one wide beam for small cell radii leads to the optimal solution.

For large cells, unicast transmissions represent the optimal solutions to the multicast problem. There is a narrow range between these two extremes, reported to be $225 - 275\,\mathrm{m}$ for the considered system parameters, where the optimal solution is non-trivial.

ACKNOWLEDGMENT

The authors gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska Curie grant agreement No. 813278 (A-WEAR: A network for dynamic wearable applications with privacy constraints, http://www.a-wear.eu/).

REFERENCES

[1] H. Holma, A. Toskala, and T. Nakamura, *5G Technology: 3GPP New Radio*. John Wiley & Sons, 2020.

[2] J. T. Penttinen, *5G Second Phase Explained: The 3GPP Release 16 Enhancements*. John Wiley & Sons, 2021.

[3] X. Lin, J. Li, R. Baldemair, J.-F. T. Cheng, S. Parkvall, D. C. Larsson, H. Koorapaty, M. Frenne, S. Falahati, A. Grovlen *et al.*, "5G New Radio: Unveiling the Essentials of the Next Generation Wireless Access Technology," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 30–37, 2019.

[4] G. Araniti, A. Iera, S. Pizzi, and F. Rinaldi, "Toward 6G Non-Terrestrial Networks," *IEEE Network*, vol. 36, no. 1, pp. 113–120, 2021.

[5] T.-K. Le, U. Salim, and F. Kaltenberger, "An Overview of Physical Layer Design for Ultra-Reliable Low-Latency Communications in 3GPP Releases 15, 16, and 17," *IEEE access*, vol. 9, pp. 433–444, 2020.

[6] F. Rinaldi, A. Raschella, and S. Pizzi, "5G NR System Design: A Concise Survey of Key Features and Capabilities," *Wireless Networks*, vol. 27, no. 8, pp. 5173–5188, 2021.

[7] I. Farris, A. Orsino, L. Militano, A. Iera, and G. Araniti, "Federated IoT Services Leveraging 5G Technologies at the Edge," *Ad Hoc Networks*, vol. 68, pp. 58–69, 2018.

[8] I. Qualcomm Technologies, "Pioneering 5G Broadcast. Building on Multiple Generations of Cellular Broadcast Technology Leadership," Tech Rep, Tech. Rep., 2021.

[9] S. Pizzi, C. Suraci, A. Iera, A. Molinaro, and G. Araniti, "A Sidelink-aided Approach for Secure Multicast Service Delivery: From Human-Oriented Multimedia Traffic to Machine Type Communications," *IEEE Transactions on Broadcasting*, vol. 67, no. 1, pp. 313–323, 2020.

[10] 3GPP, "New Work Item on NR support of Multicast and Broadcast Services," 3GPP RP-193248, Tech. Rep., December 2019.

[11] N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Efficient Management of Multicast Traffic in Directional mmWave Networks," *IEEE Transactions on Broadcasting*, vol. 67, no. 3, pp. 593–605, 2021.

[12] 3GPP, "Multimedia Broadcast/Multicast Service (MBMS); Stage 1 (Release 16)," 3GPP TS 22.146 V16.0.0, Jul 2020.

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

[14] H. Bagheri, M. Noor-A-Rahim, Z. Liu, H. Lee, D. Pesch, K. Moessner, and P. Xiao, "5G NR-V2X: Toward Connected and Cooperative Autonomous Driving," *IEEE Communications Standards Magazine*, vol. 5, no. 1, pp. 48–54, 2021.

[15] O. Chukhno, N. Chukhno, O. Galinina, Y. Gaidamaka, S. Andreev, and K. Samouylov, "Analysis of 3D Deafness effects in Highly Directional mmWave Communications," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[16] X. Lin and N. Lee, *5G and Beyond: Fundamentals and Standards*. Springer Nature, 2021.

[17] N. Chukhno, O. Chukhno, G. Araniti, A. Iera, A. Molinaro, and S. Pizzi, "Challenges and Performance Evaluation of Multicast Transmission in 60 GHz mmWave," in *International Conference on Distributed Computer and Communication Networks*. Springer, 2020, pp. 3–17.

[18] H. Park, S. Park, T. Song, and S. Pack, "An Incremental Multicast Grouping Scheme for mmWave Networks with Directional Antennas," *IEEE Communications Letters*, vol. 17, no. 3, pp. 616–619, 2013.

[19] S. Sun and T. S. Rappaport, "Multi-Beam Antenna Combining for 28 GHz Cellular Link Improvement in Urban Environments," in *2013 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2013, pp. 3754–3759.

[20] W. Hong, Z. H. Jiang, C. Yu, J. Zhou, P. Chen, Z. Yu, H. Zhang, B. Yang, X. Pang, M. Jiang *et al.*, "Multibeam Antenna Technologies for 5G Wireless Communications," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6231–6249, 2017.

[21] E. Garro, M. Fuentes, J. Carcel, H. Chen, D. Mi, F. Tesema, J. Gimenez, and D. Gomez-Barquero, "5G Mixed Mode: NR Multicast-Broadcast Services," *IEEE Transactions on broadcasting*, vol. 66, no. 2, pp. 390–403, 2020.

[22] M. Saily, C. Barjau, D. Navratil, A. Prasad, D. Gomez-Barquero, and F. B. Tesema, "5G Radio Access Networks: Enabling Efficient Point-to-Multipoint Transmissions," *IEEE Vehicular Technology Magazine*, vol. 14, no. 4, pp. 29–37, 2019.

[23] H. Park and C.-H. Kang, "A Group-aware Multicast Scheme in 60GHz WLANs," *TIIS*, vol. 5, no. 5, pp. 1028–1048, 2011.

[24] A. Biason and M. Zorzi, "Multicast Transmissions in Directional mmWave Communications," in *European Wireless 2017; 23th European Wireless Conference*. VDE, 2017, pp. 1–7.

[25] ——, "Multicast via Point to Multipoint Transmissions in Directional 5G mmWave Communications," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 88–94, 2019.

[26] K. Sundaresan, K. Ramachandran, and S. Rangarajan, "Optimal Beam Scheduling for Multicasting in Wireless Networks," in *Proceedings of the 15th annual international conference on Mobile computing and networking*, 2009, pp. 205–216.

[27] H. Zhang, Y. Jiang, K. Sundaresan, S. Rangarajan, and B. Zhao, "Wireless Multicast Scheduling with Switched Beamforming Antennas," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1595–1607, 2012.

[28] E. Aryafar, M. A. Khojastepour, K. Sundaresan, S. Rangarajan, and E. Knightly, "ADAM: An Adaptive Beamforming System for Multicasting in Wireless LANs," *IEEE/ACM Transactions on Networking*, vol. 21, no. 5, pp. 1595–1608, 2013.

[29] N. Chukhno, O. Chukhno, D. Moltchanov, A. Molinaro, Y. Gaidamaka, K. Samouylov, Y. Koucheryavy, and G. Araniti, "Optimal Multicasting in Millimeter Wave 5G NR with Multi-beam Directional Antennas," *IEEE Transactions on Mobile Computing (Early Access)*, 2021.

[30] N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Unsupervised Learning for D2D-Assisted Multicast Scheduling in mmWave Networks," in *2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2021, pp. 1–6.

[31] L. Feng, Z. Yang, Y. Yang, X. Que, and K. Zhang, "Smart Mode Selection Using Online Reinforcement Learning for VR Broadband Broadcasting in D2D Assisted 5G HetNets," *IEEE Transactions on Broadcasting*, vol. 66, no. 2, pp. 600–611, 2020.

[32] Y. Wang, M. Narasimha, and R. W. Heath, "MmWave Beam Prediction with Situational Awareness: A Machine Learning Approach," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[33] Y. Heng and J. G. Andrews, "Machine Learning-Assisted Beam Alignment for mmWave Systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 4, pp. 1142–1155, 2021.

[34] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. " O'Reilly Media, Inc.", 2017.

[35] M. Merenda, C. Porcaro, and D. Iero, "Edge Machine Learning for AI-enabled IoT devices: A Review," *Sensors*, vol. 20, no. 9, p. 2533, 2020.

[36] 3GPP, "NR; Physical Channels and Modulation (Release 15)," 3GPP TR 38.211, Dec 2017.

[37] R. Kovalchukov, D. Moltchanov, Y. Gaidamaka, and E. Bobrikova, "An Accurate Approximation of Resource Request Distributions in Millimeter Wave 3GPP New Radio systems," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Springer, 2019, pp. 572–585.

[38] 3GPP, "Multimedia Broadcast/Multicast Service (MBMS); Stage 1," 3GPP TR 22.146, March 2022.

[39] ——, "Technical Specification Group Radio Access Network; Study on Channel Model for Frequency Spectrum above 6 GHz (Release 14)," 3GPP TR 38.900 V14.2.0, Tech. Rep., December 2016.

[40] G. R. MacCartney, T. S. Rappaport, and S. Rangan, "Rapid Fading Due to Human Blockage in Pedestrian Crowds at 5G Millimeter-Wave Frequencies," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–7.

[41] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, E. Aryafar, S.-p. Yeh, N. Himayat, S. Andreev, and Y. Koucheryavy, "Analysis of Human-Body Blockage in Urban Millimeter-Wave

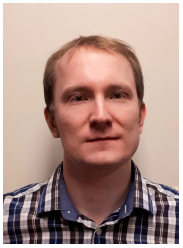
Cellular Communications," in *2016 IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–7.
[42] A. B. Constantine, *Antenna Theory: Analysis and Design*. Wiley-Interscience, 2005.
[43] ITU-T Rec. Y.3172, "Architectural Framework for Machine Learning in Future Networks Including IMT-2020," 2020.
[44] ITU-T Rec. Y.3174, "Framework for Data Handling to Enable Machine Learning in Future Networks Including IMT-2020," 2020.
[45] ITU-T Rec. Y.3176, "Machine Learning Marketplace Integration in Future Networks Including IMT-2020," 2020.
[46] A. D. Kulkarni and B. Lowe, "Random Forest Algorithm for Land Cover Classification," 2016.
[47] ITU-R, "Guidelines for Evaluation of Radio Interface Technologies for IMT-2020," M.2412-0, July 2017.

Nadezhda Chukhno is an Early Stage Researcher at A-WEAR and Doctoral Researcher at Mediterranea University of Reggio Calabria, Italy and Jaume I University, Spain. She graduated from RUDN University, Russia, and received her B.Sc. in Business Informatics (2017) and M.Sc. in Fundamental Informatics and Information technologies (2019). Her current research activity mainly focuses on wireless communications, 5G+ networks, multicasting, D2D, and wearable technologies.

Olga Chukhno is an Early Stage Researcher within H2020 MCSA ITN/EJD A-WEAR project and a PhD student at Mediterranea University of Reggio Calabria, Italy and Tampere University, Finland. She received M.Sc. (2019) in Fundamental Informatics and Information Technologies and B.Sc. (2017) in Business Informatics from RUDN University, Russia. Her current research interests include wireless communications, social networking, edge computing, and wearable applications.

Dmitri Moltchanov received the M.Sc. and Cand.Sc. degrees from the St. Petersburg State University of Telecommunications, Russia, in 2000 and 2003, respectively, and the Ph.D. degree from the Tampere University of Technology in 2006. Currently he is University Lecturer in with the Laboratory of Electronics and Communications Engineering, Tampere University, Finland. He has (co-)authored over 150 publications on wireless communications, heterogeneous networking, IoT applications, applied queuing theory. In his career he has taught more than 50 full courses on wireless and wired networking technologies, P2P/IoT systems, network modeling, queuing theory, etc. His current research interests include research and development of 5G/5G+ systems, ultra-reliable low-latency service, industrial IoT applications, mission-critical V2V/V2X systems and blockchain technologies.

Anna Gaydamaka received the B.Sc. (Hons) degree in Business Informatics from the Peoples' Friendship University of Russia (RUDN University) in 2018 and the M.Sc. degree in Computer Science (program Data Science and Business Informatics) from Universita di Pisa, Italy in 2021. Currently, she is pursuing a Ph.D. in Computing and Electrical Engineering and working as a Doctoral Researcher at Tampere University, Finland. Her research interests include resource planning of the fifth and sixth-generation wireless network, mathematical models for performance KPIs optimization and machine learning techniques.



Andrey Samuylov received the Ms.C. degree in applied mathematics and the Cand.Sc. degree in physics and mathematics from RUDN University, Russia, in 2012 and 2015. Currently he is pursuing a Ph.D. degree with the Unit of Electrical Engineering, Tampere University, Finland. His research interests include P2P networks performance analysis, performance evaluation of wireless networks with enabled D2D communications, and mmWave-band communications.

Antonella Molinaro graduated in Computer Engineering (1991) at the University of Calabria, received a Master degree in Information Technology from CEFRIEL/Polytechnic of Milano (1992), and a Ph.D. degree in Multimedia Technologies and Communications Systems (1996). She is currently an associate professor of telecommunications at the University Mediterranea of Reggio Calabria, Italy. Her research activity mainly focuses on wireless and mobile networking, vehicular networks, and future Internet.

Yevgeni Koucheryavy received the Ph.D. degree from the Tampere University of Technology (TUT), Finland. He is currently a Professor at the Laboratory of Electronics and Communications Engineering, TUT. He is the author of numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects in heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, and nanocommunications. He is Associate Technical Editor of the IEEE Communications Magazine and an Editor of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS.

Antonio Iera graduated in computer engineering from the University of Calabria in 1991, and received a Master's degree in IT from CEFRIEL/Politecnico di Milano in 1992 and a Ph.D. degree from the University of Calabria in 1996. From 1997 to 2019 he has been with the University Mediterranea, Italy, and currently holds the position of full professor of Telecommunications at the University of Calabria, Italy. His research interests include next generation mobile and wireless systems, and the Internet of Things.

Giuseppe Araniti (Senior Member, IEEE) received the Laurea degree and the Ph.D. degree in electronic engineering from the University Mediterranea of Reggio Calabria, Italy, in 2000 and 2004, respectively. He is currently an Assistant Professor of telecommunications with the University Mediterranea of Reggio Calabria. His major area of research is on 5G/6G networks and it includes personal communications, enhanced wireless and satellite systems, traffic and radio resource management, multicast and broadcast services, device-to-device (D2D), and machine-type communications (M2M/MTC).