electronics

Perspective

Modelling of ML-Enablers in 5G Radio Access Network-Conceptual Proposal of Computational Framework

Malgorzata Tomala * and Kamil Staniec

Telecommunications and Teleinformatics Department, Wroclaw University of Science and Technology, 50-370 Wroclaw, Poland
* Correspondence: m.tomala@pwr.edu.pl

Abstract: The fifth generation (5G) of mobile networks connects people, things, data, applications, transport systems, and cities in smart networked communication environments. With the growth in the amount of generated data, the number of wirelessly connected machines, traffic types, and associated requirements, ensuring high-quality mobile connectivity becomes incredibly difficult for technology suppliers. Mobile operators and network vendors enrolling in 5G face far more rapid demands than any technology before, and at the same time need to ensure efficiency and reliability in the network operations. In fact, intelligent forecasting and decision-making strategies are several of the centerpieces of current artificial intelligence research in various domains. Due to its strong fitting ability, machine learning is seen to have great potential to be employed to solve telecommunication networks' optimization problems that range from the design of hardware elements to network self-optimization. This paper addresses the question of how to apply artificial intelligence to 5G radio access control and feed ML techniques with radio characteristic-based automatic data collection to achieve ML-based evaluation of 5G performance. The proposed methodology endorses ML tools for the 5G portfolio scenarios requirements assessment and integrates into the mature methods for network performance optimization: self-organizing networks (SON) and minimization of drive tests (MDT). In this context, the proposed treatment guides future network deployments and implementations adopted on a 3GPP standard basis.

Keywords: 5G; *eMBB*; *URLLC*; *mMTC*; MDT; SON; AI; ANN; CNN; GNN; ML

Citation: Tomala, M.; Staniec, K. Modelling of ML-Enablers in 5G Radio Access Network-Conceptual Proposal of Computational Framework. *Electronics* **2023**, *12*, 481. https://doi.org/10.3390/electronics12030481

Academic Editors: Toan-Van Nguyen, Thien Huynh-The and Muhammad Atif Ur Rehman

Received: 29 November 2022
Revised: 3 January 2023
Accepted: 10 January 2023
Published: 17 January 2023

1. Introduction

1.1. Background

The earliest technological solutions in mobile networks aimed at enabling wireless connectivity transfer for cellular data voice mainly, whereas the 5G telecommunication network is hailed as the digitalization and connection of everything and everyone, with the goal of automating much of life [1,2]. Blind spots remain, however, 95 per cent of the world's population is now living within range of a mobile broadband network, more than 90 per cent of the population owns a mobile phone, and an estimated 4.9 billion people are using the Internet [1]. The data volumes in mobile networks have greatly exceeded the voice volumes, hence the subsequent generations of mobile network systems are being designed to further boost the data rates and network capacities.

The International Telecommunication Union (ITU), working towards providing successful telecommunication network deployments at the regional and international levels, established a high-level framework and overall objectives for 5G development [3–5]. To help meet the ever-increasing demands for wireless communication, the ITU envisioned three main use cases for 5G networks: enhanced mobile broadband (*eMBB*), ultra-reliable and low-latency communications (*URLLC*), and massive machine-type communications (*mMTC*). The main drivers behind this classification address the undergoing trends in mobile networks. Aside from applications generating and requiring a very high data rate, there are expectations of instantaneous connectivity without waiting times (i.e., a single click and an immediate response) and handling a vast number of connected devices simultaneously. Though some need the transmission of high-definition data while on the move, others create stationary congestion. The associated requirements of the supporting 5G system lead to a broad variety of capabilities (Figure 1). For example, *eMBB* addresses the human-centric use cases for access to multimedia content, services requiring high data rates with improved performance, and an increasingly seamless user experience. The services include video streaming, gaming, or a broad variety of mobile data generated at smart home buildings. *URLLC* examples include wireless support for industrial manufacturing or production processes, remote medical surgery, transportation safety, etc., aiming toward stringent requirements for capabilities such as throughput, latency, and availability. In general, the profile aids to support industrial automation, mission-critical applications, or self-driving cars. *mMTC* is characterized by a very large number of connected devices typically transmitting a relatively low volume of nondelay-sensitive data. This profile covers smart city applications, as well as mission-critical connections.

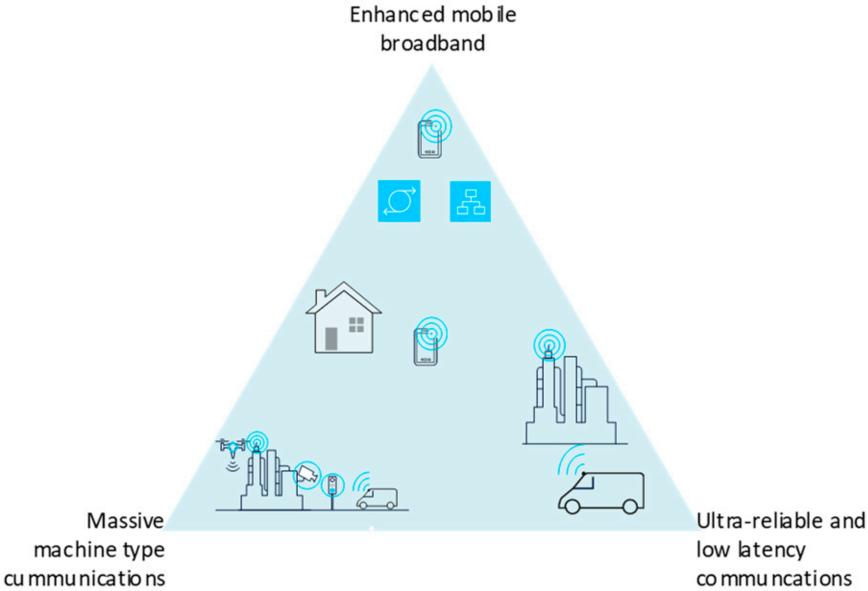

Figure 1. Usage scenarios for 5G and beyond [3].

Network vendors, supplying telecom operators offering their subscribers 5G with ambitious goals, need to navigate the standardized instruction, according to ITU requirements. The 3rd Generation Partnership Project (3GPP), the standard defining 5G technical specifications for telecommunication networks, includes overall 5G network infrastructure architecture [6] and the 5G New Radio (5G NR) concept [7]. In addition to the key architectural aspects, protocols, and interface designs, the standard defines methods for helping to facilitate 5G deployment, inheriting various optimization methods practiced already in the previous mobile networks' generations. The advanced SON (self-organizing network) technology is one example of a solution to enable operators to improve network efficiency. The functionality facilities automated network operations through self-commissioning (RAN node plug and play), self-healing, and self-configuration referring to network autonomous configuration and monitoring, its parameters tuning, and appropriate steering of the network traffic and load, leading to better overall network performance [8]. MDT (minimization of drive tests) is a method that utilizes users' equipment (UE) to collect radio measurements and associated location information, in order to assess network performance whereas reducing the operational expenditure (OPEX) associated with traditional (manual) drive tests. It enables operators to collect feedback from regular users' devices in the field (including indoor users) without the need to run offline and trial data collection sessions to observe network performance in a specific area or initiate customer care dedicated to a specific user [9]. SON targets the network's operational efficiency. MDT specifically aims at improving the end-user experience. The two features are complementary. Both address major network optimization use cases. Once applied together, they have synergized aims to achieve better network performance and satisfied subscribers.

Automated approaches exist, however, 5G merging and maintenance appear much more complex due to the 5G traffic variation characteristic of different users, flexible network deployments, and varying implementations and capabilities. For instance, 5G enables using multiple input–multiple output (MIMO) antennas, many base stations including macro sites operating in cooperation with smaller stations, employing a variety of radio technologies depending on service needs, use cases, and/or the spectrum available. Yet, 5G networks may support network slicing. Primarily, network slices may correspond to the 5G usage scenarios, i.e., *eMBB*, *URLLC*, and *mMTC*, and enable an end-to-end solution, in which multiple independent and dedicated virtual subnetworks (network instances) may be created within the substantially same infrastructure to run services that have different requirements on latency, reliability, throughput, and mobility.

Overall, the coexistence of an extremely large number of connection options and the need to process very high volumes of data become a bottleneck for the effectiveness and efficiency of known autonomous solutions for network optimization. Although AI (artificial intelligence) and ML(machine learning) methods demonstrate a promising ability to resolve difficult classification problems in big data collection sets, difficulty in integrating AI/ML with the existing tools and infrastructure, and the compatibility of AI/ML across different devices, are exemplary challenges.

For that reason, 3GPP has conducted studies on AI/ML usage in collecting data in mobile networks in [10]. The standardization support for AI/ML is focused on building the functional architecture of using AI/ML (i.e., data acquisition and exposure), and the high-level requirements of operating an ML-based RAN (i.e., the ML algorithms are assumed to be up to implementation). The studies recommend a framework, which guides further development of technical solutions and practical implementations.

1.2. Aim and Scope

Within this scope, this paper is a contribution proposing a method helping to:

- Identify ML-algorithms input data pertaining to different 5G traffic types: *eMBB*, *URLLC*, and *mMTC*;
- Develop an ML-based data collection model that classifies the importance of the exploited 5G traffic characteristics in dependence on the RAN optimization use cases;
- Apply different strategies to radio interface conditions evaluation;
- Design prediction tools for network optimization.

The presented methodology determines the data training and execution principles involved in AI/ML schemes. The objective is to achieve an ML-based solution to be explicitly associated with the traffic characteristics and the intended gain.

This paper is organized in the following manner: Section 2 provides a generic description of ML approaches, the basics of the existing network optimization automated tools, and an AI/ML-based RAN architecture. Section 3 presents an analysis of the key characteristics of 5G traffic for network optimization use cases and how they qualify for ML algorithms. Then analyses of data sets determining ML data training and the execution principles in ML algorithms adopted into 5G RAN network deployments are provided. Section 4 discusses the proposed method's utility and value. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Machine Learning Approaches

Operations based on AI have received tremendous attention in research and technology industry. ML, that allows the applications to function in an "intelligent" manner, is a core tool of AI and plays a vital role in discovering insights from the data. Various types of ML algorithms such as supervised, unsupervised, semisupervised, and reinforcement learning exist in the area [11]. Categories of data being processed in ML also vary from unstructured, semistructured, well-structured, and metadata. Depending on the principles and targets of various ML algorithms, their applicability in various domains differs also (Figure 2). Unsupervised learning analyses unlabeled datasets which can be supplied with various input data and is used for identifying meaningful trends and structures. It requires a large amount of data to find hidden patterns from the given data, group the input data together, and create arbitrary categories. Alternatively, the supervised learning approach needs a small amount of data to estimate the necessary parameters and quickly achieve the expected result (e.g., cluster or classify data). This approach is typically used to map an input to an output based on showing the machine the connection between different variables and known outcomes. It uses labeled data collection and predicts an outcome based on a 'guide' supervising the learning process. It is preferable for solving basic or straightforward problems. Conversely, the reinforcement learning method automatically evaluates the optimal behaviors in a particular context or environment by insight into the related data based on the software-enabled agents that apply a reward/penalty rule based on the observation. The goal of this type of learning is to use insights obtained from the environment to take an action, e.g., to increase the reward or minimize the risk or failure. They are specifically considered a powerful tool for training AI models that facilitate automation or optimize the operational efficiency of sophisticated systems, and for solving numerous real-world problems in various fields [12].

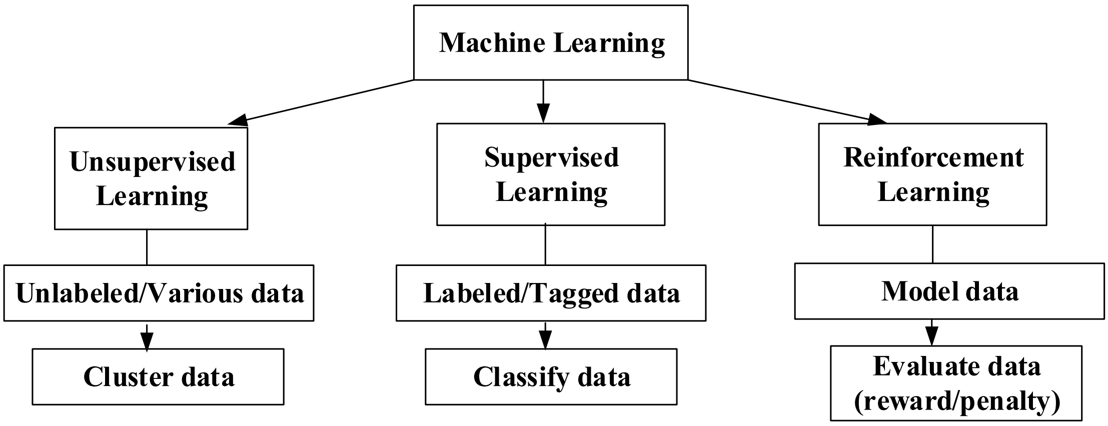

Figure 2. Different machine learning approaches.

The classifications algorithms often instantiate hybrid methodologies. Nevertheless, the right approach selection, including the appropriate identifying and grouping of related data, is a key to build effective ML-based solution. In this way, the outcome of an ML method can be more accurate and meaningful.

2.2. Architecture of Artificial Neural Networks

Artificial neural networks (ANN) are advanced methods that utilize ML-based approaches (and their combinations). They denote statistical models that emulate the function of a network of human neurons.

The models are based on a collection of connected units or nodes called artificial neurons. Combining the biological principle with advanced statistical methods, neural network models are viewed as functions solving problems in domains, such as pattern recognition.

ANN is characterized by a parallel processing ability. It is capable of performing more than one task at the same time. Deep learning, which is a part of ANN-based machine learning approaches, provides a computational architecture by combining several processing layers: input, hidden, and output layers with connected neurons, to learn from data (Figure 3). The primary advantage of deep learning over basic machine learning methods is its better performance in learning from large datasets [12].

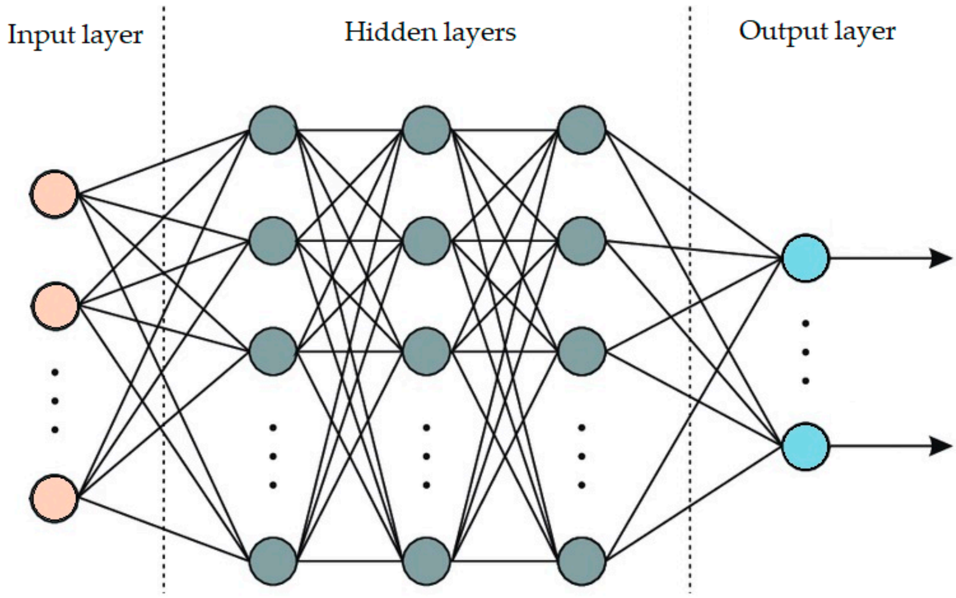

Figure 3. Model of Artificial Neural Network.

In the more detailed subject of deep learning, there are several variants of ANNs. Convolution neural networks (CNNs) are multilayered ANNs enabling convolution operation and are responsible for detecting the most important features. Another class of ANN is a graph neural network (GNN) that enables the processing of data into a graph representation. These approaches can implement different flavors of visualized representation of the observed data.

2.3. AI/ML Adoption in RAN-Architecture

The usefulness and applicability of ML-based methodologies in future deployments of mobile networks have been recognized by mobile network operators (MNO) and the industry overall. An architecture that organizes intelligence in RAN, studied by 3GPP in [10] has been developed according to Figure 4. The study has been concluded with a recommendation to define technical specifications according to the baseline, which will become 3GPP Release 18 [10].

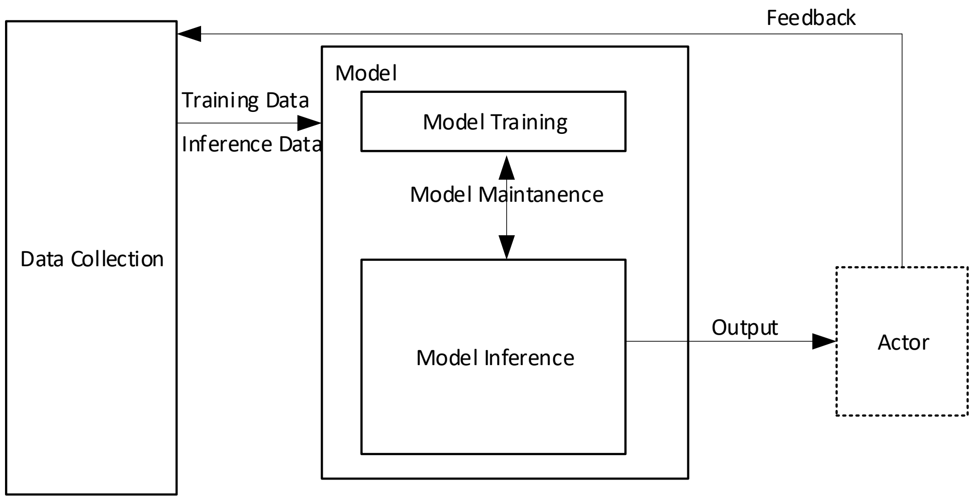

Figure 4. Functional framework for RAN intelligence [10].

The architecture for RAN implementing ML groups RAN entities into different functions:

- Data collection that provides input data to model training and model inference functions. The data collection plays the role of ML algorithm-specific data preparation;
- Model training that performs the AI/ML model training, validation, and testing which may generate model performance metrics as part of the model testing procedure. The Model Training function is also responsible for data preparation;
- Model inference that provides AI/ML model inference output (e.g., predictions or decisions);
- An actor that receives the output from the model inference function and triggers or performs corresponding actions. The actor may trigger actions directed to other network entities or to itself.

The functional framework realization in RAN needs to be compatible with radio interface procedures, as depicted in Figure 5. The data collection function can be split into the UE and RAN nodes. If UE measurements are needed by a gNB for a certain task, they are triggered by a RAN command requesting the UE to perform measurements (e.g., radio signal level and experienced delay). In addition, optionally, the measurements can be also performed by the gNB itself (marked as data collection* in Figure 5). Based on a joint input (the UE indicated measurements and the gNB generated data), the gNB conducts ML model training. The combined data training in a task-driven operation in RAN requires continued observation for a particular period, which typically is followed by an updated instruction on data collection. The typical strategy of such a learning process from data collection is to achieve a predictive analysis and an intelligent decision. This is realized by the model inference and leads to an action. These high-level principles guide a solution to be worked out as a normative standardized method in 3GPP Release 18.

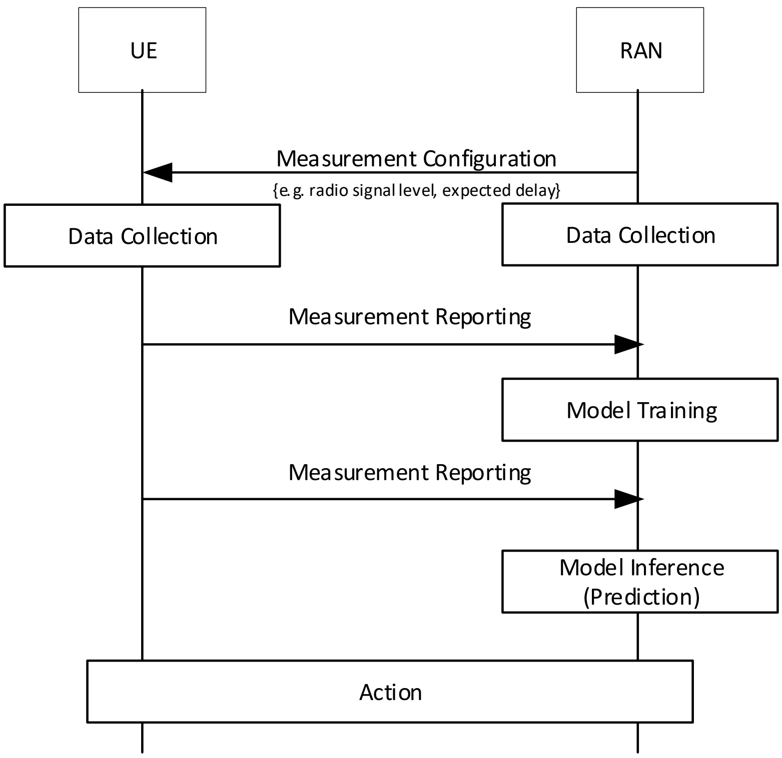

Figure 5. RAN signaling for ML-based operations.

2.4. RAN Optimization Use Cases

Mobile networks currently support various optimization methods. Self-organizing network (SON) and minimization of drive test (MDT) were initiated over a decade ago by the next generation mobile networks (NGMN) alliance and later standardized by the 3GPP. These are continuously evolving features on the enhancements' roadmap for 5G.

The mechanisms aim at automating the huge efforts the operators spent on collecting performance measurements, adjusting network configuration parameters, network planning, and monitoring tools configurations, including multivendor scenarios. Depending on a scenario, SON ensures autonomous adaptation of the network by an optimization algorithm tailored to a specific need, detailed solution, and requirement (e.g., for interference control, the network may autonomously change the settings of frequency/power restrictions and preferences for the resource usage in the different cells based on interference level). MDT specifically facilitates the automated collection of performance measurements, especially from the users. Both functions support the following fundamental use cases:

- Coverage optimization, which consists of detecting and optimizing poor cell coverage and over coverage collectively, coverage maps planning, cell coverage, and capacity monitoring according to the operator-specific deployment strategy and requirements on traffic models and coverage holes detection;
- Mobility Optimization, which consists of mobility events observation (handovers), including the detection of mobility events resulting in errors, causing radio link failure (RLF) due to unsuccessful handovers, detection, and correction of errors in the mobility configuration, etc.;
- *QoS* Optimization, which consists of the user's quality of experience monitoring, verification of whether the quality of service experienced by the end user is in line with the performance target defined in the planning strategy, data rate, and cell throughput observations, etc.

These optimization processes reveal synergy with inferring optimal behavior from ML-based operations. For that reason, it is envisioned that the computational abilities of appropriate ML techniques can empower SON and MDT with intelligence and shape new applicability of the SON and MDT features in continuously developed 5G networks.

3. Applying AI/ML into 5G RAN: Proposed Methodology

3.1. Strategy

The proposed strategy is to model an ML-based technique for 5G RAN performance evaluation that interacts with the radio interface environment. It allows an actor (e.g., RAN) to learn by trial-and-error observations. The observations are performed based on the input data received from the end user through the interactive information exchange. Following the ML paradigm, it is proposed to scale the data collection to further look into the 5G services characteristics and construct an ML data selection model per associated 5G traffic requirement.

The proposed ML scheme incorporates the following steps (also depicted in Figure 6):

1. Input Data observation by the Actor (e.g., RAN node) derived in a parameterized form, where:
    a. The parametrization request classifies multivariate data sets into two different data types. One set includes metrics that have a high influence on exploiting 5G traffic (5G slice) and the second set includes metrics that determine RAN performance;
    b. An update to the input data structure can be identified based on the data filtering and/or data processing actions (e.g., through the model update process).
2. Decision-making function triggering the actor to perform the evaluation of the outcome based on the multiple criteria, where:
    a. The criteria associate raw 5G traffic characteristics with intended RAN optimization uses cases;
    b. The evaluation is based on a decision tree, with ANN-based algorithms adoption for parallel processing of the use cases.

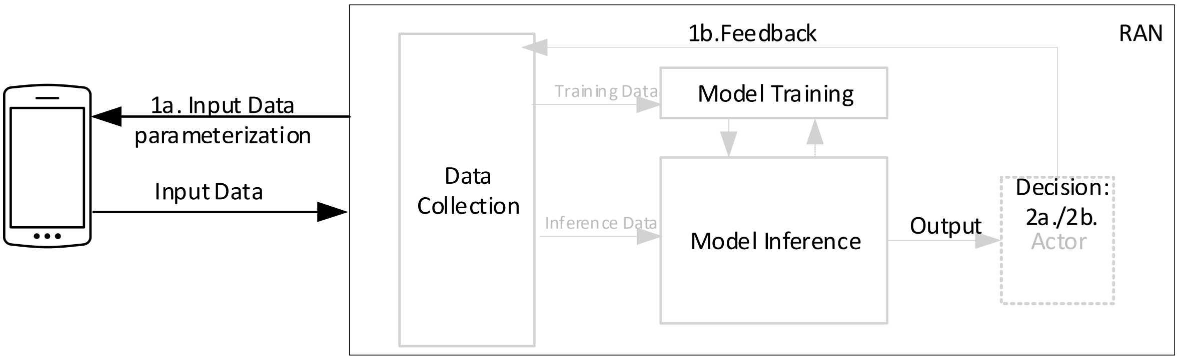

Figure 6. Applying AI/ML-based 5G RAN performance evaluation based on a 3GPP-assumed paradigm.

3.2. Input Data—Classification Analysis

3.2.1. Key Performance Indicators for 5G Traffic Types

The 5G-connected devices are bound to have varying levels of capabilities importance, such as data rate, connection density, or latency (Figure 7). The key minimum technical performance requirements with corresponding metrics definitions for the 5G radio interface are defined in [13]. In addition to these metrics, the report in [14] maps the three 5G usage scenarios (*eMBB*, *mMTC*, and *URLLC*) into evaluation criteria. With these documents, the radio interface evaluation procedure is designed in such a way that the overall performance of the radio interface can be fairly and equally assessed on a technical basis, ensuring that the overall 5G objectives are met.

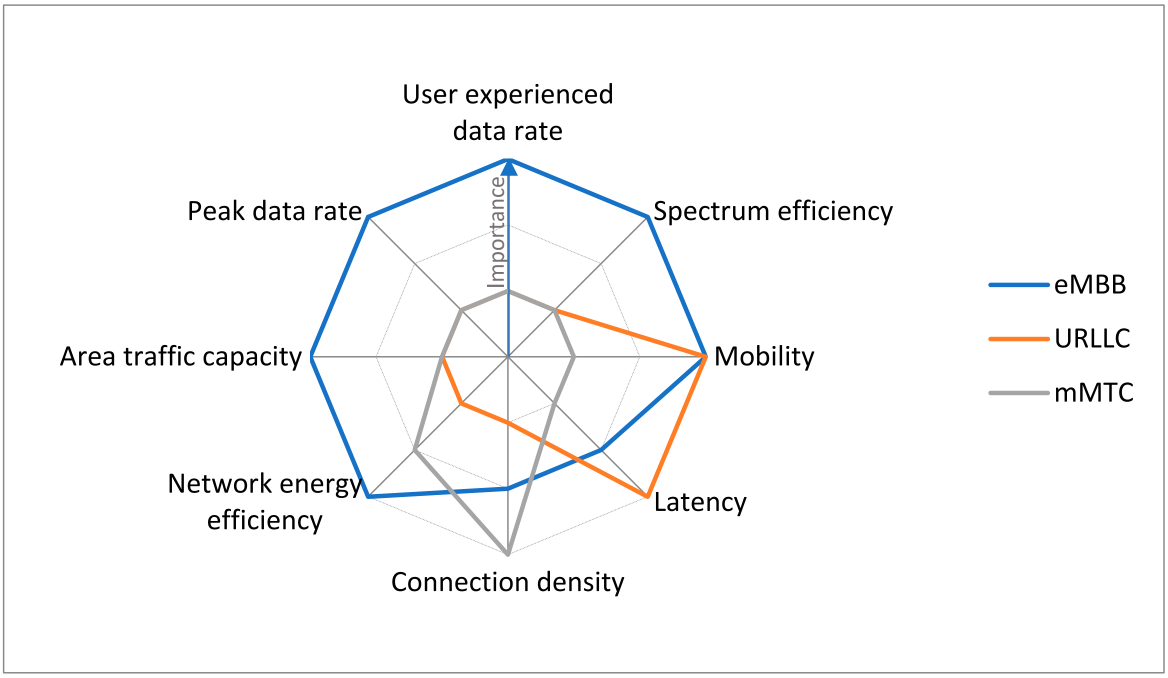

Figure 7. The importance of key capabilities in different 5G usage scenarios [3].

There are many drivers behind the anticipated importance of the 5G characteristics, depending on the detailed circumstances. The enhanced mobile broadband usage scenario comes with new application areas and requirements for improved performance and an increasingly seamless user experience. This usage scenario includes wide-area coverage and hotspot, which have different requirements. For the hotspot case, i.e., for an area with high user density, very high traffic capacity is needed, whereas the requirement for mobility is low and the user data rate is higher than that of wide area coverage. For the wide area coverage case, seamless coverage but medium to high mobility are desired, with a much-improved user data rate. However, the data rate requirement may be relaxed compared to a hotspot. Overall, the variety of smart devices (e.g., smartphones, tablets, televisions, etc.) and a wide range of applications requiring a large amount of data traffic, accelerate demand for wireless data traffic. Since the *eMBB* is expected to accommodate the increasing traffic demand and to provide gigabit-per-second user data rate services, a data rate becomes a pillar measure for the capability assessment [3].

The support for ultralow latency and high-reliability connections corresponds to human- and machine-centric communication. A human's expectation of instantaneous connectivity wherein applications need to exhibit "flash" behaviour without waiting times pertains to communication which becomes an enabler for the development of applications, e.g., in health, safety, office, entertainment, and other sectors. For applications based on machine-centric communication with real-time constraints (e.g., driverless cars, enhanced mobile cloud services, real-time traffic control optimization, emergency, and disaster response, emergency rescue services, control of industrial manufacturing or production processes, remote medical surgery, e-health, etc.), reliability understood as the capability of transmitting a given amount of traffic within a predetermined time duration with high success probability is a key demand [10].

The massive machine type communication corresponds to the presence of a large number of concurrent users, for example in a crowd with a high traffic density per unit area and a large number of handsets and machines/devices per unit area. Hence, primarily the use case envisions a certain number of devices per unit area fulfilling a specific quality of service [10], which should be measurable by the network as a certain level of connection density.

These fundamental 5G characteristics can be measured according to simulation, analytical, or inspection-based 5G performance evaluation methods [12].

To select the 5G traffic-specific indicators for ML-driven data collection, the key performance indicators defined specifically for analytical evaluation methodology [12] apply (Table 1).

Table 1. Key performance indicators of 5G per traffic type.

eMBB	*URLLC*	*mMTC*
Peak data rate	Latency	Connection density
User experience data rata	Mobility interruption time	

As an example, the network function parametrizing radio data collection for an ML algorithm may admit receiving measurement data from a specific service (Table 1), referring also to a network slice (e.g., an *eMBB* slice, a *URLLC* slice, or an *mMTC* slice). For the purpose of training a service-specific ML model, the input data function may target data set A to be denoted as:

$$A = \{eMBB, URLLC, mMTC\}, \tag{1}$$

where *eMBB*, *URLLC*, and *mMTC* denote key performance measurements taken per traffic type.

The network function parametrizing radio data collection for an ML algorithm for the particular service (e.g., *eMBB*), may be targeting measurement results taken only for the *eMBB* service, which is a subset of dataset A:

$$a_{MBB} = \{PeakDataRate, UserExpDataRate\}, \tag{2}$$

Consequently, the network function parametrizing radio data collection for an ML algorithm for combined traffic (e.g., *eMBB* and *URLLC*) may be targeting measurement results taken for both services, that is a dataset denoted as:

$$a_{MBB,URLLC} = \{PeakDataRate, UserExpDataRate\} \cup \{Latency, MobilityInterTime\} \quad (3)$$

As a result, any metric or metrics combinations collected from dataset *A* will represent key performance indicators of 5G traffic or a slice.

3.2.2. Key Performance Indicators for RAN Optimization Use Cases

For RAN performance optimization, three use cases are addressed: coverage optimization, mobility optimization, and *QoS* optimization.

Radio coverage is essential information for network planning, network optimization, and backend network management activities, such as network dimensioning, CAPEX/OPEX planning, and marketing. A detection of coverage problems (e.g., coverage holes, pilot pollution, low user throughput, etc.) in specific areas, e.g., based on customer complaints and along roads or train lines, is also crucial for network operators to be aware of the coverage shortages in the networks they provide. Coverage is easily noticed by a customer through the terminal uplink signal (e.g., strong radio signal or out-of-service area indication). Downlink signal perception is also an important criterion by which many customers judge the performance of the network, as poor coverage will impact the user experience in terms of call setup failure, call drop, or voice quality.

Mobility optimization is an important part of network operation, as information about mobility problems or failures can also be used to identify localized lack of coverage or the need to adapt the network parameters setting, e.g., in order to avoid too early or too late handover and to improve the handover success rate and overall network performance.

Verification of the quality of service (*QoS*) is another key objective of the network performance analysis, as it allows the detection of critical conditions and determining the need to change the network configuration, parameter settings, or capacity extension. Especially, in order to check if the quality of service experienced by the end user is in line with the performance target defined in the planning strategy and offered to the end user. The *QoS* verification allows the assessment of the user experience, e.g., in terms of the user throughput. Hence, it has become an important key performance indicator (KPI) in network deployments for operators.

Accordingly, the SON and MDT functionalities gather autonomous data collection from customers for network coverage, mobility, and *QoS* optimization [9].

Technically, the coverage level collected in MDT is expressed by recorded measurement quantities, such as received signal reference power (*RSRP*), received signal reference quality (*RSRQ*), or signal-to-interference and noise ratio (*SINR*), including the measurements' values recorded in the event of radio link failure (RLF). The closer to the failure, the lower the value of the radio signal level is perceived by the user.

The same measurement quantities check is fundamental in mobility events management (i.e., handovers), as it indicates a potential drop of the serving signal and hints the network to react. An additional measure for 5G traffic assessment, mobility optimization, is mobility interruption time, which denotes the shortest time duration supported by the system during which a user terminal cannot exchange user plane packets with any base station during transitions. It includes the time required to execute any radio access network procedure, radio resource control signaling protocol, or other message exchanges between the mobile station and the radio access network. The mobility interruption time corresponds to the evaluation of the latency introduced by the system, for verifying whether or not it is small enough in terms of satisfying some predefined use-case-dependent latency requirements.

For *QoS* Optimization, latency and throughput are metrics explicitly characterizing the stringent *QoS* requirements on the reliability of a target user experience.

The outlined measurements for RAN optimization are accessible for collection through MDT procedures, wherein RRC (radio resource control) signaling and RRC data are utilized. To increase the chances to supply automated data collection for a particular target use case, the radio measurements accessible through MDT configuration and reporting procedures are identified as applicable to the AI/ML framework also [10]. Such a set of the key performance indicators for RAN Optimization specific use cases, for ML-driven data collection (accessible via RRC procedures), can be grouped as indicated in Table 2.

Table 2. Key performance indicators per RAN Optimization use case.

Coverage Optimization	Mobility Optimization	*QoS* Optimization
RSRP, *RSRQ*, *SINR*	*RSRP*, *RSRQ*, *SINR* Interruption time	Latency Throughput

For the purpose of the ML-based evaluation method, we consider the network function requests that parametrized input data to train an ML model that is RAN optimization use case specific (as per Table 2). Consequently, the input data function may target the data set to be denoted as:

$$B = \{Coverage, Mobility, QoS\}, \tag{4}$$

where coverage, mobility, and *QoS* denote key performance measurements taken per RAN optimization use case. Network function targeting measurement results taken specifically only for a single RAN optimization use case (e.g., coverage optimization) admits a subset of dataset B:

$$b_{Coverage} = \{RSRP, RSRQ, SINR\}. \tag{5}$$

As a result, each metric collected as an individual entry, or a subset of dataset B will represent a key performance indicator of the RAN Optimization use case.

3.3. Data Selection Model

The availability of the data is considered the key to train any ML model. The most desirable way for handling data collection is to use measurements reports from the user devices. The MDT feature, considered as the functionality of enabling a variety of radio measurements and performance characteristics collection, can provide the data for 5G traffic evaluation, along with the data for RAN optimization.

Consequently, to adapt data collection for an ML algorithm for a joint evaluation of RAN optimization and 5G traffic, the 5G RAN parametrizing input data is proposed to trigger the MDT configuration procedure to collect any combination of the A and B dataset:

$$Input\ Data \in A \cup B \tag{6}$$

In one exemplary scenario, when the operator decides a coverage optimization is desirable, it will instruct any device to collect data on the radio signal level on a network-centric basis, i.e., *RSRP*/*RSRQ*/*SINR* in a unit area (e.g., cell). However, in the case the 5G-NR system performance is further deemed to be optimized in a user-centric manner, the particular user-specific performance indicator should be monitored (as per Table 1). For such user-centric observability, where the user's performance represents involvement in different types of 5G traffic, combined requirements need to be reliably fulfilled. For instance, *eMBB* rate maximization and *URLLC* latency minimization. The input data for such a mixed scenario (i.e., for RAN coverage optimization for the user with *eMBB* and *URLLC* service) can be denoted as follows:

$$Input\ Data = a_{MBB} \cup a_{URLLC} \cup b_{Coverage} \tag{7}$$

3.4. Data Filtering

Many algorithms have been proposed to limit data collection in the machine learning and data science literature [11]. For instance, principal component analysis (PCA) is a well-known mathematical technique that transforms a set of correlated variables into a set of uncorrelated variables known as principal components. Thus, PCA can be used as a 5G radio characteristic extraction technique to reduce the dataset to fewer dimensions. Technically, PCA identifies the given vectors transformed with the highest eigenvalues of a covariance matrix and uses those to project the input parameters into a new subspace of a lower number of indicators.

The RAN optimization-driven task, with *RSRP*, *RSRQ*, and *SINR* parameters (i.e., the input data denoted as $b_{Coverage}$ and $b_{Mobility}$), qualifies to apply the correlation coefficient measure due to the normalized meaning of the parameters. Table 3 shows the different values of the parameters, which correspond to varying levels of radio signal quality, where a very good signal level is associated with an excellent perception of the network by the end user (e.g., seamless coverage, close to the base station), while a very poor signal level is associated with a network perception having insufficient quality, lacking continuous coverage (e.g., at cell edge far from the base station).

Table 3. Key performance indicators value ranges for coverage and mobility optimization use cases.

Radio Conditions	*RSRP* (dBm)	*RSRQ* (dB)	*SINR* (dB)
Excellent	≥ -80	≥ -10	≥ 20
Good	-80 to -90	-10 to -15	13 to 20
Medium	-90 to -100	-15 to -20	0 to 13
Poor	≤ -100	< -20	≤ 0

Each parameter that determines the quality of the 5G signal can help to read the detailed characteristics of the connection. *RSRP*—the average power of the received pilot signals (reference signal) or the level of the received signal from the base station. For instance, when *RSRP* = -120 dBm and below, the connection may be completely unavailable. *RSRQ* indicates the quality of the received reference signal. If its range is below -20 dB (poor), it indicates typically high cell load or high noise level. *SINR* is the 5G signal-to-noise ratio. For *SINR*, the higher the value, the better the signal quality. At *SINR* values below 0, the connection perception will be very poor (e.g., very slow connection), since this value characterizes the situation when there is more noise in the received signal than the useful part, and the probability of losing a connection is high [15].

A correlation coefficient for *RSRP*, *RSRQ*, and *SINR* can result in proof that the performance indicators in some scenarios depend on each other [16]. Thus, in the proposed data selection model, practically, the function of data filtering can be performed on a data set, B, to reduce the data dimension in RAN optimization-related data collection.

3.5. Data Processing/Weighting

The differentiated purposes of the optimization purposes can be handled by adopting the parametrization of the input data to the ML algorithm, based on the ANN's model. Data collection can be formed as a sequence of layers, representing 5G traffic performance (user-centric) and RAN optimization use case (network-centric) generally denoted as:

$$Hidden\ Layer\ Data = f(A, B) \tag{8}$$

A hidden layer can represent a layer in between input layers (data set A and data set B) and output layers (prediction of 5G traffic characteristics and prediction of RAN optimization characteristics). Emerging from this deep learning setup, the artificial neurons implicate any combination of the parameters belonging to data set A and data set B. An algorithm in the network can apply the function as a set of weighted parameters and produce an output through the activation of a model with a complex data relationship. Activation of the function for generic RAN optimization, with equal importance of measuring coverage, mobility and *QoS* would target, in the intermediate layer, all key performance measurements taken per RAN optimization use case (according to Table 2: *RSRP*, *RSRQ*, *SINR*, interruption time, latency, and throughput) with equal weight (e.g., $w_4 = 1$, $w_5 = 1$, $w_6 = 1$). The function applied to discover the relationship of the RAN optimization use case with *eMBB* service would take a set of *eMBB*-specific parameters (according to Table 1: peak data rate and user experience data), equally weighted too (e.g., $w_1 = 1$). For this particular scenario, the applied function can calculate the hidden layer with suppressed relationship for *URLLC* and *mMTC* services, by weighting the relevant parameters with zero (e.g., $w_2 = 0$ and $w_3 = 0$).

In this novel setup, defining ANN principles for running the ML model in 5G RAN, input, and hidden layers formed by sets of measurement quantities performed according to the measurement configuration are distinguished (see Figure 5). The configuration message sent to the UE through RRC signaling determines and requests the intended datasets. The specific performance indicators measured by the UE and received by the RAN can be weighted and treated as hidden layers (Figure 8).

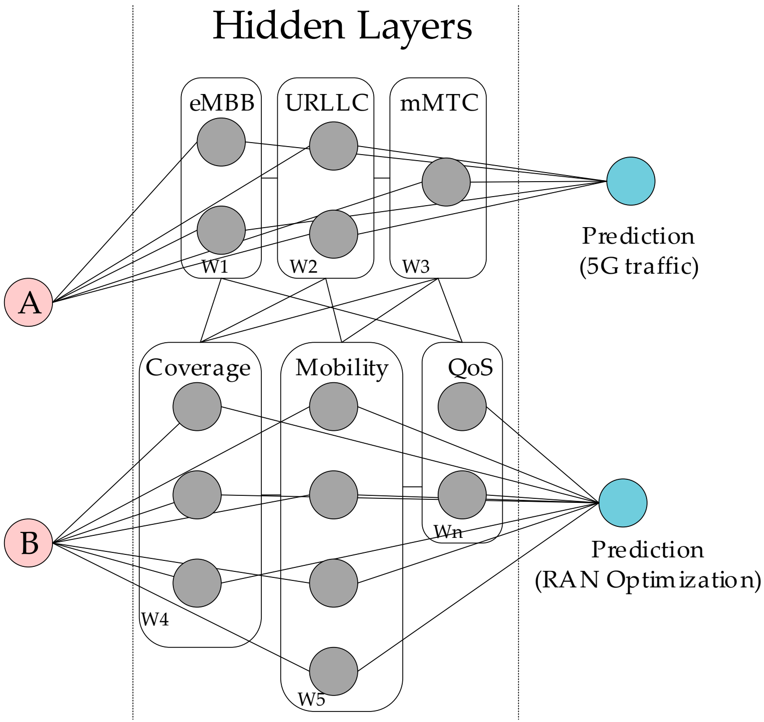

Figure 8. The model for training data collection in different 5G usage scenarios.

The input layer corresponds to the generic classification of the two main streams (5G traffic performance input data A and/or RAN optimization input data B). The individual metrics in the preceding and the succeeding sets of the hidden layer are in this model connected by weighted connections that facilitate datasets training according to the varying importance of the key performance indicators in the undertaken evaluation process.

Figure 9 depicts what such dataset processing may look like. Let us assume the optimization purpose is RAN coverage optimization use case for an *eMBB* user. At the first iteration of ML model training, the RAN takes an action to order a dataset for a specific user, with *eMBB* service, for which good coverage should be ensured. To facilitate such action the UE may be provided by the RAN with the RRC configuration to collect *RSRP*, *RSRQ*, *SINR*, peak data rate, and experienced data rate. With the MDT feature, the UE can report the metrics periodically in real time or log the measurements to provide periodic data collectively in a deferred way.

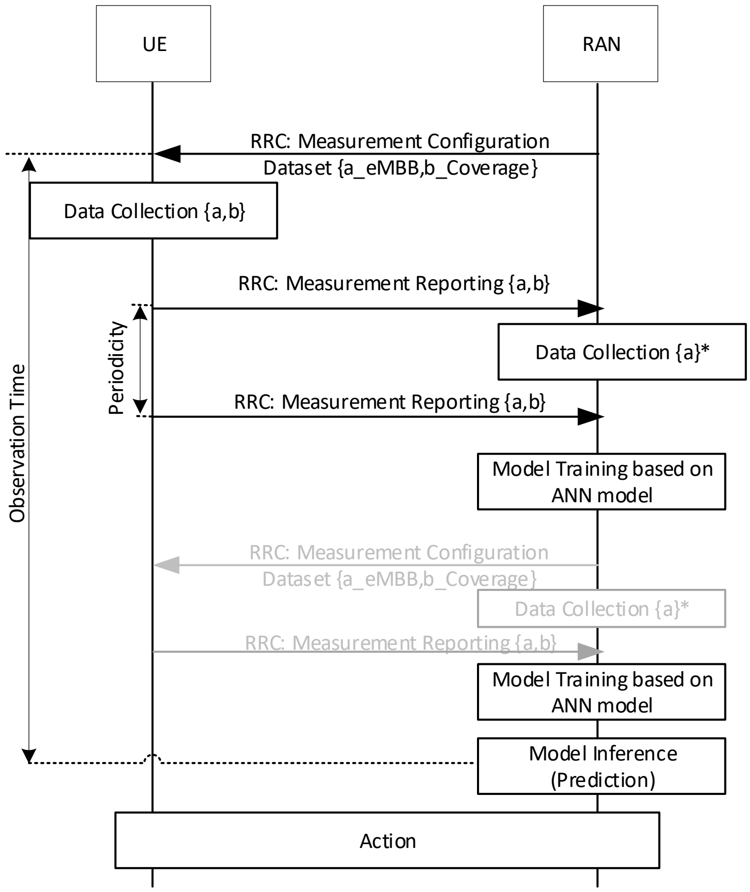

Figure 9. The model for training data collection for coverage optimization for *eMBB* traffic.

In the observation phase, the gNB performs training of the scheduled datasets based on the user input (a,b) or complementary datasets recorded by itself {a}*, according to the equivalent metrics available in the gNB itself, as per [17]. Model training based on the ANN principle enables the application of different weighting conditions and importance on data subsets, so that particular components (a,b) are collected with a varying periodicity or varying evaluation target. In the assumed scenario, the applied ANN-based algorithm, in the gNB, requires user input for the *eMBB* service (a_{MBB}) and coverage optimization ($b_{Coverage}$). The algorithm plays a determinant role to set appropriate weights for the parameters representing these input data (e.g, w_1 = 1 and w_4 = 1). Since the parameters that measure *eMBB* service can be observed by the gNB itself, the measurement results for the *eMBB* service (a_{MBB}) can be replaced by the input from the gNB {a_{eMBB}}*. The user input can be superseded, with no meaningful difference in the context of the parameter origin (user input or network measure). For simplicity, equal weights can denote the equal importance of the parameters from different data sets. However, hidden layers in the ANN model enable the discovery of more complex dependencies among each and every subset of the parameters and features, by assigning different importance and different weights to each of them.

The shadowed step is used to indicate an incremental step for the measurement configuration modification (change the data set or prolong observation time). The process guarded by certain observation rules (e.g., time), let the gNB collect the data sets and observe how/if the collected values of parameters from datasets meet the target values of the requested performance indicators and what the relationship between observed parameters is.

3.6. Data processing/Feedback

The role of the feedback function in the ML algorithm located in RAN is to process the input data collection toward a prediction. For that purpose, the actor may associate each parameter with a target value. For the analyzed example scenario, for RAN coverage optimization, the network may set an observation rule: the "*RSRP*" parameter does not drop below the "target threshold". In the undertaken optimization tasks, if the observation is performed for a single user, the drop may indicate a "Feedback: Bad coverage". The feedback forecasts the associated prediction and typically implies an action. However, in practice, a single indication in a radio network on a signal drop should not trigger a prediction that would result in network parameters changes. The remedy for that can be a cell change for the user (i.e., trigger mobility), by setting mobility parameters in a way the user switches to a neighboring cell (overlaying radio layer).

Before making a decision on the following action, a more detailed observation can rely on a different weighting setting to the RAN optimization parameters. For instance, an iteration on joined feedback on the *RSRP* + *SINR* may be required, where the complementary parameter *SINR*, can indicate whether a drop in signal is caused by noise interferences or purely weak coverage. In this case, the measurement configuration should be followed by the updated configuration with a request to provide a more complete dataset based on the received feedback.

If the user feedback is repetitively provided for an observation period, it becomes more reliable, leading to a certain set of decisions, such as change cell, adjust radio parameters, adjust the antenna, change scheduling policy, release resources for *eMBB* users, increase modulation scheme, MIMO layers, and add serving carriers (set carrier aggregation).

3.7. Outcome Evaluation

The association of the input data with various optimization uses cases, followed by reliable feedback, intends to achieve a predictive outcome. The result aims at the prediction of 5G traffic performance and avoidance of the 5G network performance dropping below expectations. Such a procedure is built on preceding tailored datasets selection and their appropriate processing that enables observation of whether the expected performance targets are met. To provide a good prediction for the observed 5G traffic characteristics, the joined evaluation with intended RAN optimization use cases sets different targets. Table 4 summarizes the expected performance depending on the nature of traffic. It can be observed that the expected target values vary for the same 5G traffic type in different scenarios. For ML-driven 5G RAN evaluation outcome, it determines a factor to make a decision in a general structure of an ML-based predictive model for 5G traffic in a given RAN optimization use case.

Table 4. Target values of the 5G key performance indicators depending on RAN optimization use case.

Performance Indicator	Use Case	Target Value
Peak Data Rate	*eMBB*	Downlink 20 Gbit/s Uplink: 10 Gbit/s
User Experience Data Rate (Throughput)	*eMBB* *QoS*	Downlink 100 Mbit/s Uplink: 50 Mbit/s
Mobility interruption time	*URLLC* Mobility	0 ms
Latency	*URLLC* *QoS*	Control Plane: 20 ms User Plane: 1 ms
Connection density	*mMTC*	1 000 000 devices per km^2.
Radio signal level	Coverage Mobility	$RSRP > -90$ dBm $RSRQ > -15$ dB $SINR > 13$ dB

In a combined evaluation methodology, where the evaluation is applied to a set of use cases, the outcome can be classified according to a decision tree using a random forest (RF) classifier. This decision tree is developed as an ensemble classification technique that uses "parallel ensembling" which fits several decision tree classifiers in parallel on different data set subsamples, and uses majority voting or averages for the outcome or final result [11]. Such methodology in RAN can apply to a generic network performance assessment, without strict focus or strength requirements for a specific optimization use case (coverage, mobility, or *QoS*). In the scenario, where datasets are produced as inputs from a few domains and overall, reasonable performance is deemed. The random forest-based algorithm establishes a baseline to predict averaged and aggregated assessment (Figure 10).

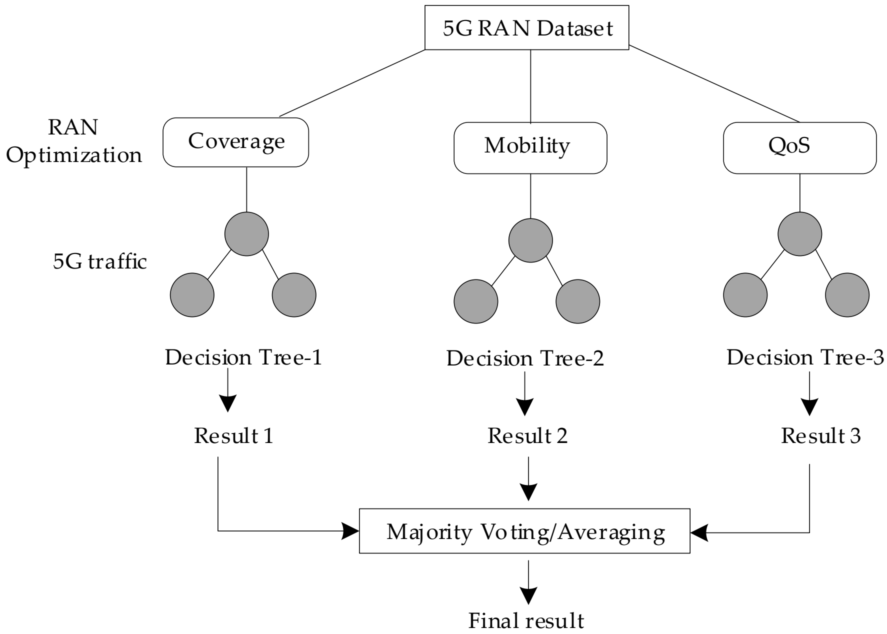

Figure 10. An example of a random forest-based decision for 5G traffic evaluation in association with multiple RAN optimization use cases.

The ANN model can be further considered beyond observability of an individual user or a single area unit. As studies in [18–23] have shown, applying convolutional neural networks (CNN) and graph neural networks (GNN) in the wireless network may reveal abilities to achieve optimal radio resource allocation in the network, predict channel stability, minimize spectrum allocation costs and maximize radio resource utilization. [18] and [19] teach CNN use for network slicing.

Once the data collection from the above method limits the input data function to data set A, it can be considered jointly with methodologies developing ML-based radio resource allocation in the network, as in [19]. It can offer a means to minimize spectrum allocation costs for the tenants of selected slices and guarantee desired service level agreements (SLA).

If suitably defined, network slicing observability and optimization can be applied together with GNN-based methods [20]. Graph-based deep learning can help to employ visualization of end-to-end network performance [21].

4. Discussion

The general idea of using AI/ML as a component of network optimization, undertaken by 3GPP, is adopting an intelligence to automated data collection through the intended RAN optimization use case. The outlined architecture for AI/ML adoption in RAN possesses generalization abilities in the context of the three essential types of 5G communication (*eMBB*, *URLLC*, and *mMTC*). Applying ML practices without considering the usage scenarios for 5G traffic can run into a problem with suboptimal optimization tasks or ML algorithm outcomes. For instance, maximally optimizing cell coverage does not guarantee *URLLC* service, characterized by stringent requirements such as reliability and latency. The ability to adapt the ML-based RAN framework for the converged evaluation criteria (i.e., 5G traffic requirements and RAN optimization use cases requirements) is envisioned to improve the robustness of ML algorithms. The applied data classification with the ANN technique can help to reduce losses and avoid redundancy in data collection campaigns, thus saving computing cost and optimization processing time, while raising accuracy. The proposed evaluation technique can help to distinguish data assessment for a single user and dedicated 5G traffic type in a unit area, where the significance of different needs, varies. The distinction of the approaches can be applicable and may serve different purposes, such as the collective traffic evaluation in a certain area or the user's specific evaluation. Both are based on the key performance metrics generated and collected from 5G users. While the availability of the real network data from 5G traced traffic appears challenging, the method can be applied to data generated in 5G testbeds and prototypes [24–34].

5. Conclusions

The presented method describes a holistic concept of the ML-driven optimization in 5G RAN that establishes a novel approach for 5G traffic evaluation with the RAN optimization use case-based classification. It yields a conceptual contribution for computational evaluation analysis and validation of the assumptions in future research work. Additionally, by taking the AI/ML-based RAN architectural framework, defined in the 3GPP standard, the method paves the way to applying an ML-based algorithm in a standard and intervendor-compliant manner.

Author Contributions: Conceptualization, M.T. and K.S.; methodology, M.T.; formal analysis, M.T.; investigation, M.T.; writing—original draft preparation, M.T.; writing—review and editing, M.T. and K.S.; visualization, M.T.; supervision, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Wroclaw University of Science and Technology under the grant number 8211104160/K34W04ND03.

Acknowledgments: The authors would like to express their appreciation for funding by Wroclaw University of Science and Technology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Telecommunication Union; Measuring Digital Development, Fact and Figures 2021. Available online: https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx (accessed on 1 January 2023).
2. International Telecommunication Union; 5G-Fifth Generation of Mobile Technologies. Available online: https://www.itu.int/en/mediacentre/backgrounders/Pages/5G-fifth-generation-of-mobile-technologies.aspx (accessed on 1 January 2023).
3. International Telecommunication Union Recommendation ITU-R M.2083-0. IMT Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond. Available online: https://www.itu.int/rec/R-REC-M.2083 (accessed on 1 January 2023).
4. International Telecommunication Union Recommendation ITU-R M.1645. Framework and Overall Objectives of the Future Development of IMT-2000 and System Beyond IMT-2000. Available online: https://www.itu.int/rec/R-REC-M.1645/en (accessed on 1 January 2023).
5. International Telecommunication Union Recommendation ITU-R M.2150-1. Detailed Specifications of the Terrestrial Radio Interfaces of International Mobile Telecommmunications-2020 (IMT-2020). Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2150-1-202202-I!!PDF-E.pdf (accessed on 1 January 2023).
6. 3rd Generation Partnership Project Technical Specification: 3GPP TS23.501. System Architecture for the 5G System. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144 (accessed on 1 January 2023).
7. 3rd Generation Partnership Project Technical Specification: 3GPP TS38.300. NR; NR and NG-RAN Overall Description; Stage-2. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3191 (accessed on 1 January 2023).

8. 3rd Generation Partnership Project Technical Specification: 3GPP TS 36.902. Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-Configuring and Self-Optimizing Network (SON) Use Cases and Solutions. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2581 (accessed on 1 January 2023).
9. 3rd Generation Partnership Project Technical Specification: 3GPP TS37.320. Radio Measurement Collection for Minimization of Drive Tests (MDT); Overall Description; Stage 2. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2602 (accessed on 1 January 2023).
10. 3rd Generation Partnership Project Technical Specification: 3GPP TR 37.817. V1.0.0 Study on Enhancement for Data Collection for NR and ENDC. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3817 (accessed on 1 January 2023).
11. Mohammed, M.; Khan, M.B.; Bashier Mohammed, B.E. *Machine Learning: Algorithms and Applications*; CRC Press: Boca Raton, FL, USA, 2016.
12. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [CrossRef] [PubMed]
13. NGMN. NGMN Recommendation on SON and O&M Requirements. Available online: https://www.ngmn.org/publications/ngmn-recommendation-on-son-and-om-requirements.html (accessed on 1 January 2023).
14. International Telecommunication Union Recommendation ITU-R M.2410. Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s). Available online: https://www.itu.int/pub/R-REP-M.2410 (accessed on 1 January 2023).
15. Zyxel Networks. 2022. Available online: https://support.zyxel.eu/hc/en-us/articles/4406391493778-5G-signal-quality-parameters (accessed on 1 January 2023).
16. Afroz, F.; Subramanian, R.; Heidary, R.; Sandrasegaran, K.; Ahmed, S. *SINR*, *RSRP*, RSSI and *RSRQ* Measurements in Long Term Evolution Networks. *Int. J. Wirel. Mob. Netw.* **2015**, *7*, 113–123. [CrossRef]
17. 3rd Generation Partnership Project Technical Specification: 3GPP TS38.314 NR, Layer 2 Measurements. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3671 (accessed on 1 January 2023).
18. Debbabi, F.; Rihab, J.; Chaari, L.; Aguiar, R.L.; Gnichi, R.; Taleb, S. Overview of AI-Based Algorithms for Network Slicing Resource Management in B5G and 6G. In Proceedings of the 2022 International Wireless Communications and Mobile Computing (IWCMC), Dubrovnik, Croatia, 30 May 2022–3 June 2022; 2022; pp. 330–335.
19. Azimi, Y.; Yousefi, S.; Kalbkhani, H.; Kunz, T. Applications of Machine Learning in Resource Management for RAN-Slicing in 5G and beyond Networks: A Survey. *IEEE Access* **2022**, *10*, 106581–106612. [CrossRef]
20. He, S.; Xiong, S.; Ou, Y.; Zhang, J.; Wang, J.; Huang, Y.; Zhang, Y. An overview on the application of graph neural networks in wireless networks. *IEEE Open J. Commun. Soc.* **2021**, *2*, 2547–2565. [CrossRef]
21. Jiang, W. Graph-based Deep Learning for Communication Networks: A Survey. *Comput. Commun.* **2022**, *185*, 40–54. [CrossRef]
22. Hui, N.; Sun, Q.; Wang, Y.; Zhang, Z.; Tian, L.; Feng, C.; Guan, Z. Wireless Resource Allocation Based on Multiplexing and Isolation in Sliced 5G Networks. In Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, 10–13 April 2022; 2022; pp. 1629–1634.
23. Rkhami, A.; Hadjadj-Aoul, Y.; Outtagarts, A. Learn to Improve: A Novel Deep Reinforcement Learning Approach for beyond 5G Network Slicing. In Proceedings of the 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 9–12 January 2021; 2021; pp. 1–6.
24. 5G Campus Networks: A First Measurement Study. Available online: https://ieee-dataport.org/open-access/5g-campus-networks-measurement-traces (accessed on 1 January 2023).
25. Available online: https://www.free5gc.org/ (accessed on 1 January 2023).
26. Available online: https://github.com/free5gc/free5gc (accessed on 1 January 2023).
27. Available online: https://github.com/open5gs (accessed on 1 January 2023).
28. Available online: https://github.com/my5G/my5G-RANTester (accessed on 1 January 2023).
29. Available online: https://github.com/5g-empower (accessed on 1 January 2023).
30. Available online: https://5g-empower.io/ (accessed on 1 January 2023).
31. Available online: https://www.5g-induce.eu/ (accessed on 1 January 2023).
32. Available online: http://simu5g.org/ (accessed on 1 January 2023).
33. Available online: https://5g-pagoda.aalto.fi/ (accessed on 1 January 2023).
34. Available online: http://5g-transformer.eu/ (accessed on 1 January 2023).