Annals of Telecommunications (2024) 79:15–33
https://doi.org/10.1007/s12243-023-00980-9

Generative AI in mobile networks: a survey

Athanasios Karapantelakis[1] · Pegah Alizadeh[2] · Abdulrahman Alabassi[1] · Kaushik Dey[3] · Alexandros Nikou[1]

Received: 20 June 2023 / Accepted: 26 July 2023 / Published online: 17 August 2023

Abstract

This paper provides a comprehensive review of recent challenges and results in the field of generative AI with application to mobile telecommunications networks. The objective is to classify the literature using an approach that encompasses the type of generative AI technology employed, the functional purpose, and the specific component of the mobile network that each solution targets. Moreover, performance requirements for generative AI applications are considered. Thereafter, state-of-the-art generative AI algorithms and an examination of their use cases across various industry verticals are presented. The discussion extends to the current level of AI integration in telecom standardization bodies, such as the 3rd Generation Partnership Project (3GPP). Finally, the open research challenges that the generative AI technology aims to address are thoroughly investigated.

Keywords Generative AI · Telecom · 6 G · 5 G · Survey · Application

1 Introduction

Recent breakthroughs in generative Artificial Intelligence (AI), such as transformer-based Large Language Models (LLMs), have attracted attention from both the scientific community and industry alike. Generative Pretrained Transformers (GPTs), for example, such as OpenAI's GPT−3.5 and recently GPT-4, have formed the basis for a number of Natural Language Processing (NLP) applications ranging from text summarizers and generators to language translators. Finetuned versions of these transformers have evolved into advanced chatbots that allow easy accessibility of vast corpora of knowledge to humans. Teaching and sales assistants, product troubleshooters, and assistants for upskilling of employees and personnel are all use cases where generative AI has recently found good popularity.

1.1 Generative AI and relevance to mobile networks

The telecommunications industry and in particular Fifth Generation of Mobile Networks (5 G) use AI algorithms to analyze, optimize, and troubleshoot network functionality, from resource allocation and sharing to anomaly detection and path loss prediction [1]. Such AI algorithms are known as "discriminative," as they train machine learning (ML) models using network data that is either readily available, in form of historical datasets ("offline learning"), or is retrieved in real time from the mobile network ("online learning").

Sixth Generation of Mobile Networks (6 G) is expected to serve more demanding use cases in terms of connectivity, capacity, data rates, latency, mobility, and reliability [2]. These requirements also introduce a set of challenges for use of Machine Learning (ML) models.

One challenge is about *data observability*, meaning the collection and transport of data needed for training of ML models. The overall process consumes computational, storage, and transport resources that can otherwise be used to serve mobile network traffic. The challenge compounds when

✉ Athanasios Karapantelakis
athanasios.karapantelakis@ericsson.com

Pegah Alizadeh
pegah.alizadeh@ericsson.com

Abdulrahman Alabassi
abdulrahman.alabassi@ericsson.com

Kaushik Dey
deykaushik@ericsson.com

Alexandros Nikou
alexandros.nikou@ericsson.com

[1] Ericsson Research, Ericsson AB, Torshamnsgatan 21, Kista, Stockholm 164 83, Sweden

[2] Ericsson Research, Ericsson France, 25 Carnot, Massy 91300, France

[3] Ericsson Research, Ericsson India, MGR Salai, Perengudi, Chennai, Tamil Nadu 600096, India

considering that model training is not an isolated process happening once; instead, models may need to be retrained several times during their lifetime. Furthermore, data may not always be available, for example, in cases where data collection involves User Equipment (UE). Generative AI models may be used in this case to provide data for ML model training and can be deployed close to where training takes place.

Another challenge is about *safe learning*, and specifically, cases where discriminative models are trained in real time, using online learning approaches such as Reinforcement Learning (RL) and Active Learning (AL). Contrary to observing data a priori, building a dataset, and using the dataset to train the model, in online learning, the model is trained serially, as data becomes available. However, in such cases, predictions of models can be inaccurate, inadvertently leading to potentially bad predictions that can negatively impact network operation. In this case, generative AI can assist in creation of Digital Twins (DTs), which provide a safe environment to improve RL-trained models.

Customer-related tasks such as creating subscription plans and product bundles given mobile subscriber requirements and network deployment configurations given mobile operator requirements are also expected to become more complex, as 6 G will diversify in terms of use cases, radio access technologies, and types of supported UE. Additionally, with strict privacy and legal requirements which often vary across countries, collecting customer usage data required for training discriminative algorithms is often a challenge. This is an area where generative AI-related solutions such as transformers can help, using digital assistants and NLP, to elicit requirements from the user, and converting those requirements into machine-readable parameters that can be further actuated upon.

The aforementioned challenges are not only limited to 6 G networks. In particular, as 5 G deployments grow denser, these challenges also apply to existing mobile networks and call for undertaking research efforts to address them.

1.2 Paper overview

The main contribution of this paper is to provide a historical perspective and highlight recent work in the field of generative AI with application to mobile telecommunications networks. For classifying the literature, we use an orthogonal system that is based on both the type of generative AI technology used and functional purpose, but also the part of the mobile network each solution is applied to. The paper also describes outstanding challenges and open research questions that can potentially be explored.

The rest of the paper is structured as follows: Section 2 provides an overview of the state of the art of generative AI algorithms and use cases for these algorithms in industry verticals. The section ends by discussing the current level of AI in telecom standardization bodies such as Third Generation Partnership Project (3GPP). Section 3 describes performance requirements for generative AI and how they are linked to 6 G requirements. Section 4 describes current research on generative AI for telecommunication networks. Section 5 describes open research questions and opportunities for future research, as well as outstanding challenges. Finally, we conclude in Section 6 with a recapitulation of the paper content.

2 Background

2.1 Generative AI algorithms

Generative approaches date back to the 1950 s, one of the first being Markov Chains, i.e., statistical models predicting the probability of an event occurring given the present state. However, it was not until the 1980 s, when neural network-based Boltzmann Machines (BMs) were introduced, that generative approaches became relevant in solving real-world applications [3]. Restricted Boltzmann Machines (RBMs) in particular, that were introduced in mid-2000 s, have found application in a wide range of fields, from natural language processing to computer vision [4]. RBMs are shallow, 2-layer generative models that are able to learn a probability distribution from their inputs. Stacking multiple RBMs one after the other is the basis of more complex Deep Belief Networks (DBNs), which compound advantages of single RBM [5].

Generative Adversarial Networks (GANs) are another architecture for training of generative neural models [6]. GANs have enjoyed success in the computer vision field, in areas such as image and texture synthesis and manipulation, image translation, but also NLP and music [7].

Variational Autoencoders (VAEs) are also generative neural models that, given highly dimensional input data, generate a more compact representation of this input data [8]. This representation is also known as “latent space” and probabilistically represents the input by describing the probability distribution of every input value. Given an “encoding,” i.e., a latent space representation, every distribution in this space can be sampled to generate new output that approximates the input. Compared to GANs, VAEs produce a less accurate representation of input (e.g., images of lower fidelity) but train faster [9].

Departing from the visual domain and moving towards linguistics, Recurrent Neural Networks (RNNs) and RNN-based Long-Short Term Memory networks (LSTMs) have long been the *de-facto* models for text generation type of use cases [10]. RNNs are sensitive to sequential data, where the prediction of the model is not only dependent on current input, but also previous predictions. Given that sentences are sequences of words, RNNs are theoretically a good match for text generation type of problems. Such problems may include, for example, text completion, translation, and summarization. In the literature, it has been acknowledged that vanilla RNNs suffer from the vanishing gradient problem, which creates challenges when learning long sequences of data, such as text [11]. LSTMs, a special type of RNN, were introduced in order to solve this type of problem [12].

However, RNNs and LSTMs suffer from performance problems, as they have to sequentially process the input. For example, considering a sentence used as input, processing is done token by token, i.e., word per word. To address these issues, the transformer architecture was introduced, allowing for parallelization of input processing and therefore resulting in reduction of training time [13]. Transformers also introduced the concept of self-attention. Traditionally, LSTM-based models tried to remember the whole input sequence, for example, the whole sentence, before producing the output (e.g., a translation of that sentence). The problem with this approach is that loss of information can occur, especially for longer sequences, as the LSTM-based models have to implicitly remember the complete input sequence. Instead, transformers focus only on relevant parts of the sequence, for example, words that are related to each other in a sentence. This allows transformers to remember larger parts of input sequences and also produce output faster. Transformer-based models such as GPT [14] and Bidirectional Encoder Representations from Transformer (BERT) [15] have demonstrated ability to capture large corpora of information, earning the name LLMs. In addition, with relatively small effort, such models can be fine-tuned to serve real-world applications. ChatGPT is one such popular example—finetuned from GPT using supervised learning and Reinforcement Learning with Human Feedback (RLHF), it provides a dialogical, Question and Answer (Q&A) type of functionality accessible to human users [16].

Figure 1 provides a taxonomy of AI algorithms, based on previous taxonomy presented in [6]. The generative algorithms are split into three main categories, depending on the technique they follow to generate content.

- Explicit density models model the probability distribution of the data, by learning the Probability Density Function (PDF) of the input. Depending on the computational complexity of the operations, explicit density models are either tractable or approximate (also known as intractable). Intractable models are better suited for complex, high-dimensional spaces, whereas tractable models, being simpler and more computationally efficient, are better suited for low-dimensional input. Some of the explicit density function learning examples are two neural network architectures, including Neural Autoregressive Density Estimation (NADE) and Masked Autoencoder for Distribution Estimation (MADE) for tractable density. However, Variational and Markov Chain Monte Carlo (MCMC) models are two examples of approximate density.
- On the other hand, implicit density function models—such as GANs—do not model the PDF. Instead, they learn them indirectly by generating output from the data distribution. As these models do not require a parametric form of PDF, they can sample from more complex, multimodal distributions than their explicit density counterparts. Under implicit density models lies Markov chain models, Generative Stochastic Network (GSN), and GAN.
- The last category of generative AI algorithms is transformers. Transformers are not explicitly density models, but instead rely on an internal attention mechanism to learn relationships between input tokens and generate new data out of these learned relationships.

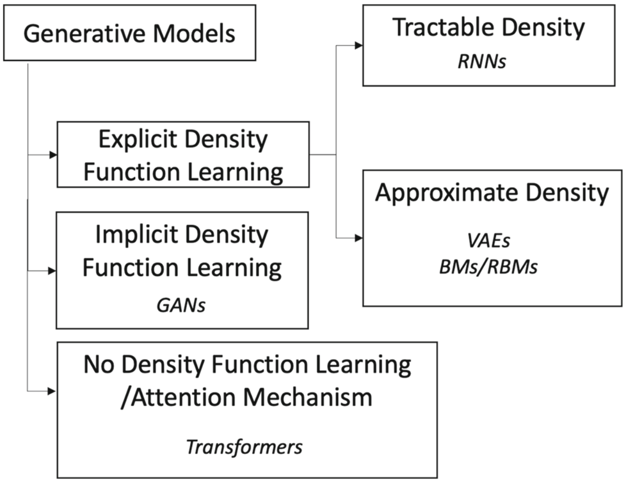
Generative Models
Explicit Density Function Learning
Implicit Density Function Learning
GANs
No Density Function Learning /Attention Mechanism
Transformers
Tractable Density
RNNs
Approximate Density
VAEs
BMs/RBMs

Fig. 1 Taxonomy of generative AI algorithms

2.2 Use cases and industrial applications

There are several use cases in which generative AI has already been applied in. In this section, we present some of the prominent areas of application.

- Text generation includes any type of text content such as summaries, suggestions, translations, answers to user queries, production of original text such as poems, speeches, or novels, captioning, etc. [17, 18]. Tools such as ChatGPT, for example, may impact all aspects of a business, from marketing and sales to operations, engineering, legal, Human Resources (HR), and employee working efficiency [19]. Text generators are not limited to natural language text, but may also generate code. Tools generating code can be used by their human

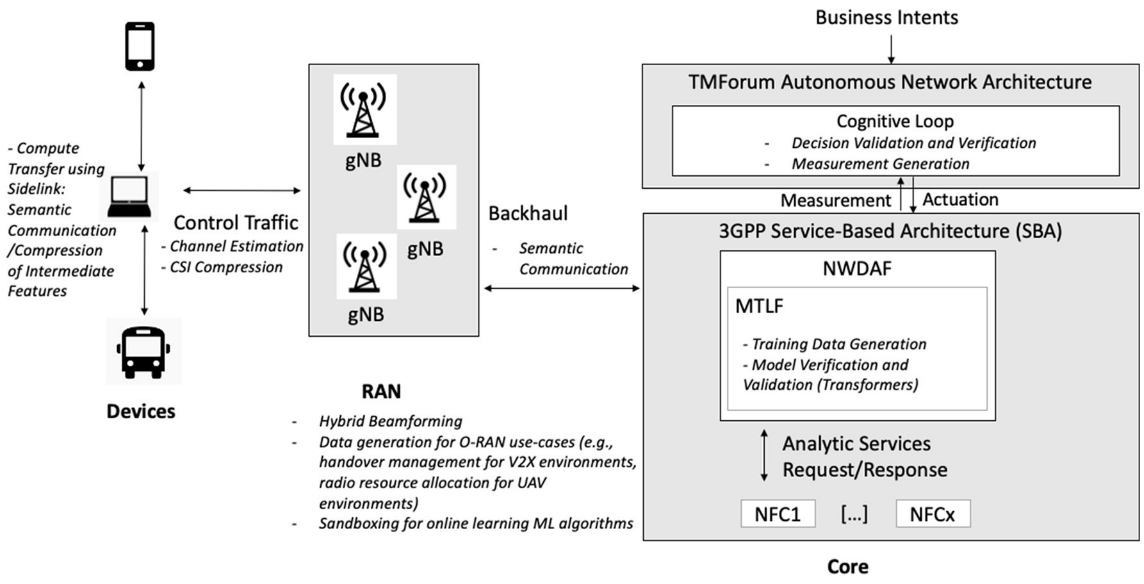
Business Intents
TMForum Autonomous Network Architecture
Cognitive Loop
- Decision Validation and Verification
- Measurement Generation
Measurement
Actuation
3GPP Service-Based Architecture (SBA)
NWDAF
MTLF
- Training Data Generation
- Model Verification and Validation (Transformers)
Analytic Services Request/Response
NFC1
[...]
NFCx
Core
- Compute Transfer using Sidelink: Semantic Communication /Compression of Intermediate Features
Devices
Control Traffic
- Channel Estimation
- CSI Compression
gNB
gNB
gNB
Backhaul
- Semantic Communication
RAN
- Hybrid Beamforming
- Data generation for O-RAN use-cases (e.g., handover management for V2X environments, radio resource allocation for UAV environments)
- Sandboxing for online learning ML algorithms

Fig. 2 Potential for generative AI application in mobile network standardization

counterparts in pair-programming sessions, increasing the overall quality of the code and turnaround time [20].

- Image generation is another area where generative AI approaches excel in. Text-to-image generation (T2I), where the generative model uses a textual description to produce an image [21], and image-to-text (I2T) generation, where the generative model produces a textual description from an image [22], are both areas that are heavily researched and wherein generative AI models enjoy wide industrial applicability.
- Audio generation is another prominent category. Audio diffusion models, for example, generate speech from text [23], while music generation models implicitly learn properties such as harmony, structure, melody, and texture as well as instrumentation and orchestration of music pieces to produce new music [24].
- Video generation, or synthesis, is still in early stages compared to other domains such as images; however, tools such as 3D RNNs show promise [25].

Another category of use cases is generation of content in a form that is not necessarily human-readable, but is compact enough to yield other benefits. One example is use of generative AI models for semantic communication that aims to increase transmission efficiency in communication networks. In conventional communication and regardless of the content being transmitted, all information for the communication is transmitted from sender to receiver. On the other hand, semantic communication transmits symbols describing the content of the transmission, which is synthesized on the receiver side. Transformer-based approaches, such as DeepSC, encode text input into complex symbol streams, decoded back to text at the receiver, saving bandwidth during transmissions [26]. We further discuss semantic communication and its application in mobile telecommunication networks in Section 4.6.1.

2.3 AI standardization activities in mobile networks

From the perspective of standardization, and at the moment of this writing, there are different AI functions being investigated, belonging to different parts of the mobile network (see Fig. 2). While the AI algorithms used to implement functionality are implementation-specific, generative AI algorithms such as transformers and GANs can be used to implement standardized functionality. Whereas this section provides an overview of AI functionality that is being already standardized, generative AI-based implementations of this functionality are discussed in greater detail in Section 4.

- In (RAN), normative (i.e., standardization) activities for AI functionality have started in 3GPP for 5 G-advanced (Release 18) accross all three layers, Layer 1 (L1) to Layer 3 (L3), of the air interface. Specifically, in L1, a study on physical layer enhancements referencing use cases such as compression of Channel State Information (CSI) reporting [27], beam management, and positioning is ongoing, supported by both UE chipset and mobile network vendors [28]. Examining the above use cases, we can identify several areas of interest for generative AI. Transformer or VAE-based architectures, for example, can be used for encoding and decoding CSI information. CSI data can be projected into a size-efficient embedding space. Within this space, relevant information can be correlated using a transformer's attention mechanism. Subsequently, this embedding space can be transmitted to the receiver. There, the complete CSI report can be generated.

In higher layers, network protocols for data collection, model training, and deployment are currently being standardized [29]. Specifically, three use cases have been defined: First, network energy saving, wherein UE and Radio Base Stations (RBSs) exchange energy-saving information. This information is used by energy-saving models to make an energy-related decision, e.g., put a cell[1] to sleep for a given amount of time. Second, load balancing, wherein AI models steer UE-destined or originated data traffic through different cells, in order to evenly allocate the load. Third, mobility optimization, which uses AI models to reduce failed handovers by predicting mobility, location, and performance of UEs.

Complementary to the 3GPP architecture for RAN is Open Radio Access Network (O-RAN), proposed by O-RAN Alliance. From an AI perspective, O-RAN standards enhance 3GPP standards by introducing Radio Intelligent Controller (RIC), a component that controls and optimizes RAN functions. A RIC consists of a near-Real-Time (RT) and a non-RT component. The latter is a centralized function, which, through applications known as "rApps," provides non real-time control of RAN aspects, with decision loops larger than a second. The latter is a distributed function, deployed at network edge, which, through another set of applications known as "xApps," enables near real-time optimization and control of RAN elements, with decision loops between 10 millisecond and a second. Within RIC, O-RAN Alliance, the organization responsible for standardizing O-RAN, has specified an AI architecture, which describes how ML models are being trained and deployed in near-RT and/or non-RT RIC [30].

Additional to the three specified use case by 3GPP above, O-RAN alliance has specified several more use cases, including control of RAN elements for dynamic environments (e.g., radio resource allocation for Unidentified Aerial Vehicle (UAV) type of UE and handover management for Vehicle-To-Everything (V2X) environments, where UE are moving vehicles, dynamic spectrum sharing, etc.) [31]. Dynamic environments are of interest to generative AI approaches. As discussed in Section 1, adapting baseline models to the local UE traffic and mobility patterns, radio channel interference, etc., will require observing data locally and potentially transferring to the non-RT RIC for training. Generative AI models deployed at the non-RT RIC can help reduce bandwidth requirements for transferring large amounts of data.

- In core network, AI-related standardization activities in 3GPP have been focusing on Network Data Analytics Function (NWDAF), which is a Network Function (NF) that provides analytic services to other NFs. These analytic services enable these NFs to take actions. For example, NWDAF may provide a prediction of when a congestion will occur, to the Policy Control Function (PCF), which in turn will take action to mitigate the possibility of congestion (e.g., by changing policies for at least some of the attached UE). In order to provide this type of service, NWDAF may use ML models, trained at and served by an internal Model Training Logical Function (MTLF) and consumed by another internal Analytics Logical Function (ANLF), which also provides the response by embedding the output of the ML-trained model. Of specific interest to generative AI is a 3GPP technical report, outlining key issues with NWDAF and the suggestions for improvement [32], many of which are already in "normative" phase (i.e., under standardization). Specifically, solutions around improving the correctness of NWDAF analytics may necessitate the use of generative AI-based DTs to evaluate the performance of MTLF-trained models.
- Beyond 3GPP and O-RAN standardization, telecom vendors and operators have shown interest in the idea of using intent-based operations to create autonomous networks. The TeleManagement Forum (TMForum)'s Autonomous Network Project describes a reference architecture for designing and realizing autonomous networks [33]. Part of this architecture are hierarchically organized intelligent agents, which participate in the so-called intent management loop [34]. Given an intent, these agents observe the network, analyze the observations, propose decisions, evaluate these decisions, and once a decision is chosen, execute a series of actions in the network to fulfill the requirements of that intent. Different phases in this loop can be of interest to generative AI. For example, the evaluation phase would need to explore several what-if scenarios in order to find the better performing proposal. Such an evaluation could not be done in a live network and would instead use a virtual environment such as a generative AI-built DT.

As part of Release 19, 3GPP also performed a study of use cases for distributed model training and execution[2] [35]. Specifically, in one set of use cases, the study describes splitting computation between UE and the network infrastructure ("application server") for inferencing of ML models

[1] Note that the terms RBS and cell are used interchangeably in the context of this publication.

[2] In context of this publication, the terms "execution" and "inferencing" are used interchangeably.

Table 1 KPI requirements for work task offloading (adapted from Table 5.1-2 of [35])

Detector	Data size	Data Rate	Upload latency	Image recognition latency
AlexNet (30 frames per second (FPS))	0.15−0.02 Mbyte per frame	4.8–65 Mbps	2ms (remote driving augmented reality (AR) gaming, robotics)	1 s
VGG-16 (30 FPS)	0.1−1.5 Mbyte per frame	24–720 Mbps	10ms (video recognition) 100ms (person identification photo enhancement)	1 s

for visual object detection. The layers of the Artificial Neural Network (ANN), for example, a Convolutional Neural Network (CNN)-based visual object detector, are split, and part of the computation is shared between the UE and is sent to the network infrastructure. The decision on where to transfer the computation is based on the available radio resources on the uplink (UL) interface of the UE running the application and latency of inferencing task. Some of these requirements are illustrated in Table 1.

Generative AI algorithms can play a role in reducing the data size required to be transferred between compute nodes. Specifically, VAEs have proven to create efficient latent space representations, resulting in smaller sizes of data to be exchanged between nodes [36]. Semantic communication, a process where a source content, such as an image, is described, is transmitted in semantic information (e.g., text) to the receiving node, which then uses generative AI to recreate the source content, is also achieved with generative AI [37]. The semantic information being transmitted is comparatively much smaller in size than the actual content, therefore saving bandwidth. In terms of the use cases in the 3GPP report, visual object detectors work by so-called feature extraction, wherein hidden layers extract intermediate features, that can be described using semantic information.

In conclusion, many standardization activities currently exist for AI functionality in RAN and core parts of mobile networks. We see several opportunities for generative AI that can be synopsized into two categories. First is content generation for training ML models to be deployed at the network edge, targeting RAN use cases. Second, safe learning, in use cases involving training ML models, wherein generative AI algorithms can be used for the creation of DTs.

3 Performance and requirements for generative AI

This section discusses methods used to evaluate the performance of generative AI algorithms. As discussed in Section 2 and shown in Fig. 1, generative AI algorithms may have tractable likelihood (e.g., RNNs), meaning that they have knowledge of the underlying probability density function, or intractable likelihood (e.g., VAEs and GANs), meaning that they do not have any knowledge of the probability density function.

In the case of tractable likelihood models, the approach follows the normal evaluation route used for discriminative AI models, where evaluation methods such as accuracy, precision, and recall (in case of classification models) and mean square error and explained variance (in case of regression models) are used.

On the other hand, for intractable generative AI models or for transformers such as LLMs relying on the attention mechanism, it is impossible to estimate the probability density function and compare it to the ground truth. Therefore, other types of evaluation methods need to be used.

3.1 Evaluation approaches for generative AI models

For GAN models, there exist a series of works in the literature for measuring the perceptual quality, diversity of samples, and generalization ability [38].

Specifically, the evaluation problem can be formalized as follows: how to learn the original distribution of a given dataset or domain to generate new samples that are realistic and close enough to the data instance.

Beyond labor-intensive, manual inspection of generated content, two common approaches for computing the quality and diversity of generated samples are Inception Score (IS) [39] and and Fréchet Inception Distance (FID) [40]. Both are limited to the visual domain and make use of a pre-existing classifier trained on the image dataset [41]. The use of a pre-trained classifier, however, also means that both FID and IS are limited by what the pre-trained classifier can detect. Test set log-likelihood is another evaluation method for evaluating GAN models, although it is not directly applicable to models from which we can only draw samples from.

Recent developments in GAN evaluation have improved on the weaknesses of the aforementioned evaluation methods and broadened their applicability beyond the image domain [42, 43]. Some examples include modifications to the original FID method to accelerate the evaluation process [44] and reduce bias [45]. FID also does not distinguish between fidelity and diversity of generated content.[3] To measure fidelity and diversity separately, other methods inspired by and building on the precision-recall metric were introduced. An example of such metric is density and coverage [46].

In addition to the above quantitative metrics, GAN evaluation state of the art also includes qualitative evaluation methods. These methods focus on the perception of the generated content from human audience and therefore include humans-in-the-loop [47].

Transformers are models with encoder-decoder structures that use the attention mechanism. The encoder generates the encoding of inputs, and the decoder generates the output using all the encodings and their incorporated contextual information. In the original transformers, the attention mechanism utilizes softmax to capture the long-range dependencies in the sequence of tokens (for texts) or of patches (for images).

The transformers are widely used in BERT and GPT models. BERT-like models [48] adopt two objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM is a fill-in-the-blank task consisting of predicting tokens of the input sequence that have been masked, whereas NSP is a binary classification task to predict whether two segments are adjacent in the original text. Some models eliminate the NSP task in their pre-training to address its ineffectiveness, such as XLM-R, CamemBERT, and FlauBERT models. Some models (such as FrALBERT) pre-train MLM and Sentence Order Prediction (SOP) to predict whether a sentence order in a given sentence pair is swapped.

Transformer-based models are mostly focused on language understanding at the beginning and, more recently, on image processing; therefore, they have been widely evaluated on text representations for natural language understanding. Some of their important benchmarks are crowdsourced questions derived from Wikipedia articles (SQuAD) [49], multiple-choice reading (RACE) [50], and diverse set of text-based tasks covering paraphrasing, sentiment, and linguistic acceptability (GLUE) [51]. For LLMs in particular, their performance is determined by multiple factors such as the layers specification, the dataset size [52], or the pre-training objectives. These objectives define several language models that can be evaluated by the common or separate methods.

For use cases involving text (e.g., translation, generation, summarization) in a telecommunication context, [53] has proposed several benchmark datasets, including TeleQuAD and mTeleQuAD, to evaluate *telecome question answering* task. The data is mainly collected from 3GPP specifications.

On the other hand, image-based transformers focus on a series of tasks, including image classification, object detection, and multi-modal learning. Different loss functions have been used to train these models considering the data and label types. As an example, cross-entropy [54], distillation loss based on KL-divergence [55], negative log-likelihood (of masked patches), Hungarian loss [56], etc. The same evaluation measures can be used for applications such as channel estimation and passive beamforming since these use cases can be modeled as images or videos [57].

Training and pre-training transformers in a self-supervision manner play a crucial role in their scalability and generalization. In order to take advantage of pre-trained transformers and fine-tune them to unseen data, GPT models have been developed. They contain two main components: pre-training and fine-tuning (it can be improved by the help of human in the loop). Although each component's performance depends on the target task, there exist a series of evaluations to asses them regardless of the selected task. For instance, perplexity, sample generation quality, transfer learning, and diversity and coherence of generated samples are important for evaluating the pre-traning part. On the other hand, cross-validation, human evaluation, error analysis, and generalization are the list of important parameters for fine-tuning a GPT model [14].

The VAE is a balance between two components: the log-likelihood that improves the reconstruction quality in the latent space and the Kullback-Liebler component that acts as a regulizer pushing the inference distribution towards the desired distribution. Log-likelihood and KL-divergence can be frequently balanced using a regulator parameter as a normalizing factor for the reconstruction component. Preferring the reconstruction component improves the quality of the reconstruction regardless of the generation effects and the latent space shape, while preferring the KL-divergence improves the smoothness and normalization of the latent space with more detached features

4 Applications of generative AI in mobile networks

In this section, we describe existing approaches of applying generative AI algorithms to mobile networks. Figure 3 illustrates a general categorization of such approaches based on their functionality and part of the mobile network they are applied in. In the context of this publication, we consider a mobile network to include a network management and exposure layer. The former includes all the necessary functions to operate the network throughout its lifetime, while the latter exposes functionality of the network to applications used by mobile subscribers such as private individuals and enterprises. In addition, we consider UEs to be part of the mobile network as well, as 3GPP increasingly recognizes their role

[3] Fidelity refers to the degree to which the generated content matches the real content. Diversity measures whether the generated content covers the full variability of the real content.

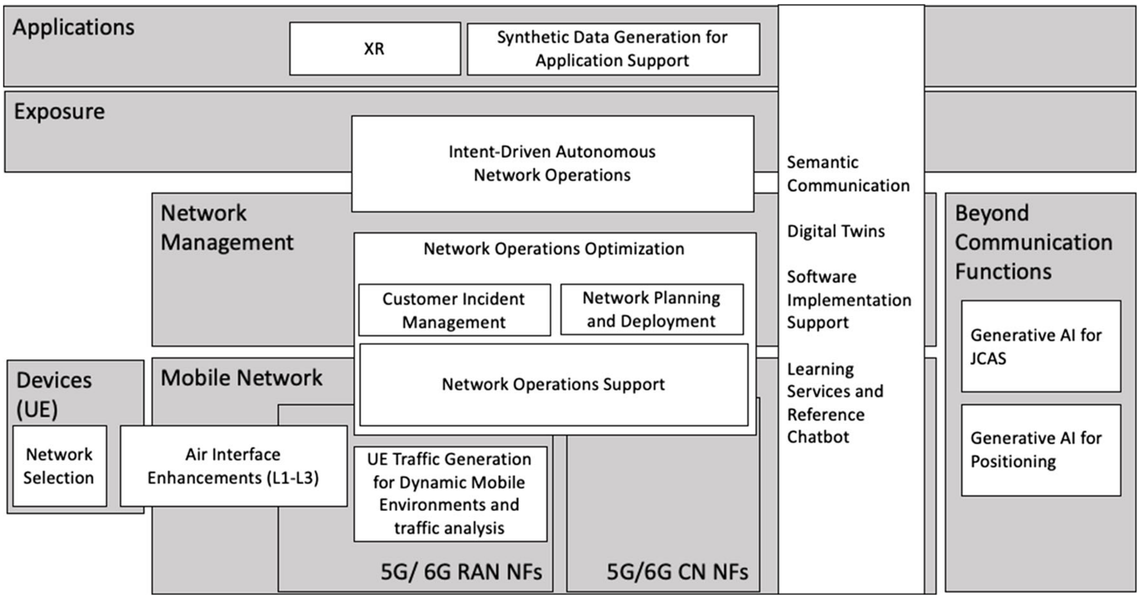
Applications
XR
Synthetic Data Generation for Application Support
Exposure
Intent-Driven Autonomous Network Operations
Semantic Communication
Digital Twins
Software Implementation Support
Learning Services and Reference Chatbot
Network Management
Network Operations Optimization
Customer Incident Management
Network Planning and Deployment
Beyond Communication Functions
Generative AI for JCAS
Generative AI for Positioning
Devices (UE)
Network Selection
Mobile Network
Air Interface Enhancements (L1-L3)
Network Operations Support
UE Traffic Generation for Dynamic Mobile Environments and traffic analysis
5G/ 6G RAN NFs
5G/6G CN NFs

Fig. 3 Taxonomy of generative AI algorithms for 5 G and envisioned 6 G mobile networks

as compute nodes in execution of distributed AI algorithms (see Section 2.3).

4.1 Generative AI in RAN

In this subsection, we present approaches of generative AI algorithms aiming to improve several aspects of RAN.

We start with air interface enhancements, as illustrated also in Fig. 3. As “air interface,” we define the network between the UE and RBSs in a mobile communications network. It is arguably the most valuable resource, as it is finite, but its use is also very unpredictable as it depends on factors external to the mobile network (e.g., UE traffic and interference from the environment). Starting from the physical (PHY) layer, also known as L1, there are several problems which are addressed using generative AI:

- *Wireless channel modeling*: A first set of approaches aims to learn implicit probability distributions of multiple-input and multiple-output (MIMO) channels. Learning these models is important as they are widely used for benchmarking and creation of simulation scenarios. Rather than gathering measurements of the wireless channel and fitting them to a known model such as Tapped Delay Line (TDL) or Clustered Delay Line (CDL), a number of research works use a GAN to learn the channel distribution, resulting in a more accurate channel model [58–60].
- *Spectrum sensing* is the process where a radio monitors a frequency at certain bandwidth to gather information about spectrum availability. Finding or even predicting unused spectrum dynamically is important as it can be reallocated for use in other communication tasks. Deep learning has already been applied to prediction of free spectrum by use of CNNs; however, collecting a training dataset for all spectrum conditions is expensive, while the network requires retraining should the underlying spectrum environment changes. Instead, authors in [61] develop a GAN generating synthetic data that can be used later to create a classifier that predicts available spectrum. In [62], authors use an Enhanced Capsule Generation Adversarial Network (ECGAN) that itself uses a VAE to estimate the wireless channel occupancy.
- *Channel quality estimation* is the process of determining the state of a wireless channel in a mobile network. This is typically achieved by means of RBSs, either periodically or on event basis, sending requests for transmission of reference signals to attached UE. Subsequently, this information is used by the RBS to determine correct modulation and code rate, select the correct beam in a beamforming scenario, etc. The information reported by the UE is known as Channel State Information (CSI). CSI is represented as a high-dimensional matrix which is large, resulting in high computational and storage requirements. In [63], the authors propose a transformer-based architecture that by learning correlations between subcarriers using the attention mechanism compresses an original CSI to an efficient representation that can be reconstructed at the destination. The computational requirements are lower than other deep learning-based solutions, while the size of the latent space is kept compact. In another approach, authors use VAEs to estimate channel state in massive MIMO scenarios [64]. The model is found to perform well not only in optimal

conditions, but also in the presence of additive white Gaussian noise.

- *Hybrid beamforming (HBF)* is a solution used in 5 G and expected to be used in 6 G that uses digital precoding with analog beamforming, in order to support throughput-demanding, multi-user environments. HBF optimization is a challenge, as the process is non-convex and highly complex on a high-dimensional space that require a large number of iterations in order to find the optimal precoders needed for data communication [65]. In addition, optimization methods require near-perfect CSI feedback that may be difficult to obtain in practice. In [66], authors suggest use of a GAN to create a low-dimensional representation of the high-dimensional search space, then perform a search for the optimum precoders in this low-dimensional space. By incorporating domain knowledge of each iteration in the neural network, the authors show that the complexity of HBF optimization is decreased, while at the same time decreasing CSI feedback overhead. In [63], authors propose using a tranformer model that outperforms existing or CNN-based methods suggested in the literature in terms of CSI feedback overhead and spectral efficiency. Spectral efficiency is expressed in the paper as sum-rate of concurrent transmissions.
- *Network traffic generation:* In addition to the above, as mentioned in Section 2.3, O-RAN Alliance has also specified a number of use cases that relate to use of ML models in highly dynamic scenarios such as V2X and UAV communication. In many cases, training such models may require large amounts of data, which are expensive to observe and transport at the point of model training. Therefore, network traffic data generation is one field that is of interest, as it can contribute to reducing the bandwidth and computational requirements for discriminative ML model training. In [67], the authors leverage several GAN variants to generate realistic network traffic data, which match the original data traffic distributions. In some cases, especially in RAN, some ML models may require data with such fine granularity that the observation framework in place may be unable to provide. For example, algorithms for Radio Resource Management (RRM) may require data gathered in millisecond granularity (e.g., 1 or 2 ms), whereas the observation framework collects aggregated data measurements every minute. In [68], authors combine a Zipper Network (ZipNet) with a GAN to propose a solution that generates network traffic data on a granularity that is two orders of magnitude higher than normal probing, while maintaining remarkable accuracy. Network traffic analysis is another aspect that is useful in many use cases.
- *Network traffic analysis and anomaly detection* is also an area of interest as detecting or predicting anomalies may trigger corrective action from the mobile network. In [69], the authors propose the use of a VAE together with unsupervised deep learning for anomaly detection. Contrary to other approaches, the solution proposed in this paper does not require large amounts of labeled data for training, thus making it likely to detect “zero-day” attacks. In addition, VAE creates a comparatively small latent space representation and therefore allows scalability of the solution both in terms of data size as well as number of features that can be considered.
- *Network selection* may also be a future application areas for UE with multi-connectivity capabilities. In [70], authors propose a GAN-based load estimation algorithm, which provides input to a network selection algorithm, that is used by UEs to select networks with lighter loads.

4.2 Network management

In this section, we present existing approaches of generative AI that target network management, including network planning, deployment, and operation. Some of the works presented in this section are not exclusively limited to 3GPP mobile networks and are applicable to other network types.

4.2.1 Customer incident management

In this section, we refer to previous approaches to solutions contributing to automation of customer incident management, from detection of the incident, to drafting of customer support requests and trouble reports, to resolution.

- *Incident detection* refers to methods used to detect anomalous behavior using as source machine readable data such as logs (e.g., alarms, program traces). In [71], the authors transform raw logs into multi-dimensional numerical vectors with the help of transformer models. Next, a RNN-based LSTM parses those embeddings in sequence of arrival in order to identify anomalous behavior that could be considered as an incident. Results show robustness of approach to alteration of log content (e.g., due to a software upgrade), but also good accuracy of detecting anomalous behavior. In [72], the authors propose LogFit, a BERT-based LLM, finetuned on log data patterns, and used for log anomaly detection. Results show that the model is also robust to vocabulary changes in the logs and can also outperform legacy methods for log anomaly detection.
- *Trouble report drafting* refers to methods used to automatically create trouble reports on detection of an incident. One aspect of this drafting is the ability to parse logs, i.e., to convert them from raw-format to a more compact, machine readable format, either to be included in the trouble report as is or to be used as prompt to another model for summarization. In [73], authors present LogPPT, a

LLM that uses a novel prompt tuning method to recognize keywords and parameters from a few labelled log data selected by an adaptive random sampling algorithm.
- *Trouble report classification* concerns automatically classifying customer incidents specified in trouble reports or customer support requests in order of severity, so the most critical incidents are addressed first. In [74], the authors train a number of classifiers including deep learning models but also BERT-type transformers. The results showed that the transformer-based models outperformed all other approaches.
- Given an existing customer support request, *Trouble Report similarity search* concerns finding similar previous customer support tickets by comparing ticket contents. Similar tickets can be solved in similar ways or can be dealt by the same support team. In [75], the author used BERT-type transformers to create embeddings of trouble reports, then using clustering techniques for grouping trouble reports with similar embeddings in clusters, and finally labeling the clusters using an extractive summarization model.

4.2.2 Network planning and deployment

This section describes generative AI approaches to facilitate network planning, but also deployment and configuration of network nodes once planning is complete. Specifically, we distinguish the following areas:

- *Radio map estimation*: A radio map is an important tool for network planning, as it provides information about the spatial distribution of signal strength and can therefore identify network blind spots or areas that have weak coverage. Creation of radio maps is a challenging task, as radio measurements are sparse due to environment and UE availability. In [76], the authors describe a GAN-based approach for estimating radio maps based on sparse measurements. The results show that this approach has better performance than state-of-the-art methods, partially using non-variational autoencoders. In [77], authors describe RadioNet, a transformer-based solution for radio map estimation, that not only reduces validation loss and increases prediction reliability when compared to the state of the art, but also increases prediction speed by 4 orders of magnitude.
- *Cell load and traffic estimation* is another area where generative AI algorithms are used, in the absence of a UE-dependent, rich measurement set. In [78], the authors use GANs to produce realistic samples of user spatial distribution in various scenarios, based on the day, week, and time of day. This information is then provided in a simplified form to a control plane of Software Defined Network (SDN) controllers that take actions on the mobile network (e.g., traffic routing and load balancing). Results show that with this approach improves network coverage and spectrum utilization over the state of the art.
- *Configuration of network elements* is another area where generative AI algorithms can be of good use. Specifically, many network elements, either virtual (e.g., NFs, SDN controllers, virtual routers) or physical (e.g., baseband boards, radios), may have a number of configuration parameters that need to be adjusted manually prior to deployment of this equipment on a production environment. However, this kind of parameter tuning also requires a deep knowledge of the system that may be beyond the abilities of the average user. In pioneering work of [79], authors introduce “ACTGAN,” an algorithm that leverages a GAN that generates configurations by learning and utilizing structures of existing good configurations. The authors test ACTGAN on a number of open-source software including databases, computational frameworks, and event streaming platforms to show that the generated configuration outperforms default configurations and state-of-the-art configuration algorithms. In another approach, authors make use of a Conditional Variational AutoEncoder (CVAE) to propose improved network configurations for sustaining end-user quality of experience (QoE) in a video streaming scenario [80].

4.2.3 Network operations support

In this subsection, we describe several areas of network operations where generative AI algorithms can be of service.

- *Fault diagnosis*. In 3GPP Release 10, Model Drive Test (MDT) functionality was introduced, which allows operators to use UE to collect mobile network data, thus reducing need for drive testing. However, in practice, MDT reports are sparse as they depend on UE availability. In [81], the authors counter the spase availability of MDT reports by using an image translation GAN (also known as “Pix2Pix”) to generate MDT coverage maps from sparsely available data, which are subsequently used as input to a CNN classifier that provides a potential fault diagnosis.
- *Resource allocation for network slicing* is another area where generative AI algorithms can be of assistance. In [82], the authors use double deep-Q learning RL algorithm to train a GAN to allocate certain bandwidth to each network slice, given the number of data packets arriving for every network slice. The authors show that the hybrid RL-GAN solution they call “GAN-DDQN” outperformes vanilla DQN. In [83], the authors suggest using a GAN for forecasting resource utilization.

- *Network security* is an important area of network management that needs to be part of everyday network operations. There are several active resarch areas within network security for generative AI.
  - One body of work considers the problem of *malware detection*. Malware is software installed on a computing device without the consent of the owner and may perform various malicious actions. State-of-the-art, deep learning-based malware detection models are trained to detect malware. Generative AI approaches challenge the detection capability of these models by using GANs to generate examples that can be used with purpose to avoid detection [84, 85]. Generated malware examples can be subsequently used to train new malware detectors.
  - Another body of work considers the problem of *rogue device detection*, wherein identifying and localising malicious devices that might threaten the security of the mobile network or its users. In [86] and [87], the authors use a GAN to detect rogue radio frequency (RF) transmitters. Detection of rogue devices could be especially valuable in multi-access, heterogeneous 6 G networks, or in Internet of Things (IoT) type of scenarios where not all terminal devices beyond the IoT gateway are authenticated to the mobile network.

4.3 Requirements engineering

In this section, we describe the set of approaches that given requirements from the exposure layer generate network-specific configurations to fulfill these requirements. As discussed in Section 2.3, TMForum-led intent-driven autonomous network management is a stepping stone of network evolution towards a fully autonomous paradigm. As described in [88], intent is an abstract notion of a high-level policy that is interpreted in greater detail in lower layers. This is a process also known as requirements extraction and is a natural place for generative AI algorithms that can interpret natural language models and transform them to machine readable specifications.

In [89], the authors describe a GPT-finetuned transformer model for extracting use cases, actors, and their relationships from a specification document in natural language. In [90], the authors use a GPT-J transformer to extract technical requirements from natural language sources.

In the context of a mobile network, an intent can be considered as requirements from mobile subscribers (e.g., UE applications) about quality of service (QoS), billing plan, feature list, etc. Oftentimes, these requirements can be expressed in natural language or in an abstract format that can be translated to technical specifications. These technical specifications can be further translated into actionable network configuration instructions using a GAN such as the one mentioned in Section 4.2.2. The potential applications are presented in greater detail in Section 5 as future research direction. To the best of our knowledge, no work exists yet in context of telecommunication networks.

4.4 Applications

In this section, we cover the use of generative AI algorithms for potential future applications. The generative capabilities of these algorithms may allow for reduced bandwidth demands and lower communication latency. The reader should note that the use cases presented in this section are non-exhaustive, but may function as an inspiration for using generative AI in similar use cases in the future.

4.4.1 Optimization of (XR) network datastreams

XR, including AR and Virtual Reality (VR), is one of the applications expected to drive demand for next-generation networks. In [91], authors describe the use of GANs to generate content for XR applications, based on observed patterns of everyday use. Such content generation algorithms can be placed close to network edge and therefore reduce the bandwidth requirements and communication latency. In another example, generative AI algorithms can be used to also compress XR datastreams. In [92], authors use a GAN for compressing high-quality video, outperforming previous methods.

4.4.2 Data generation for industry applications

- *Generative AI for robotic surgeries and remote diagnosis:* Improved connectivity using 5 G has opened doors for robotic surgery and remote diagnosis. This makes it possible to conduct a preplanned surgery by an expert physician for a patient in another part of the world. Today, most of such procedures are conducted on an experimental basis [93] with low latency [94] visual and robotic control. However, to replicate the surgeon's physical proximity, such a surgery requires real-time transfer of at least four senses—haptic, tactile, audio, and visual signals—with low latency and high levels of assurance. For the same, sophisticated control and precise coordination across all domains of the network are needed, which today is a challenge to achieve in a commercial rollout. However, with generative AI, information exchange of all 5 senses may become feasible as the data produced by edge devices may be recreated at the user's (doctor's) end. This would enable remote surgeries and interventional medicine to be carried out in scale. In one approach, multiple streams of data (e.g., from all 5 senses) can be multiplexed and represented efficiently in latent space using a transformer, later to be decoded on the user end.

- *Vehicle teleoperation* is another application, where generative AI can be used. In this application, a vehicle is driven by a remote operator, using mobile network communication. During teleoperation, there can be situations where, due to poor network performance, video feed coming from the vehicle and control feedback sent back to the vehicle is poor (e.g., due to coverage, load, or environmental reasons). In [95], for example, the authors presented real-world measurements in a semi-urban environment, which revealed network blind spots, where teleoperation would not have been possible. In such situations, teleoperation data can be compressed on the originating end and regenerated on the receiving end.

4.5 Beyond communication functions

4.5.1 Generative AI for JCAS

Joint communication and sensing (JCAS) is a new technology expected to be commercialized in the 6 G time-frame [96, 97]. In this technology, the network (e.g., gNodeB (gNB)) is capable of providing a map of the surrounding sensed objects using radio frequency beams. This could help in many applications, e.g., autonomous driving and channel modeling. Generation of radio rays in an efficient manner is key to the success of JCAS technology. Generative AI could help in generating relevant rays to capture and sense the surrounding. For instance, this could be done by generating the right distribution of Azimuth and Elevation Angles of the radio frequency beam transmitted out of the node (gNB or UE), see, for example, a GAN-based approach at [98].

4.5.2 Generative AI for positioning and localization

Positioning and localization are functions used by many industry vertical use cases, for example, asset tracking type of use cases in logistics and autonomous navigation in robotics, such as UAV and Automated Guided Vehicle (AGV) type of devices.

Solutions use Received Signal Strength Indicator (RSSI) and CSI UE measurements, also known as *fingerprints*, and apply further processing to estimate the position of a UE. In [99], authors propose the use of a GAN to correlate the signal strength of a device, measured as RSSI, and the position of the device in space. Given an RSSI measurement from the UE, a gNB can localize this UE in space. In [100], authors use a Conditional Generative Adversarial Network with Auxiliary Classifier, or "AC-GAN," to generate realistic fingerprint measurements to augment real measurements. The latter are expensive to obtain—thus, the GAN presented saves time, bandwidth, and computational resources.

4.6 Overarching aspects

4.6.1 Semantic communication

Semantic communication is the process of encoding information at the sender in a compact representation that sufficiently describes this information. Subsequently, this compact representation is sent from the sender to a receiver in place of the raw information itself. Based on the compact representation, the receiving end can apply an algorithm to decode the information that should be identical as to the raw information encoded at the sender. The advantage of using this type of communication is that it has the potential to save a significant amount of bandwidth. Especially considering the "air interface" between UE and gNB, these bandwidth savings are very important as they free bandwidth for use for other type of communication. We have already mentioned in Section 2.3 that 3GPP is standardizing functionality on CSI compression, but also in this section of several places in the network that can benefit from compressed content.

In the literature, DeepSC, a transformer-based system designed for semantic communication, outperforms other encoding and decoding methods [63, 101]. In addition, special purpose-built transformers, variants of DeepSc, are presented for different data modalities, such as speech [102] and images [103]. Given the volume and recency of published works, we conclude that semantic communication is an open research area in context of mobile networks.

4.6.2 Generative AI for digital twining

One application of generative AI is creating or assisting in creating DTs. A DT is a virtual representation of a physical object, process, or system. It is created using digital data and simulation software to model the behavior, performance, and characteristics of the physical object or system.

In the context of this work, a DT is considered to virtually represent any network entity, from a NF in RAN or core network, to a gNB, to a complete RAN or core network. In the most complex case, an end-to-end DT may represent the totality of the mobile network, including transport and infrastructure. On the other end of the scale, DT may represent one or more wireless protocols from a protocol stack.

In terms of usage, a DT can be utilized to analyze, optimize, and monitor, e.g., the factory radio environment, system key performance indicators (KPIs), and its life cycle, from design and development to operation and maintenance. An accurate abstraction of DT can enable improved network performance, better optimization, and efficient management. For instance, a DT of gNodeBs (gNBs) and antennas can help in the coverage, capacity, and management of networks. Also, DTs can provide enhanced observability and visibility of wireless network KPIs via emulating the behavior of different basic functions of the network. Such functionality allows for a major role of DT in 5 G and 6 G realization. For instance, it allows for a faster time of deploying new features to market via speeding up the validation process. It also allows for more efficient risk discovery and management via realizing risky actions or designs.

In [104], authors suggest the suitability of conditional GANs (cGANs) as digital twins, as they are able to learn the distribution or distribution of quantity or quantities of interest. With generative AI, GANs are a black box, meaning that instead of an approach where the behavior of a DT is *programmed*, the approach here is that instead the behavior is *learned*, and new behavior can be generated based on what was learned that is consistent with reality. In [105], authors present a network digital twin that uses GAN as a data generator for other network functions to use. In [106], authors present a GAN that reduces the training time of 5 G DTs, by learning the tail behavior of the train dataset distribution (i.e., all rare events), with only a few samples.

4.6.3 Software development, learning services, and reference agents

OpenAI's release of ChatGPT, a finetuning of their own GPT model, has revolutionized the way humans interact. In addition, the process of finetuning GPT and ChatGPT has yielded applicability of expert systems in different domains, with relatively little effort. Such expert systems provide a dialogic interface in which users can ask single questions or chains of questions, and the expert system provides responses.

Transformer models can be used as assistants to humans in different parts of the network operations. First, transformers can generate code and can, together with their human counterparts, make software development more efficient. One part of transformers works with code generation and code completion [107, 108].

Second, transformers can function as reference and learning agents. In this case, transformers can provide text generation, translation, and summarization services in a number of subjects, ranging from product documentation for troubleshooting or upgrades or installation to 3gpp standards. For this type of use cases, [53] has proposed several benchmark datasets, including TeleQuAD and mTeleQuAD. The data is mainly collected from 3GPP specifications.

5 Research directions

This section discusses open problems and research trends for generative AI. The aim is to provide the reader with a general impression on where research efforts on application of generative AI in mobile networks are focused on. We present each general challenge in the relative subsection.

5.1 Performance versus resource requirements

As discussed in Section 4, several approaches using generative AI networks were found to perform better than the state of the art. At the same time though, these approaches come at a computational cost that is often higher than the methods used before. For example, looking at semantic communication, the computational cost of encoding and decoding the information, in addition to the storage cost of keeping the encoder/decoder in memory, may be significant. Resource cost especially impacts RAN applications, where resource-constrained UE may be involved. Finding ways to compress the generative AI algoritms as in [109] or distribute the computation would be critical for mass adoption of generative AI models especially at the network edge.

5.2 Enhancing network exposure functions

Network exposure is the function of the network that exposes network information to third parties. Currently, 3GPP-standardized nodes, such as Network Exposure Function (NEF) and Service Capability Exposure Function (SCEF), expose certain limited information in the form of APIs. In the future, this interface will be bidirectional, where third party requests can be sent to the network and information on these requests can be pushed from the network back to the third party. Also, information can be pushed by the network to the third party without the latter's request.

The capability of transformer networks in particular to transform natural language prompts to other representations would be useful for mobile networks to acquire intent-driven autonomous network functionality. Specifically, the interpretation of intents expressed in natural language to generate network QoS policies, billing plans, subscription plans, etc. GANs can also be trained to generate customer offers using network consumption data (e.g., suggestion of new billing plan).

5.3 Generalization of RAN optimization techniques

Radio environments are not stationary and change often with mobility and user behavior. Time series prediction or anomaly detection are quite relevant use cases, and here, the current data is often correlated with previous time steps. GANs have proved to be effective in capturing temporal correlations [110] and thereby identifying anomalies in such time series data. However, as the environment changes often, existing solutions suffer from poor generalization abilities. Recent advances in Generative AI [111] make it possible to develop algorithms which generalize to such changes in time series distributions.

Similarly, for RL policies, any change in the dynamics of the system renders the policy ineffective in many regions of the state space. Using Generative AI, we can build correspondence models [112] such that a cross-domain alignment can be created between the source and unseen environment. Such correspondences, once established, can be effective in addressing generalization of RL-based RAN KPI optimization problems specially as environment dynamics changes [113].

Another example of generative AI is to incorporate generalization abilities on pre-existing models through a hybrid approach. Such an approach has some similarities with mixture of expert models, but in such cases, the existing models may be specific in nature and introduce the necessary inductive bias into the inference process. The attention mechanisms, on the other hand, possess minimal inductive bias but can exceed the performance of traditional supervised neural nets when trained on large data sets. So if the outputs from both the existing model and the attention mechanism can be fed into an feed-forward network (decoder), then the algorithms would generalize better to new contexts as well as be able to reuse the inductive bias from existing models. Such a hybrid mechanism [114] may also benefit from a Gated-Positional Self-Attention mechanism which helps it to decide when to leverage the inductive bias. In the context of telecom applications, the industry possesses a plethora of pre-existing models which have been trained on moderate data volume but contains the relevant domain knowledge in the form of predictive inferences for specific problems. Today, with large volumes of data available in cloud storage, it is possible and also important to adapt such algorithms to a wider variety of situations without losing the necessary context or the inductive bias. Also, for use cases like KPI optimization or predictions in the core, Layer 2 (L2) or L3 layers (defined in [115]), available models can be combined with attention mechanisms that can ingest even newer sources of data and, in conjunction, can provide more generalized solutions.

5.4 Digital twinning

In this section, we focus on the potential of applying generative AI within Digital Twin for Protocol and Connectivity (DT-PC).

One benefit of such application would be to simplify high-complexity optimization problems, such as modeling the interplay between Operation, Administration, and Maintenance (OAM) and RAN. In this case, validation of new OAM functionalities would require “realistic” and “relevant” abstraction of RAN functionalities via generative AI based on DTing of such functions. Realistic abstraction means that the accuracy of the DTPC (constructed by generative AI) is high and close to the performance of the actual product (e.g., RAN) functions. Relevant abstraction means that the RAN-DT-PC that interfaces with OAM represents a relevant function of RAN to the targeted evaluation and/or validation of OAM.

The above proposal could be realized by dividing the main representation of RAN into simple-blocks representation using generative AI to realize higher accuracy of the generative AI sub-blocks of RAN-DT-PC. Such simple-blocks representation of the system could be, for instance, the system’s traffic model, arrival and departure of UEs, radio small or large scale channels, or deployment models.

In this context, another application is enabling an interactive dataset, where the data can be a dynamic organism. Such an application could be used to realize offline RL, where the vector of “state, action, reward, and next-state” could be generated via the generative AI model [116]. This data could then be used to train the model offline devoid of challenges for online exploration.

DTs created by generative AI techniques also enable safe exploration for control-based algorithms. A huge variety of use cases in the radio domain which includes L2 and L3 applications, Core and Orchestration functions could benefit from autonomous control. This is especially true in intent-based Future Networks like 6 G, where many actions, either in software or hardware, may need continuous tuning. However, RL-based control needs exploration in real networks to learn optimal actions. This is a challenge, and service providers are understandably wary given that this may lead to a negative customer experience. Given this constraint, which is typical of Telecom environments, a DT could enable close-to-optimal training of the policy whereby the policy is trained in the DT and quickly fine-tuned through Domain Adaptation/Sim2Real techniques once deployed to production. Additionally, it is seen in [117] that offline RL may be benefited from using generative models, which approximate a policy using fixed data, and Action-conditioned Q-Learning (AQL). The proposed method employs a residual generative model to reduce policy approximation error for offline RL.

Another advantage of such a DT would be prediction-based systems. Often in networks, the model of the system is too complex, and hence, it requires a large amount of data to train a prediction model. Hence, even in supervised settings, it is a challenge to get enough of labeled data for all states of the network. A DT could either simulate the environment in totality or could generate a realistic distribution of data across all states of the network. Such a distribution could enable supervised algorithms to play a bigger role in optimization problems.

6 Conclusion

5 G and 6 G mobile networks are expected to heavily rely on AI- and ML-trained models for their operation. However, there still exist challenges when widely deploying ML models. In summary, such challenges include training data observability, efficient and cost-effective data transfer to edge nodes, and safe learning in network environments are key challenges in implementing closed-loop AI-native functions in future networks, which the generative AI techniques may be able to address.

The paper categorizes generative AI algorithms between tractable and intractable. Several evaluation approaches for performance measurement of generative AI algorithms are highlighted.

The paper also presents a taxonomy to illustrate evidence in the existing literature around generative AI techniques across all domains of the network. Generative AI techniques have been explored across RAN L1 to L3 functions, Core network, Network Management and Exposure and Application layers in order to create effective representations, generate configurations, and data distributions by learning complex structures and interactions between entities, create synthetic data for effective classification or clustering in light of data sparsity, provide dimensionality reduction for effective search in high-dimensional space, for anomaly detection and similarity search, and also extracting use cases, actors, and their relationships from a specification document in natural language.

The paper also reviews current and presents potential future mobile network applications from a generative AI lens. Such applications include XR, Robotic Surgeries, Remote Diagnostics, and Vehicle Teleoperation by optimizing network datastreams and generative AI-based semantic communications.

We also highlight beyond communication functions, and in particular how digital twins can be created to provide an effective representation of the real networks using generative AI. Such a digital Twin could facilitate ML model training, data collection, and safe execution.

Finally, we provide some future directions and highlight how model generalization, continuous learning, and reuse of existing ML assets could be a possibility with new state-of-the-art approaches in generative AI.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Morocho Cayamcela ME, Lim W (2018) Artificial intelligence in 5G technology: a survey. In: 2018 International Conference on Information and Communication Technology Convergence (ICTC), pp 860–865. https://doi.org/10.1109/ICTC.2018.8539642
2. Tataria H, Shafi M, Molisch AF, Dohler M, Sjöland H, Tufvesson F (2021) 6G wireless systems: vision, requirements, challenges, insights, and opportunities. Proceedings of the IEEE 109(7):1166–1199. https://doi.org/10.1109/JPROC.2021.3061701
3. Hinton GE, Sejnowski T (1986) Learning and relearning in Boltzmann machines. Parallel distributed processing: explorations in the microstructure of cognition. MIT Press, Cambridge, MA, pp 282–317
4. Salakhutdinov R, Mnih A, Hinton G (2007) Restricted Boltzmann machines for collaborative filtering. In: Proceedings of the 24th International Conference on Machine Learning. ICML '07, pp 791–798. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1273496.1273596
5. Hinton GE (2009) Deep belief networks. Scholarpedia 4(5): 5947
6. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ (eds.) Advances in Neural Information Processing Systems, vol 27. https://proceedings.neurips.cc/paperfiles/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
7. Gui J, Sun Z, Wen Y, Tao D, Ye J (2023) A review on generative adversarial networks: algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineering 35(4):3313–3332. https://doi.org/10.1109/TKDE.2021.3130191
8. Kingma DP, Welling M (2019) An introduction to variational autoencoders. Found. Trends Mach. Learn. 12(4):307–392. https://doi.org/10.1561/2200000056
9. Vaithilingam P, Zhang T, Glassman EL (2022) Expectation vs. experience: evaluating the usability of code generation tools powered by large language models. In: Chi Conference on Human Factors in Computing Systems Extended Abstracts, pp 1–7
10. Iqbal T, Qureshi S (2022) The survey: text generation models in deep learning. Journal of King Saud University - Computer and Information Sciences 34(6,Part A), 2515–2528. https://doi.org/10.1016/j.jksuci.2020.04.001
11. Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 6(2): 107–116. https://doi.org/10.1142/S0218488598000094
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput. 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp 6000–6010. Curran Associates Inc., Red Hook, NY, USA
14. Radford A, Narasimhan K, Salimans T, Sutskever I, et al (2018) Improving language understanding by generative pre-training. Preprint, published by OpenAI
15. Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (Eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp 4171–4186. https://doi.org/10.18653/v1/n19-1423

16. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A et al (2022) Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35:27730–27744
17. de Rosa GH, Papa JP (2021) A survey on text generation using generative adversarial networks. Pattern Recognition 119:108098. https://doi.org/10.1016/j.patcog.2021.108098
18. Junyi L, Tang T, Zhao W, Wen J-R (2021) Pretrained language model for text generation: a survey, pp 4492–4499. https://doi.org/10.24963/ijcai.2021/612
19. Chui M, Roberts R, Yee L (2022) Generative AI is here: how tools like ChatGPT could change your business. Quantum Black AI by McKinsey
20. Vaithilingam P, Zhang T, Glassman EL (2022) Expectation vs. experience: evaluating the usability of code generation tools powered by large language models. In: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. CHI EA'22. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3491101.3519665
21. Lee H, Ullah U, Lee J-S, Jeong B, Choi H-C (2021) A brief survey of text driven image generation and manipulation. In: 2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), pp 1–4. https://doi.org/10.1109/ICCE-Asia53811.2021.9641929
22. Li S, Tao Z, Li K, Fu Y (2019) Visual to text: survey of image and video captioning. IEEE Transactions on Emerging Topics in Computational Intelligence 3(4):297–312. https://doi.org/10.1109/TETCI.2019.2892755
23. Zhang C, Zhang C, Zheng S, Zhang M, Qamar M, Bae S-H, Kweon IS (2023) A survey on audio diffusion models: text to speech synthesis and enhancement in generative AI. arXiv. arXiv:2303.13336
24. Hernandez-Olivan C, Beltrán JR (2023) In: Biswas A, Wennekes E, Wieczorkowska A, Laskar RH (Eds.) Music composition with deep learning: a review, pp 25–50. Springer, Cham. https://doi.org/10.1007/978-3-031-18444-4_2
25. Aldausari N, Sowmya A, Marcus N, Mohammadi G (2022) Video generative adversarial networks: a review. ACM Comput. Surv. 55(2). https://doi.org/10.1145/3487891
26. Zhang H, Shao S, Tao M, Bi X, Letaief KB (2023) Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data. IEEE Journal on Selected Areas in Communications 41(1):170–185. https://doi.org/10.1109/JSAC.2022.3221991
27. Guo J, Wen C-K, Jin S, Li X (2022) AI for CSI feedback enhancement in 5G advanced. IEEE Wireless Communications 1–8. https://doi.org/10.1109/MWC.010.2200304
28. 3GPP: Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface. Technical Report (TR) 38.843, 3rd Generation Partnership Project (3GPP) (2022). Version 0.1.0
29. 3GPP: study on enhancement for data collection for NR and ENDC. Technical Report (TR) 38.817, 3rd Generation Partnership Project (3GPP) (2022). Version 17.0.0
30. O-RAN Working Group 2: AI/ML workflow description and requirements. Technical report (tr), O-RAN Alliance (2021). Version 1.03
31. O-RAN working group 1: use cases detailed specification. Technical specification, O-RAN Alliance (2023). Version 10.00
32. 3GPP: study of enablers for network automation for the 5G system (5GS); Phase 3. Technical Report (TR) 23.700-81, 3rd Generation Partnership Project (3GPP) (December 2022). Version 18.0.0
33. TMForum: autonomous networks technical architecture. Technical Report (TR) IG1230 (December 2022). Version 1.1.1
34. Niemöller J, Szabö R, Zahemzky A, Roeland D (2022) Creating autonomous networks with intent-based closed loops. Whitepaper, Ericsson
35. 3GPP: study on AI/ML model transfer-phase 2. Technical Report (TR) 38.843, 3rd Generation Partnership Project (3GPP) (March 2023). Version 1.0.0
36. Zhou L, Cai C, Gao Y, Su S, Wu J (2018) Variational autoencoder for low bitrate image compression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops
37. Grassucci E, Barbarossa S, Comminiello D (2023) Generative semantic communication: diffusion models beyond bit recovery. arXiv preprint arXiv:2306.04321
38. Gulrajani I, Raffel C, Metz L (2020) Towards GAN benchmarks which require generalization. arXiv preprint arXiv:2001.03653
39. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. Advances in neural information processing systems 29
40. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in neural information processing systems 30
41. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A largescale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255. IEEE
42. Borji A (2019) Pros and cons of GAN evaluation measures. Computer Vision and Image Understanding 179:41–65
43. Borji A (2022) Pros and cons of GAN evaluation measures: new developments. Computer Vision and Image Understanding 215
44. Mathiasen A, Hvilshoj F (2020) Fast frëchet inception distance. CoRR abs/2009.14075
45. Chong MJ, Forsyth DA (2019) Effectively unbiased FID and inception score and where to find them. CoRR abs/1911.07023
46. Naeem MF, Oh SJ, Uh Y, Choi Y, Yoo J (2020) Reliable fidelity and diversity metrics for generative models. CoRR abs/2002.09797
47. Zhou S, Gordon M, Krishna R, Narcomey A, Fei-Fei LF, Bernstein M (2019) Hype: a benchmark for human eye perceptual evaluation of generative models. In: Wallach H, Larochelle H, Beygelzimer A, Alchë-Buc F, Fox E, Garnett R (Eds.) Advances in Neural Information Processing Systems, vol 32. https://proceedings.neurips.cc/paperfiles/paper/2019/file/65699726a3c601b9f31bf04019c8593c-Paper.pdf
48. Xia P, Wu S, Van Durme B (2020) Which* bert? a survey organizing contextualized encoders. arXiv preprint arXiv:2010.00854
49. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 2383–2392. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1264 . https://aclanthology.org/D16-1264
50. La G, Xie Q, Liu H, Yang Y, Hovy E (2017) RACE: large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 785–794. Association for Computational Linguistics, Copenhagen, Denmark. https://doi.org/10.18653/v1/D17-1082 . https://aclanthology.org/D17-1082
51. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) Glue: a multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461
52. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020) Scaling laws for neural language models. arXiv preprint arXiv:2001.08361

53. Gunnarsson M (2021) Multi-hop neural question answering in the telecom domain. Master's thesis, LTH, Lund University
54. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T. Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
55. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jëgou H (2021) Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp 10347–10357. PMLR
56. Zhu X, Su W, Lu L, Li B, Wang X, Dai J (2020) Deformable detr: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159
57. Soltani M, Pourahmadi V, Mirzaei A, Sheikhzadeh H (2019) Deep learning-based channel estimation. IEEE Communications Letters 23(4):652–655
58. Yang Y, Li Y, Zhang W, Qin F, Zhu P, Wang C-X (2019) Generative-adversarial-network-based wireless channel modeling: challenges and opportunities. IEEE Communications Magazine 57(3):22–27. https://doi.org/10.1109/MCOM.2019.1800635
59. Orekondy T, Behboodi A, Soriaga JB (2022) MIMO-GAN: generative MIMO channel modeling
60. O'Shea TJ, Roy T, West N (2019) Approximating the void: learning stochastic channel models from observation with variational generative adversarial networks. In: 2019 International Conference on Computing, Networking and Communications (ICNC), pp 681–686. https://doi.org/10.1109/ICCNC.2019.8685573
61. Davaslioglu K, Sagduyu YE (2018) Generative adversarial learning for spectrum sensing. In: 2018 IEEE International Conference on Communications (ICC), pp 1–6. https://doi.org/10.1109/ICC.2018.8422223
62. DCM, Reddy, BVR, (2023) Enhanced capsule generative adversarial network for spectrum and energy efficiency of cooperative spectrum prediction framework in cognitive radio network. Transactions on Emerging Telecommunications Technologies 34(4):4736. https://doi.org/10.1002/ett.4736
63. Wang Y, Gao Z, Zheng D, Chen S, Gunduz D, Poor HV (2022) Transformer-empowered 6G intelligent networks: from massive MIMO processing to semantic communication. IEEE Wireless Communications, pp 1–9. https://doi.org/10.1109/MWC.008.2200157
64. Hussien M, Nguyen KK, Cheriet M (2022) PRVNet: a novel partially-regularized variational autoencoders for massive MIMO CSI feedback. In: 2022 IEEE Wireless Communications and Networking Conference (WCNC), pp 2286–2291. https://doi.org/10.1109/WCNC51071.2022.9771642
65. Nguyen NT, Ma M, Shlezinger N, Eldar YC, Swindlehurst AL, Juntti MJ (2023) Deep unfolding hybrid beamforming designs for THz massive MIMO systems. arXiv:2302.12041
66. Balevi E, Andrews JG (2021) Unfolded hybrid beamforming with GAN compressed ultra-low feedback overhead. IEEE Transactions on Wireless Communications 20(12):8381–8392. https://doi.org/10.1109/TWC.2021.3092350
67. Anande TJ, Al-Saadi S, Leeson MS (2023) Generative adversarial networks for network traffic feature generation. International Journal of Computers and Applications 45(4):297–305. https://doi.org/10.1080/1206212X.2023.2191072
68. Zhang C, Ouyang X, Patras P (2017) ZipNet-GAN: inferring fine-grained mobile traffic patterns via a generative adversarial neural network. In: Proceedings of the 13th International Conference on Emerging Networking EXperiments and Technologies. CoNEXT'17, pp 363–375. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3143361.3143393
69. Nguyen QP, Wai L, Divakaran DM, Low K, Chan M (2019) GEE: a gradient based explainable variational autoencoder for network anomaly detection. https://doi.org/10.1109/CNS.2019.8802833
70. Leng C, Yang C, Chen S, Wu Q, Peng Y (2022) GAN for load estimation and traffic-aware network selection for 5G terminals. IEEE Internet of Things Journal 9(17):16353–16362. https://doi.org/10.1109/JIOT.2022.3152729
71. Ott H, Bogatinovski J, Acker A, Nedelkoski S, Kao O (2021) Robust and transferable anomaly detection in log data using pre-trained language models. In: 2021 IEEE/ACM International Workshop on Cloud Intelligence (CloudIntelligence), pp 19–24. https://doi.org/10.1109/CloudIntelligence52565.2021.00013
72. Almodovar C, Sabrina F, Karimi S, Azad S (2022) Can language models help in system security? Investigating log anomaly detection using BERT. In: Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association, pp 139–147. Australasian Language Technology Association, Adelaide, Australia. https://aclanthology.org/2022.alta-1.19
73. Le V-H, Zhang H (2023) Log parsing with prompt-based few-shot learning. arXiv preprint arXiv:2302.07435
74. Ahmed S, Singh M, Doherty B, Ramlan E, Harkin K, Bucholc M, Coyle D (2023) An empirical analysis of state-of-art classification models in an it incident severity prediction framework. Applied Sciences 13(6). https://doi.org/10.3390/app13063843
75. Alexander B (2022) Automated trouble report labeling in the Telecom industry. Thesis at Uppsala University, Department of Information Technology, ISSN 1401–5749
76. Zhang S, Wijesinghe A, Ding Z (2023) RME-GAN: a learning framework for radio map estimation based on conditional generative adversarial network. IEEE Internet of Things Journal 1–1. https://doi.org/10.1109/JIOT.2023.3278235
77. Tian Y, Yuan S, Chen W, Liu N (2021) Transformer based radio map prediction model for dense urban environments. In: 2021 13th International Symposium on Antennas, Propagation and EM Theory (ISAPE), vol 1, pp 1–3. https://doi.org/10.1109/ISAPE54070.2021.9753644
78. Maksymyuk T, Gazda J, Luntovskyy A, Klymash M (2018) Artificial intelligence based 5G coverage design and optimization using deep generative adversarial neural networks. In: 2018 International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), pp 1–4. https://doi.org/10.1109/UkrMiCo43733.2018.9047611
79. Bao L, Liu X, Wang F, Fang B (2019) ACTGAN: automatic configuration tuning for software systems with generative adversarial networks. In: 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp 465–476. https://doi.org/10.1109/ASE.2019.00051
80. Ickin S (2021) Recommending changes on QoE factors with conditional variational autoencoder. In: Proceedings of the 4th FlexNets Workshop on Flexible Networks Artificial Intelligence Supported Network Flexibility and Agility. FlexNets'21, pp 20–25. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3472735.3473387
81. Rizwan A, Abu-Dayya A, Filali F, Imran A (2022) Addressing data sparsity with GANs for multi-fault diagnosing in emerging cellular networks. In: 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp 318–323. https://doi.org/10.1109/ICAIIC54071.2022.9722696
82. Hua Y, Li R, Zhao Z, Zhang H, Chen X (2019) GAN-based deep distributional reinforcement learning for resource management in network slicing. In: 2019 IEEE Global Communications Conference (GLOBECOM), pp 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9014217
83. Abbas K, Afaq M, Ahmed Khan T, Rafiq A, Song W-C (2020) Slicing the core network and radio access network domains

through intent-based networking for 5G networks. Electronics 9(10). https://doi.org/10.3390/electronics9101710

84. Hu W, Tan Y (2022) Generating adversarial malware examples for black-box attacks based on GAN. In: Tan Y, Shi Y (eds) Data Mining and Big Data. Springer, Singapore, pp 409–423
85. Peng X, Xian H, Lu Q, Lu X (2021) Semantics aware adversarial malware examples generation for black-box attacks. Applied Soft Computing 109:107506. https://doi.org/10.1016/j.asoc.2021.107506
86. Roy D, Mukherjee T, Chatterjee M, Pasiliao E (2019) Detection of rogue RF transmitters using generative adversarial nets. In: 2019 IEEE Wireless Communications and Networking Conference (WCNC), pp 1–7. https://doi.org/10.1109/WCNC.2019.8885548
87. Chen Z, Peng L, Hu A (2021) Fu H (2021) Generative adversarial network-based rogue device identification using differential constellation trace figure. EURASIP Journal on Wireless Communications and Networking 1:72. https://doi.org/10.1186/s13638-021-01950-2
88. Behringer MH, Pritikin M, Bjarnason S, Clemm A, Carpenter BE, Jiang S, Ciavaglia L (2015) Autonomic networking: definitions and design goals. RFC Editor. https://doi.org/10.17487/RFC7575
89. Soman S, G RH (2023) Observations on LLMs for Telecom domain: capabilities and limitations. arXiv:2305.13102
90. Gräsler I, Preus D, Brandt L, Mohr M (2022) Efficient extraction of technical requirements applying data augmentation. In: 2022 IEEE International Symposium on Systems Engineering (ISSE), pp 1–8 . https://doi.org/10.1109/ISSE54508.2022.10005452
91. Allam S (2023) AI-based use-pattern generative hybrid spaces for indoor and outdoor activities. In: 2023 20th Learning and Technology Conference (L&T), pp 54–58. https://doi.org/10.1109/LT58159.2023.10092345
92. Mentzer F, Agustsson E, Ballë J, Minnen D, Johnston N, Toderici G (2022) Neural video compression using GANs for detail synthesis and propagation. [eess.IV] arXiv:2107.12038
93. Moustris G, Tzafestas C, Konstantinidis K (2023) A long distance telesurgical demonstration on robotic surgery phantoms over 5G. International Journal of Computer Assisted Radiology and Surgery. https://doi.org/10.1007/s11548-023-02913-2
94. Xu S, Perez M, Yang K, Perrenot C, Felblinger J, Hubert J (2014) Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dV-Trainer (r) simulator. Surgical endoscopy 28. https://doi.org/10.1007/s00464-014-3504-z
95. Inam R, Schrammar N, Wang K, Karapantelakis A, Mokrushin L, Feljan AV, Fersman E (2016) Feasibility assessment to realise vehicle teleoperation using cellular networks. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp 2254–2260. https://doi.org/10.1109/ITSC.2016.7795920
96. Zhang JA, Rahman ML, Wu K, Huang X, Guo YJ, Chen S, Yuan J (2022) Enabling joint communication and radar sensing in mobile networks-a survey. IEEE Communications Surveys & Tutorials 24(1):306–345. https://doi.org/10.1109/COMST.2021.3122519
97. Li X, Chen M, Liu Y, Zhang Z, Liu, D, Mao S (2023) Graph neural networks for joint communication and sensing optimization in vehicular networks. arXiv:2302.02878
98. Rahnemoonfar M, Johnson J, Paden J (2019) AI radar sensor: creating radar depth sounder images based on generative adversarial network. Sensors 19(24) https://doi.org/10.3390/s19245479
99. Serreli L, Nonnis R, Bingöl G, Anedda M, Fadda M, Giusto DD (2021) Fingerprint-based positioning method over LTE advanced pro signals with GAN training contribute. In: 2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp 1–5. https://doi.org/10.1109/BMSB53066.2021.9547015
100. Wei W, Yan J, Wan L, Wang C, Zhang G, Wu X (2021) Enriching indoor localization fingerprint using a single AC-GAN. In: 2021 IEEE Wireless Communications and Networking Conference (WCNC), pp 1–6. https://doi.org/10.1109/WCNC49053.2021.9417513
101. Xie H, Qin Z, Li GY, Juang B-H (2021) Deep learning enabled semantic communication systems. IEEE Transactions on Signal Processing 69:2663–2675. https://doi.org/10.1109/tsp.2021.3071210
102. Weng Z, Qin Z (2021) Semantic communication systems for speech transmission. IEEE Journal on Selected Areas in Communications 39(8):2434–2444. https://doi.org/10.1109/JSAC.2021.3087240
103. Yoo H, Jung T, Dai L, Kim S, Chae C-B (2022) Demo: Real-time semantic communications with a vision transformer. In: 2022 IEEE International Conference on Communications Workshops (ICC Workshops), pp 1–2. https://doi.org/10.1109/ICCWorkshops53468.2022.9914635
104. Tsialiamanis G, Wagg DJ, Dervilis N, Worden K (2021) On generative models as the basis for digital twins. Data-Centric Engineering 2:11. https://doi.org/10.1017/dce.2021.13
105. Mozo A, Karamchandani A, Gómez-Canaval S, Sanz M, Moreno JI, Pastor A (2022) B5GEMINI: AI-driven network digital twin. Sensors 22(11) https://doi.org/10.3390/s22114106
106. Baldvinsson JR, Ganjalizadeh M, AlAbbasi A, Björkman M, Payberah AH (2022) IL-GAN: rare sample generation via incremental learning in GANs. In: GLOBECOM 2022 - 2022 IEEE Global Communications Conference, pp 621–626. https://doi.org/10.1109/GLOBECOM48099.2022.10001069
107. Svyatkovskiy A, Deng SK, Fu S, Sundaresan N (2020) Intellicode compose: code generation using transformer. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ESEC/FSE 2020, pp 1433–1443. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3368089.3417058
108. Sun Z, Du X, Song F, Wang S, Ni M, Li L (2023) Don't complete it! Preventing unhelpful code completion for productive and sustainable neural code completion systems. arXiv:2209.05948 [cs.SE]
109. Ganesh P, Chen Y, Lou X, Khan MA, Yang Y, Sajjad H, Nakov P, Chen D, Winslett M (2021) Compressing large-scale transformer-based models: a case study on BERT. Transactions of the Association for Computational Linguistics 9:1061–1080. https://doi.org/10.1162/tacl_a_00413
110. Geiger A, Liu D, Alnegheimish S, Cuesta-Infante A, Veeramachaneni K (2020) TadGAN: time series anomaly detection using generative adversarial networks
111. Li Y, Peng X, Zhang J, Li Z, Wen M (2023) DCT-GAN: dilated convolutional transformer-based GAN for time series anomaly detection. IEEE Transactions on Knowledge and Data Engineering 35(4):3632–3644. https://doi.org/10.1109/TKDE.2021.3130234
112. Zhang Q, Xiao T, Efros AA, Pinto L, Wang X (2021) Learning cross-domain correspondence for control with dynamics cycle-consistency. In: International Conference on Learning Representations. https://openreview.net/forum?id=QIRlze3I6hX
113. Dey K, Perepu SK, Dasgupta P, Das A (2023) Domain adaptation of reinforcement learning agents based on network service proximity. In: 2023 IEEE 9th International Conference on Network Softwarization (NetSoft), pp 152–160 (2023). https://doi.org/10.1109/NetSoft57336.2023.10175507
114. d'Ascoli S, Touvron H, Leavitt ML, Morcos AS, Biroli G (2022) Sagun L (2022) Convit: improving vision transformers with soft convolutional inductive biases. Journal of Statistical Mechanics:

Theory and Experiment 11. https://doi.org/10.1088/1742-5468/ac9830

115. TSG RAN - radio access network. https://www.3gpp.org/3gpp-groups/radio-access-networks-ran
116. Wu W, Yang B, Wang D, Zhang W (2020) A novel trajectory generator based on a constrained GAN and a latent variables predictor. IEEE Access 8:212529–212540. https://doi.org/10.1109/ACCESS.2020.3039801
117. Wei H, Ye D, Liu Z, Wu H, Yuan B, Fu Q, Yang W, Li Z (2021) Boosting offline reinforcement learning with residual generative modeling. In: Zhou Z-H (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pp 3574–3580. Main Track. https://doi.org/10.24963/ijcai.2021/492

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.