Machine Learning in Communications: A Road to Intelligent Transmission and Processing

Shixiong Wang, Geoffrey Ye Li

Abstract

Prior to the era of artificial intelligence (AI) and big data, research into wireless communications primarily followed a conventional route involving problem analysis, model building and calibration, algorithm design and tuning, and holistic and empirical verification. However, this methodology often faced limitations when dealing with large-scale and complex problems and managing dynamic and massive data, resulting in inefficiencies and limited performance of traditional communication systems and methods. As such, wireless communications have embraced the revolutionary impact of AI and machine learning (ML), leading to the development of more adaptive, efficient, and intelligent systems and algorithms. This technological shift paves the way to intelligent information transmission and processing. This paper discusses the typical roles of ML in intelligent wireless communications, as well as its features, challenges, and practical considerations.

Keywords

machine learning, intelligent transmission, intelligent processing

1 Introduction

Since the 19th century, radio communications have started a new era of information transmission for human society. The early stages of radio transmission technology, such as Morse codes and telegraph machines, relied heavily on manual operations, which limited the efficiency and reliability of information exchange. Aiming to automate the task of information transmission and processing at a sophisticated level, the first-generation concept of "intelligent transmission and processing" emerged. Subsequently, the 20th century witnessed significant developments in module-based communication systems. These systems encompass essential modules such as source coding, channel coding, modulation, transmit beamforming, wireless channel transmission, receive beamforming, demodulation, signal detection, channel decoding, and source decoding [1–3]. Methodologically, the development of module-based wireless communications and signal processing follows a systematic research trajectory that includes problem analysis, model development and calibration, algorithm design and optimization, and empirical validation, feedback, and improvement. Notably, every step in this methodological loop demands large volumes of human intellectual endeavors.

In the 21st century, wireless communication systems are expected to deliver extremely vast amounts of data in various formats, such as audio, video, and text, while ensuring low latency, high data rates, and reliability. Furthermore, the incorporation of new network topologies (e.g., internet-of-things networks, unmanned-aerial-vehicle relay networks) and cutting-edge functions (e.g., integrated sensing and communications [ISAC], integrated computing and communications [ICAC]) has added complexity to the design of modern communication systems. This complexity is particularly evident in the following three aspects:

- Addressing different types of modeling uncertainties in not only holistic systems but also individual modules
- Leveraging various forms of big data generated by user equipment and base stations
- Solving challenging algorithmic problems in realizing the networks

Traditional design methods rely heavily on the intensive human intellectual efforts and are proven inadequate for managing large-scale and complex issues and handling dynamic extensive data. This inadequacy results in inefficiencies and limited performance of information transmission and processing. In response, wireless communications and signal processing have embraced the transformative potential of artificial intelligence (AI) and machine learning (ML) [4]; for comprehensive surveys of ML on communications, see [5–11]. This technological and methodological shift has enabled the development of more adaptive, efficient, robust, and intelligent systems and algorithms. Consequently, the second-generation concept of "intelligent transmission and processing" is emerging, aiming to significantly reduce the need for human intellectual efforts and improve the integrated performances of communication systems.

Figure 1 illustrates the philosophical connotations of intelligent transmission and processing. A technical visualization of ML-empowered intelligent transmission and processing is shown in Figure 2.

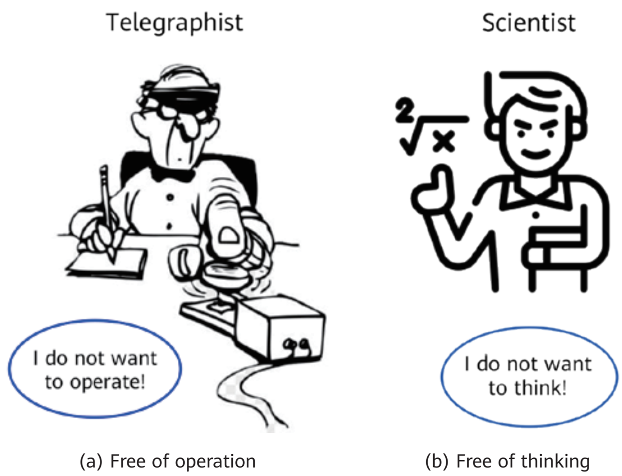

Figure 1 Connotations of intelligent transmission and processing (Icon credit: CLEANPNG.com and FLATICON.com)

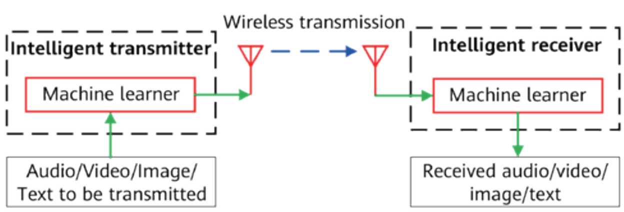

Figure 2 An end-to-end structure of intelligent transmission and processing systems. The intelligent transmitter and receiver act as end-to-end information processors f, where each f is learned by machines from data; the transmitter and receiver can automatically adapt to the real-time characteristics of wireless channels. The intelligence is reflected in the sense that human intellectual efforts are no longer explicitly required to study large-scale, dynamic, and uncertain information transmission mechanisms and processing solutions.

To showcase the power of ML techniques in enabling intelligent transmission and processing, this paper reviews trending ML applications in communication systems and methods, including physical-layer communications [6, 12], semantic communications [13, 14], resource allocations in communications [15], ICAC [16], and ISAC [17]. However, the ambition of this paper is not to offer an exhaustive list of all existing works in the area. Rather, we aim to illuminate the path towards intelligent transmission and processing.

Although ML has the potential to reform the theory and practice of wireless communications, the challenges and disadvantages of utilizing ML-based approaches accompany its opportunities and advantages [4], for example, the reliability issue due to the lack of interpretability of black-box learning methods (e.g., deep learning), the generalization issue due to the limited training data and the non-stationarities of the underlying data-generating laws, and the resource deficits in training and storing large ML models (e.g., deep learning); see Figure 3 for a motivational understanding. In addition to the three primary challenges exemplified, other instances may also arise, e.g., the scalability issue caused during the reconfiguration of system topology or hardware (e.g., removing or adding antennas; which can be seen as a kind of generalization problem) and the security and privacy issue in networked learning [16]. The main message is that in advancing communication theories and developing communication systems, the role of ML should not be overstated: ML (especially data-driven deep learning) can be a valuable factor to consider rather than an absolute rule to follow; problem analyses and mechanism modeling are always important; see [18–20] for technical investigations and justifications; see also the example below for a motivational understanding.

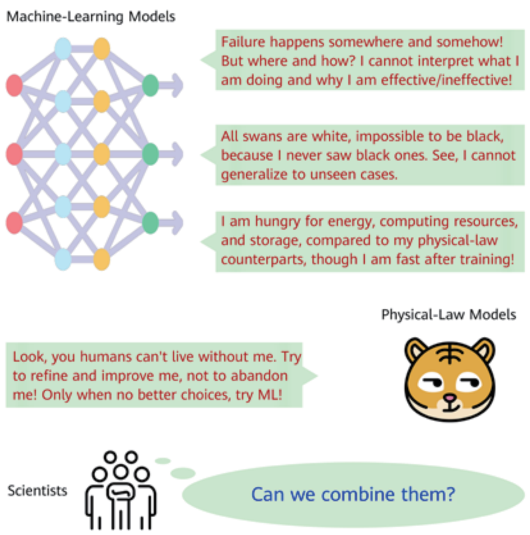

Figure 3 To choose an ML model or a physical-law model, this is a question! Nothing is free, although some are cheap! Choose the one that fits your situation and expectations well! Whenever possible, combine them to improve the overall system performance. (Icon credit: FLATICON.com)

Example (Ice Cream Sales and Shark Attacks): Regression analysis using historical data shows a positive relationship between ice cream sales and shark attacks, which is illogical. However, the primary factor driving this correlation is temperature: higher temperatures lead to increased ice cream sales and beach attendance; more beach visitors result in more shark attacks [21]. Hence, mechanism modeling is vital.

Before digging into ML applications in communications, we quickly review the essentials of ML concepts and methods in Section 2, especially those of trustworthy ML. The aim is to highlight the primary considerations, including philosophical and technical facets, of using ML in wireless communications.

2 Machine Learning Concepts and Methods

ML is concerned with discovering hidden information and patterns from data. The primary advantage is its ability to explain data automatically, thus freeing humans from studying the underlying data-generating mechanisms. This feature inherently enables machine intelligence in the practice of communications, specifically, in the transmission and processing of information [5, 7, 8, 11, 22].

Depending on the characteristics of tasks, ML can be categorized into four genres: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Mathematically, the key to all ML tasks is to find a function f, called a hypothesis, that maps the observed data to a desired decision; see Figure 4 for a conceptual illustration. Specific examples of ML are as follows.

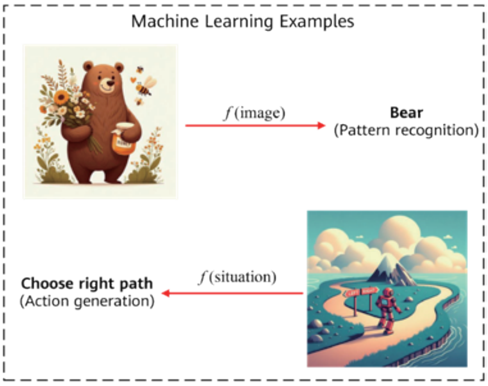

Figure 4 Conceptual illustration of ML. ML is to find a mapping f from the input data to a decision. The upper example is a supervised ML problem where an image classifier f recognizes the image as a bear. The lower example is a reinforcement learning problem where an action generator f recommends the robot to choose the right path in the current situation. (The two images are generated by Microsoft Copilot.)

- Supervised Learning: Supervised learning summarizes the hidden information of labeled data and includes regression and classification as two main types of tasks. Given the deterministic or random variable pair (x, s) where x is the feature vector and s is the continuous-valued expected response, regression aims to find a functional relation f from x to s such that predicted label $f(x)$ can be as close as possible to the target label s. The closeness is measured by a loss function L through $L[s, f(x)]$, for example, the mean-squared error $[s - f(x)]^H[s - f(x)]$. In handling the deterministic variable pair (x, s), there potentially exists a function f such that $f(x)$ is exactly equal to s for every realization of (x, s), i.e., $L[s, f(x)] \equiv 0$. As for the random variable pair (x, s), however, the exact equality cannot be generally guaranteed. Instead, the loss is calculated under the joint distribution $\mathbb{P}_{\mathrm{x},\ \mathrm{s}}$ of (x, s), for example, the expectation of the loss $\mathbb{E}_{(\mathrm{x},\ \mathrm{s}) \sim \mathbb{P}_{\mathrm{x},\ \mathrm{s}}} L[\mathbf{s}, f(\mathbf{x})]$. When s is one-dimensional and takes discrete values, we have the classification problems where $L[s, f(x)]$ is defined by, e.g., the indicator function $\mathbb{I}\{\mathrm{s} \neq f(\mathrm{x})\}$. In this case, s indexes different target classes; (e.g., for binary classification, $s \in \{-1, 1\}$).
- Unsupervised Learning: In unsupervised learning, there are no labels s for the collected data, and only feature data x is present. Therefore, we focus on discovering the hidden information from the realizations of the datum variable x. Clustering is a typical task of unsupervised learning. Different from classification which predicts the categorical labels s of new data points x based on a training dataset with known labels, clustering aims at grouping similar data points x together based on their features without predefined categorical labels. In summary, clustering is to find a function f that maps data x into a suitable group. Feature transformation is another example of unsupervised learning, which transforms original feature data x into another feature space using a learned mapping f; Specifically, $y = f(x)$ —compare this with a time-domain signal x and its Fourier transform y. Autoencoders, a type of artificial neural network, provide an excellent example of feature transformation through their encoding and decoding operations. Yet another important example of unsupervised learning is distribution estimation, i.e., to estimate the data-generating distribution that best fits (or describes) the collected data. Distribution estimation is particularly vital in generative tasks such as producing a new sample based on collected samples; for example, given a group of cat images, determining how to produce a new cat image by drawing from the fitted distribution?
- Semi-supervised Learning: Semi-supervised learning can be considered a variant of supervised learning as it extends the principles of supervised learning by incorporating a mixture of labeled data (x, s) and unlabeled data x'. While supervised learning relies entirely on labeled data to train the model, semi-supervised learning aims to improve model performance and generalization capabilities by leveraging the additional unlabeled data. The semi-supervised approach is particularly useful when acquiring large amounts of labeled data is expensive or time-consuming, while unlabeled data is abundant and easy to obtain. By leveraging the information from the unlabeled data along with that from the labeled data, semi-supervised learning can find a model f that has better prediction performance (of the label s associated with the data x) compared to purely supervised learning methods that rely solely on labeled data.
- Reinforcement Learning: Reinforcement learning is concerned with decision-making problems in a dynamic and uncertain environment. Unlike supervised learning, which uses labeled data, and unsupervised learning, which finds patterns in unlabeled data, reinforcement learning involves the agent interacting with the environment, receiving feedback in the form of rewards or penalties, and using this feedback to learn optimal behaviors or strategies over time. To be specific, the agent autonomously learns to make decisions in an environment by performing actions a, in response to current states s, in order to maximize the cumulative reward. Therefore, mathematically, an action-generating function f from state s to action a needs to be learned.

For specific applications of the four ML genres in wireless communications, refer to [8, 23].

Data-Driven and Model-Driven Learning: Considering the degree of human intelligence and domain knowledge involved, ML can be classified into data-driven and model-driven approaches. Data-driven ML relies entirely on historical data and does not involve any analysis of underlying data-generating mechanisms. In contrast, model-driven ML incorporates, to varying extents, studying the underlying physical mechanisms and data-generating models. Intelligent information transmission and processing can benefit, in terms of improving overall performance, from the collaboration between communication-systems modeling and big data discovery [12, 23]. For example, in signal detection, suppose that we have T pilot data pairs $\{(s_1, x_1), (s_2, x_2), \ldots, (s_T, x_T)\}$ where x_i are the received signals and s_i are the transmitted symbols, for $i = 1, 2, \ldots, T$. Data-driven ML directly utilizes all the data to train a detector f from x to s. In contrast, model-driven ML first considers the signal-transmission model $x = Hs + v$ where H denotes the channel matrix and v the channel noise, and then finds a detector f based on the above underlying data-generating mechanism. For detailed technical treatments and discussions, see [19, 20, 24, 25].

Hypothesis Space and Deep Learning: To locate a best decision function f, a candidate function space $\mathcal{H}$ (called hypothesis space) from which f is drawn, needs to be specified. To clarify further, for instance, supervised statistical ML can be formulated as:

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(\mathbf{x},\mathbf{s}) \sim \mathbb{P}_{\mathbf{x},\mathbf{s}}} L(\mathbf{s}, f(\mathbf{x})),$$

where the joint distribution $\mathbb{P}_{\mathrm{x},\,\mathrm{s}}$, which is unknown in practice, can be estimated using collected historical data (e.g., using empirical distribution). As an example, signal detection problems can be characterized as described earlier, where f is a detector, x is the antenna-received signal, and s is the transmitted symbol (e.g., constellation points) [20]; the loss function L can be mean-squared error or symbol-error rate. Canonical examples for hypothesis space $\mathcal{H}$ are as follows.

- Linear Function Space: $\mathcal{H}$ only includes the linear transforms of input x. In the signal detection case, $\mathcal{H}$ contains only linear detectors.
- Reproducing Kernel Hilbert Space: $\mathcal{H}$ includes all linear transforms of the nonlinearly-lifted-feature $\varphi(x)$ of the original feature x, using a feature mapping function φ. In essence, $\mathcal{H}$ contains some specific types of nonlinear functions of input x.
- Neural Network Function Space: $\mathcal{H}$ is represented (or structured, characterized) by neural networks, for instance, multi-layer perceptron, recurrent neural networks, convolutional neural networks (CNNs), radial basis neural networks, autoencoders, or transformers. Each given neural network defines a special type of function space $\mathcal{H}$. When the employed neural network has deep structures with many hidden layers being included, $\mathcal{H}$ denotes a deep-neural-network function space. Upon operating with deep neural networks, ML is referred to as deep learning.

On the other hand, with the involvement of domain knowledge and expert designs, a hypothesis space $\mathcal{H}$ can be accordingly adapted or tailored to a domain-specific problem [12, 18, 23]. Therefore, model-driven ML is to devise an ad-hoc and structured candidate space $\mathcal{H}$, by leveraging known problem characteristics and data-generating mechanisms.

Explainability, Reliability, and Sustainability: Modern ML research addresses several advanced concerns, including explainability, reliability, and sustainability of learning models [26, 27]. Explainable ML seeks to make learning models transparent, interpretable, and accountable through techniques such as feature engineering and physical modeling [28]; model-driven ML, which leverages underlying physical data-generating mechanisms, can be seen as such a scheme [12, 18]. Reliable ML focuses on creating robust and accurate learning models that generalize well to new data (that are not used in the training stage), tackling issues such as overfitting, generalization, knowledge migration, and limited-sample learning [20, 29–31]. Sustainable ML aims to develop learning models with minimal negative impact on the environment and society, addressing energy efficiency, privacy and security, and fairness and bias [16, 32]. In the context of intelligent information transmission and processing, the three considerations (i.e., explainability, reliability, and sustainability) are of natural importance and significance. Therefore, they are the primary considerations in developing ML-based solutions for wireless communications.

Centralized and Distributed Learning: ML models can be trained using various approaches depending on the structure of the data distribution and the architecture of the computation. Two primary paradigms in this context are centralized learning and distributed learning [33, 34]. Centralized learning involves collecting and storing all training data in a single central location, such as a data center or cloud server, and the ML model is trained on this aggregated dataset. Distributed learning, on the other hand, involves training ML models in a distributed manner across multiple devices (or nodes), each of which holds a portion of the data. A prime example of distributed learning is federated learning, where multiple clients (e.g., smartphones, internet-of-things devices, or different organizations) collaboratively train a model without sharing their local data. Instead, each client trains the model on its local data and only shares the model updates (gradients or weights) with a central server, which aggregates these updates to form a global model. Both centralized and distributed learning methods are beneficial for advancing future-generation communication systems because they can adapt to diverse modern communication network typologies.

3 Physical Layer Communications

Physical layer communications aim to reliably transmit raw data streams, e.g., binary bits, through physical mediums. Figure 5 presents a traditional architectural diagram of wireless communications, featuring various functional modules (or blocks) that are meticulously designed by humans in accordance with fundamental mathematical and physical principles. This block-based diagram is structurally different from the ML-empowered architectural diagram shown in Figure 2, where interconnected functional modules are taken over by end-to-end operating parts.

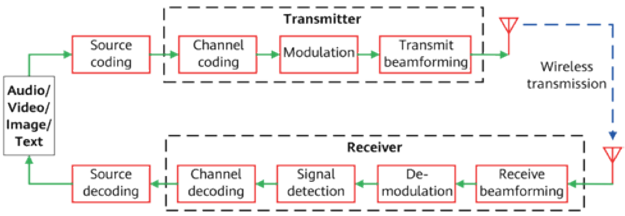

Figure 5 The module-based structure of traditional transmission and processing systems. Every block acts as an information processor f, where f is elaborately designed by scientists based on underlying physical mechanisms and mathematical laws.

In addition to the highly integrated (i.e., highly intelligent) structure in Figure 2, ML-based transmission and processing systems can also be partially intelligentized. For example, in one scenario, only the channel coding or decoding block is managed by ML, meaning that the channel coding scheme is designed by machines rather than information scientists. In another scenario, ML is used solely for the transmit beamformer or the receive beamformer. A conceptual illustration is shown in Figure 6.

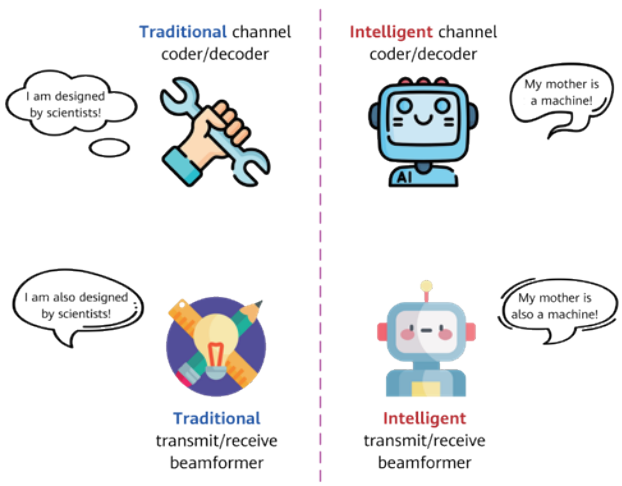

Figure 6 In contrast to the highly integrated structure in Figure 2, each individual module in Figure 5 can be augmented by ML. (Icon credit: FLATICON.com.)

In technical details, applications of ML in physical layer communications include overall end-to-end system design [35, 36] (Figure 2) and individual module design (Figure 6). The latter, to be specific, encompasses:

- coding/decoding techniques, for example, source coding [37], channel coding [38, 39] , and joint source-channel coding (JSCC) [40, 41]
- signal modulation and detection [25, 42]
- transmit and receive beamforming [20, 43–46], for example, beam alignment and beam tracking [47–50]
- channel estimation and feedback [51, 52]

among many others. For comprehensive and recent surveys, see [6, 53, 54].

Coding and decoding techniques are vital in digital communications, ensuring efficient and reliable data transmission. Recently, ML has been increasingly applied to enhance these techniques, encompassing source coding, channel coding, and JSCC. Traditionally, source coding (i.e., data compression) reduces data redundancy for efficient transmission and storage. ML techniques, such as neural-network-based autoencoders, have revolutionized this field. Autoencoders learn efficient representations of data by encoding it into a lower-dimensional space and then reconstructing it, achieving high compression rates with minimal loss of information [55]. Channel coding adds redundancy to data in order to detect and correct transmission errors caused by noisy channels. ML models, particularly deep learning techniques, have been applied to develop novel error correction codes. For example, neural decoders have been designed to decode complex schemes like low-density parity-check (LDPC) [56] and Turbo codes [57], offering improved performance over traditional algorithms, especially in highly noisy environments. JSCC integrates source and channel coding to optimize overall system performance. ML models, such as variational autoencoders (VAEs) [58], CNNs [59], and generative adversarial networks (GANs) [60], are used to jointly learn the representation and error correction codes. These models can adapt to the characteristics of both the source and the channel, achieving better compression and error resilience than traditional methods. Overall, ML-based coding and decoding techniques represent a significant advancement in digital communications. By leveraging the predictive and adaptive capabilities of ML, these techniques enhance data compression, error correction, and overall transmission efficiency. This lays the foundation for creating communications systems that are more robust and efficient.

Signal modulation and detection, which enable the transmission and interpretation of data over various channels, are fundamental processes in digital communications. Recently, ML techniques have been applied to enhance these processes, improving efficiency and reliability. Modulation involves altering a carrier signal's properties, such as amplitude, frequency, or phase, to encode information. Traditional modulation schemes include amplitude modulation (AM), frequency modulation (FM), and phase shift keying (PSK). ML techniques, particularly deep learning models, are now used to design adaptive modulation schemes. These models can dynamically adjust modulation parameters based on the channel conditions, optimizing performance in real time. For instance, neural networks can learn complex modulation patterns that maximize data throughput and minimize error rates [61]. Detection involves demodulating the received signal to recover the transmitted information. Traditional methods rely on predefined algorithms to estimate the transmitted data but often assume specific channel characteristics. ML approaches, such as fully connected deep neural networks [42] and transfer learning [62], have been employed to enhance signal detection. These models can learn from data to accurately detect signals under varying and complex channel conditions, improving robustness against noise and interference. Overall, the integration of ML techniques in signal modulation and detection represents a significant leap forward in communications technology — it enhances data transmission efficiency, resilience to noise, and overall system performance.

In wireless communication systems, transmit and receive beamforming techniques are essential for enhancing signal quality and increasing data throughput. Beamforming directs the transmission or reception of signals in specific directions using antenna arrays, improving signal strength and reducing interference. Recently, ML techniques have significantly advanced beamforming performance and adaptability. For transmit beamforming, traditional methods, such as phased array systems, use predefined algorithms to adjust the phase and amplitude of signals from multiple antennas. In contrast, ML techniques, particularly deep learning models, optimize this process by learning from environmental data. As an example, reinforcement learning can dynamically adjust beamforming patterns in real time based on feedback from the communication environment, enhancing performance in complex and changing scenarios [63, 64]. For receive beamforming, conventional methods, such as minimum variance distortionless response (MVDR) and maximal ratio combining (MRC), rely on statistical models of the signal environment. ML approaches, such as CNNs, improve upon these by learning optimal beamforming weights directly from data, allowing for more accurate and robust signal reception in diverse and dynamic environments [65, 66]. Beam alignment and tracking are crucial subcategories of beamforming, particularly important in high-frequency bands like millimeter-wave (mmWave) and terahertz communications. These techniques ensure that the transmitter and receiver maintain optimal beam alignment to maximize signal strength and data throughput [49, 50, 67]. Traditional alignment methods rely on exhaustive search or iterative algorithms, which are time-consuming and computationally intensive. ML approaches, such as supervised learning, multi-armed bandits, and reinforcement learning, provide more efficient solutions by predicting optimal beam directions from historical data, significantly reducing the search space. Beam tracking maintains alignment as the transmitter or receiver moves or as the environment changes. ML techniques, particularly deep learning models, enhance tracking by predicting beam direction changes in real time. Recurrent neural networks and long short-term memory networks, which capture temporal dependencies, are particularly effective for this purpose. For technical details on ML-based beam alignment and tracking, see [47–50, 67]. In summary, the integration of ML into beamforming, including beam alignment and tracking, is critical for next-generation networks such as 5G and beyond, because these ML-driven techniques can leverage predictive and adaptive capabilities to enhance signal quality, reduce interference, and optimize system performance.

Channel estimation and feedback techniques are of high importance in wireless communication systems for accurately characterizing the communication channel and ensuring efficient data transmission. These processes involve measuring the channel's properties and providing necessary feedback to transmitters. Recently, ML techniques have been applied to enhance these processes, offering significant improvements in accuracy and efficiency [68, 69]. Channel estimation involves predicting the state of the communication channel to optimize signal transmission and reception. Traditional methods, such as minimum mean-squared error (MMSE), rely on statistical models and require significant computational resources. ML approaches, especially deep learning models, have introduced new ways to perform channel estimation with higher accuracy, and potentially, lower computational complexity. For instance, CNNs can learn to estimate channel states directly from received signal data, providing more robust and adaptive solutions in complex environments [70]. Long short-term memory networks are particularly effective for capturing temporal dependencies in channel conditions, improving estimation accuracy [71]. Feedback mechanisms forward channel state information (CSI) from the receiver back to the transmitter, allowing for real-time adaptation of transmission setups. Traditional feedback methods often involve quantizing and encoding the CSI, which may incur delays and inaccuracies. ML techniques, such as autoencoders and CNNs, improve feedback efficiency by compressing and reconstructing the CSI with minimal loss of information [52]. This allows for more precise and timely adjustments to transmission setups. In addition, ML models can simultaneously handle channel estimation and feedback, optimizing both processes in an integrated manner [69]. This holistic approach leverages the strengths of ML to enhance overall system performance.

4 Semantic Communications

Semantic communications, unlike conventional physical-layer communications, focus on transmitting semantic information conveyed in original data (e.g., image, text, audio) rather than bit-wise raw information. The primary benefit of semantic communications is that the transmission overloads of wireless channels can be significantly reduced compared to bit-wise transmission. Consequently, the information transmission speed and efficiency can be considerably improved. For comprehensive and recent surveys in semantic communications, see [72–75].

The key to semantic communications is to extract the semantic information from raw data. Therefore, semantic communications can be realized by elegantly designing the source coding and decoding strategies. It can also be actualized through JSCC and decoding. The difficulty, however, is that the semantic information of given raw data is specific to a task (see Figure 7), due to which, a generally well-accepted mathematical analysis, modeling, and computing framework for semantic communications is still lacking; for exploring works in this direction, see [76]. Therefore, for a specified communication task, the semantic coding and decoding schemes need to be elaborated. In the context of intelligent transmission and processing, semantic communications can be implicitly realized in highly integrated end-to-end transceivers, shown in Figure 2.

ML approaches play a pivotal role in semantic communications, enabling systems to understand, process, and convey meaning more accurately. Techniques such as deep learning models, including transformers, CNNs, and recurrent neural networks, have been widely utilized to analyze and predict the semantic relevance of data, thus optimizing bandwidth usage and improving communication efficiency. To be specific, natural language processing (NLP) algorithms allow for the extraction and interpretation of semantic content from text and audio, facilitating more meaningful data compression and transmission [77, 78]; computer vision methods, on the other hand, enable the extraction and interpretation of semantic meaning from image and video [79].

Recent research has demonstrated the potential of ML-driven semantic communications in various applications. For example, in [13], transceiver neural networks have been designed to directly transmit text semantic meaning, which significantly reduces the demand on communication resources and improves the overall transmission performance. Another example is [80], in which an efficient system for video conferencing is developed to improve transmission efficiency. Most studies in semantic communications focus on JSCC to save communication resources. However, this approach requires changing the existing communication infrastructures and therefore hinders practical implementation. As such, a pragmatic approach to wireless semantic transmission through revising some modules in existing infrastructures is reported [14]. To guarantee semantic transmission reliability and communication efficiency, the spectral efficiency in the semantic domain and the semantic-aware resource allocation issues have been investigated in [81]. In addition to the above representative applications, the synergy between semantic communications and emerging technologies, such as the internet of things (IoT) [82] and edge computing [83], is fostering new opportunities for intelligent and context-aware communication systems. By leveraging distributed ML models, semantic communication systems can dynamically adapt to changing environmental conditions and user requirements, ensuring robust and efficient information exchange [84].

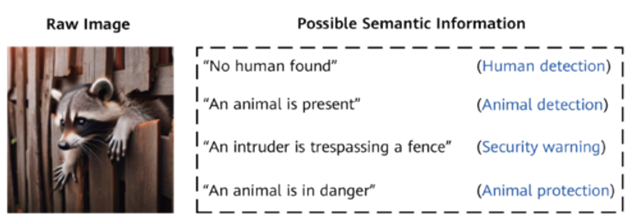

Figure 7 The semantic information of raw data is task-specific. Losslessly transmitting a high-definite image is time- and resource-consuming. However, accurately transmitting a semantic message can be relatively simpler and cheaper. (The image is generated by Microsoft Copilot.)

In summary, semantic communications, underpinned by advanced ML techniques, define a brilliant future direction for communication systems. This innovative approach promises to reform how information is transmitted and understood, offering profound implications for the efficiency and effectiveness of future communication networks.

5 Resource Allocation in Communications

In wireless communications, resource allocation is concerned with how to efficiently manage and utilize spectra, power, computing, space, and time resources, thus improving the overall communication network performance, e.g., higher throughput, lower latency, larger coverage, higher reliability, to name a few [15, 85, 86]. Typical applications encompass link scheduling, message routing, power allocation, channel selection, beamforming, spectra access and management, and division protocol design (i.e., time division, frequency division, etc.). From the mathematical programming perspective, resource allocation is often formulated as optimization problems. From the operations research perspective, assignment and scheduling are two pivotal techniques; the former handles static resource allocation problems, while the latter addresses dynamic ones; the static and dynamic features are with respect to time. From the computational and algorithmic perspective, standard and trending solution frameworks include the following:

- Continuous optimization, discrete (e.g., combinatorial, integer) optimization, and mixed optimization
- Single-objective optimization and multi-objective optimization
- Linear programming and nonlinear programming
- Convex optimization and non-convex optimization
- Smooth optimization and non-smooth optimization
- Min-Max optimization (e.g., game theory, worst-case robust analyses)
- Deterministic programming and stochastic programming (i.e., whether random variables are involved; if involved, associated distributions are considered)
- Single-stage optimization (i.e., static programming) and multi-stage optimization (i.e., dynamic programming)
- Heuristic optimization (e.g., genetic algorithm, particle swarm optimization, simulated annealing)
- Surrogate optimization which is also known as black-box optimization (e.g., Bayesian optimization)
- ML-based optimization (e.g., solution methods based on reinforcement learning and deep learning)

The canonical applications and solution frameworks of resource allocation in wireless communications are shown in Figure 8. For introductory and motivational reading on this topic, refer to [85, 87, 88].

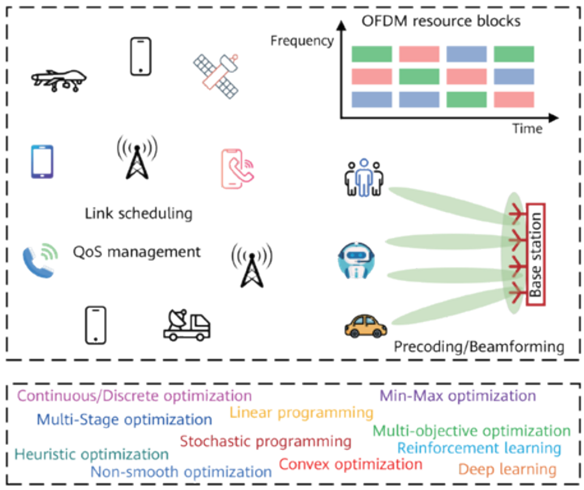

Figure 8 Typical applications and solution frameworks of resource allocation in wireless communications. OFDM: orthogonal frequency division multiplexing; QoS: quality of service. (Icon credit: FLATICON. com.)

Nowadays, the advent of ISAC [89] has slightly changed the connotation of traditional resource allocation. This is because the best resource allocation scheme for communication is not necessarily the same as (or in accordance with) that for sensing; see [90]. For example, the optimal waveforms for communication and sensing are usually dissimilar [91, 92] because the two radio functions have different or even contradicting design preferences. Therefore, diverse resources, including radio, computing, power, time, beam, etc., should be delicately allocated to satisfy the individual performance requirements of communication and sensing. The same dilemma holds for ICAC, e.g., edge computing [93] and networked control* [94], because limited resources need to be elegantly distributed to computing and communication. For further details, see [95, 96].

ML techniques have emerged as powerful tools to address the challenges brought by resource allocation in wireless communications. Recent advancements have demonstrated the potential of ML in various resource allocation tasks. For instance, deep reinforcement learning has been applied to optimize spectrum allocation, power control, and user

*Controllers are, by their nature, information processors and are therefore ad-hoc computing modules.

association in heterogeneous networks, showing significant improvements over conventional methods [22, 97–102]. Similarly, supervised learning algorithms have been used to efficiently solve complex optimization problems in resource allocation such as mixed integer nonlinear programming (MINLP) [103]. In addition, unsupervised learning techniques can also be employed to solve resource allocation problems and refine the solutions, e.g., the graph embedding trick in link scheduling [104].

Traditional resource allocation methods rely heavily on human intellect to build exact models and develop ad-hoc solution methods, which can be suboptimal and even inflexible in dynamic, complex, and large-scale scenarios. ML, particularly deep learning and reinforcement learning, offers the ability to model complex interactions with environments, predict communication-network states, and optimize decisions in a real-time manner, thereby enhancing the overall performance and adaptability of wireless networks [15]. To be specific, communication channels in practice are often time-varying, however, mathematically considering such model uncertainties is not straightforward. This is because we do not exactly know how the channels evolve over time. Even worse, the resultant mathematical programming models are computationally complex, and therefore, hard to be efficiently and optimally solved. The role of ML, in this sense, is to leverage accessible real-world data, discover the hidden knowledge and patterns that the data convey, and automatically find satisfactory resource allocation decisions. In technical details, on the one hand, ML can assist in solving computationally difficult optimization problems because resource allocation optimization can be seen as a mapping from parameters to decisions. This data-to-decision mapping benefits from the powerful function-fitting ability of supervised learning based on deep neural networks, where labeled data-decision pairs are generated by well-behaved artifact solution methods. On the other hand, ML can treat the utility function of a resource allocation problem as the loss function in the training stage. This strategy allows ML to generate high-quality resource allocation decisions without relying on legacy human-made algorithms. In addition to the above two ML schemes in resource allocation, another archetype, called algorithm unrolling [18], employs neural networks to unroll existing efficient iterative algorithms. Specifically, each neural network layer acts as an iteration step of an iterative algorithm. By cascading several layers, an iteration process of the algorithm can be mimicked. This algorithm-instructed archetype is also referred to as model-driven deep learning [12], where the architecture of a deep neural network is tailored considering domain knowledge, thus improving the generalizability of the network and reducing the required size of the training data set. The fourth promising ML paradigm in resource allocation is to use reinforcement learning to explore unknown and hard-to-model environments (e.g., complex and dynamic physical transmission channels). By interacting with environments, intelligent resource allocation solutions can be learned. The four typical roles of ML in resource allocation are summarized in Figure 9. The first benefit of using ML methods is their fast computing speed in the running stage, although the training stage might be computationally heavy (when compared with the first three schemes). The second benefit of using ML methods is the ability to respond to dynamic and uncertain (even unknown) environments without explicit physical modeling (when compared with the fourth scheme, i.e., the reinforcement learning scheme).

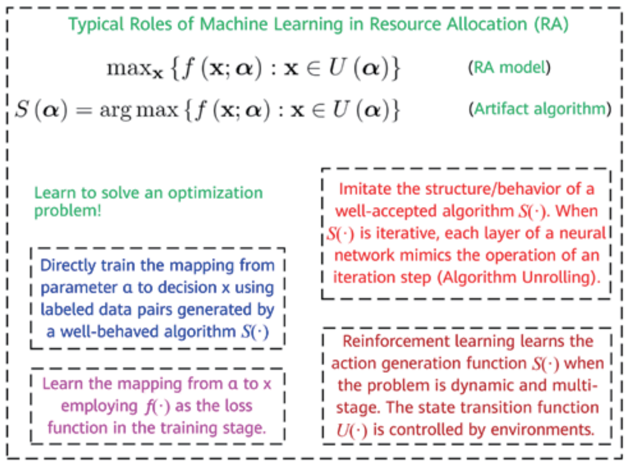

Figure 9 Four typical roles of ML in resource allocation —learning to solve optimizations.

6 Beyond Data Transmission: Sensing and Computing

Wireless communication systems are undergoing transformative changes driven by the increasing demand for low-latency and high-speed connectivity, the growing need for sensing abilities (e.g., to localize and track users) assisting high-performance communications, and the proliferation of connected devices enabling collaborative computing. This evolution has led to the development of innovative system paradigms such as ISAC [89, 105] and ICAC [106, 107], which aim to unify traditionally disparate functionalities to optimize resource usage, reduce hardware costs, and enhance overall system capabilities. For example, environmental and users' sensory data can be utilized to enhance communication performance through beam management and resource allocation, while sharing sensing data across network nodes enables real-time network monitoring and situational awareness for better sensing accuracy and larger coverage. Another example is local data processing at the edge, which can reduce latency for real-time communications, while high-speed communications enable efficient distributed computing for large-scale data analytics. As discussed in previous sections, ML (especially deep learning) techniques are indispensable in modern communication systems. These techniques offer sophisticated algorithms that can learn from vast amounts of data, and can therefore, optimize various aspects of communication networks, including resource allocation, signal processing, and fault detection. These benefits are also applicable to emerging ISAC and ICAC systems. In ISAC, deep learning models, such as CNNs and transformers, can improve sensing accuracy and robustness [108], while realizing semantic information transmission [109]. In ICAC, ML algorithms, such as federated learning, can protect users' data privacy and optimize computational tasks, facilitating efficient data processing and communication [110, 111]. In short, the synergy between ML/deep learning and the developing integrated paradigms enables more intelligent, adaptive, and efficient communication systems.

6.1 Integrated Sensing and Communication

ISAC is a paradigm that merges sensing and communication functionalities into a single system, leveraging shared infrastructure and spectral resources. This integration is essential in applications where both capabilities are crucial, such as autonomous vehicles, smart cities, and advanced surveillance systems. ISAC enhances the efficiency and performance of these systems by enabling simultaneous data acquisition and communication, thus reducing hardware costs and spectral congestion. However, this integration complicates the design of communication waveforms, the allocation of system and hardware resources, interference management, and overall network operations [112, 17]. These challenges drive the need for novel approaches to unlock the potential of ISAC systems in real-world applications. ML and deep learning techniques are, therefore, pivotal in ISAC, providing advanced data processing and decision-making capabilities. For comprehensive and motivational surveys on ML for ISAC, refer to [112, 17].

6.2 Integrated Computing and Communications

ICAC represents the convergence of computing and communication functionalities, aiming to meet the increasing computational demands of modern applications while maintaining high and robust communication performance. This integration is driven by the necessity to handle massive data processing tasks close to the source — handling tasks closer to the source helps reduce latency and improve efficiency of communications in edge computing environments, and enables intelligence of all connected devices. ICAC, which facilitates real-time data processing and analytics, is essential for applications like industrial automation, virtual reality, and the internet of things. Typical examples of ICAC include edge computing, federated learning, pervasive computing, fog computing, internet of things/vehicles, and autonomous systems. ML and deep learning are integral to ICAC, enabling dynamic resource allocation, adaptive system configurations, and real-time information analytics. These techniques ensure that computing and communication resources are utilized optimally, providing enhanced performance and responsiveness. For comprehensive and motivational surveys on ML for ICAC, refer to [16, 113, 114]. Note that, swarm intelligence and network control [94] are closely related to ICAC because controllers are, in nature, information processors (mapping the system's state signals to the system's control input signals). They are therefore ad-hoc computing modules.

7 Discussions and Conclusions

This paper discusses several pivotal aspects where ML can reform wireless communications, including but not limited to physical-layer communications, semantic communications, resource allocation, ISAC, and ICAC (e.g., federated learning, edge computing). These applications demonstrate ML's potential to upgrade various facets of communication systems, ranging from signal processing algorithms to overall network management. Nevertheless, the adoption of ML in communications is not without challenges and its role should not be overstated. Issues, such as the interpretability and troubleshooting of ML models, the need for large and rich training datasets, and the high computational resources (e.g., power, processing speed) required for training and deployment, must be addressed. In addition, particular focus should be given to the reliability and security of ML-based systems, especially in scenarios where data privacy (e.g., federated learning [115]), data freshness (e.g., few-shot learning [116, 117]), and real-time decision-making (e.g., autonomous driving) are critical. To address these challenges, the hybrid methodology, which combines the strengths of traditional physical-law models with emerging data-driven ML models, is advocated. Such a synergistic strategy can leverage the reliability and interpretability of physical mechanisms while harnessing the adaptability and learning capabilities of ML, thus enhancing overall communication system performance; see Figure 10 for features, challenges, and future considerations of intelligent transmission and processing. Among all the challenges that we can imagine, the following three items are crucial in real-world operations because they are the minimum requirements for implementing ML-based communications systems:

- How do we interpret the performance gains and failures of machine-learned models, and how do we troubleshoot and repair failures when systems are down, thus improving the overall reliability of systems? In this sense, the paradigm in Figure 6 is more reliable and manageable than that in Figure 2.
- How do we use practically limited data for better generalization and how do we integrate newly available data to improve the generalization capability [20, 31]? This consideration also includes determining how to quickly adapt the learned model to new data, for example, when the environment's data-generating laws change over time [62, 117]. In ML terminologies, data freshness, sample efficiency, and data-distributional robustness are closely related to this issue.
- How do we build domain-knowledge-informed ML models (beyond general-purpose deep neural networks such as multi-layer perception) and design computationally efficient training algorithms (beyond popular stochastic gradient descent) to diminish response times and power consumption [12, 108, 118, 119]? In addition, how do we reduce the model sizes (especially those of deep neural networks) to save storage space [120]? The three considerations above are particularly vital for embedded and edge devices.

In summary, the convergence of ML and communication systems marks a significant technological advancement, which offers the possibility for more intelligent, efficient, and reliable communication networks.

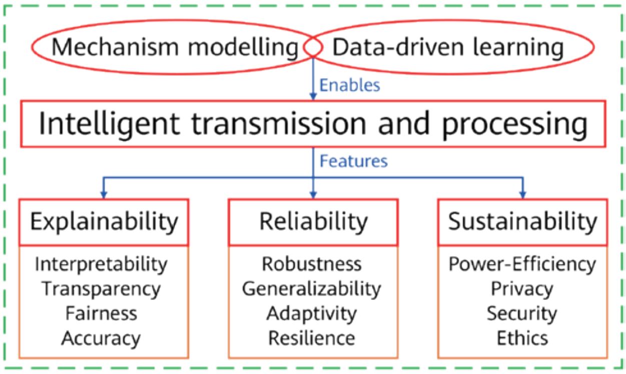

Figure 10 Features, challenges, and considerations of intelligent transmission and processing (in Figure 3) Although data-driven ML is powerful, mechanism modeling (including discovering physical/mathematical laws) is always important to improve explainability, reliability, and sustainability.

References

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[3] G. L. Stüber and G. L. Steuber, *Principles of Mobile Communication*, 4th ed. Springer, 2017.

[4] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.

[5] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2016.

[6] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 93–99, 2019.

[7] D. Gündüz, P. De Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. Van der Schaar, "Machine learning in the air," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2184–2199, 2019.

[8] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to pareto-optimal wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1472–1514, 2020.

[9] W. Yu, F. Sohrabi, and T. Jiang, "Role of deep learning in wireless communications," *IEEE BITS the Information Theory Magazine*, vol. 2, no. 2, pp. 56–72, 2022.

[10] A. Alhammadi, I. Shayea, A. A. El-Saleh, M. H. Azmi, Z. H. Ismail, L. Kouhalvandi, and S. A. Saad, "Artificial intelligence in 6G wireless networks: Opportunities, applications, and challenges," *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 8845070, 2024.

[11] A. Celik and A. M. Eltawil, "At the dawn of generative AI era: A tutorial-cum-survey on new frontiers in 6G wireless intelligence," *IEEE Open Journal of the Communications Society*, 2024.

[12] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Model-driven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, 2019.

[13] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[14] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic transmission via revising modules in conventional communications," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 28–34, 2023.

[15] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep learning-based wireless resource allocation with application to vehicular networks," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 341–356, 2019.

[16] M. A. Ferrag, O. Friha, B. Kantarci, N. Tihanyi, L. Cordeiro, M. Debbah, D. Hamouda, M. Al-Hawawreh, and K.-K. R. Choo, "Edge learning for 6G-enabled internet of things: A comprehensive survey of vulnerabilities, datasets, and defenses," *IEEE Communications Surveys & Tutorials*, 2023.

[17] S. Lu, F. Liu, Y. Li, K. Zhang, H. Huang, J. Zou, X. Li, Y. Dong, F. Dong, J. Zhu et al., "Integrated sensing and communications: Recent advances and ten open challenges," *IEEE Internet of Things Journal*, 2024.

[18] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, 2021.

[19] N. Shlezinger and T. Routtenberg, "Discriminative and generative learning for the linear estimation of random signals [lecture notes]," *IEEE Signal Processing Magazine*, vol. 40, no. 6, pp. 75–82, 2023.

[20] S. Wang, W. Dai, and G. Y. Li, "Distributionally robust receive beamforming," *arXiv preprint arXiv:2401.12345*, 2024.

[21] G. James, D. Witten, T. Hastie, R. Tibshirani et al., *An Introduction to Statistical Learning*, 2nd ed. Springer, 2021.

[22] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3072–3108, 2019.

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

[24] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465–499, 2023.

[25] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1702–1715, 2020.

[26] B. Thuraisingham, "Trustworthy machine learning," *IEEE Intelligent Systems*, vol. 37, no. 1, pp. 21–24, 2022.

[27] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

[28] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Leanpub, 2020.

[29] K. Kawaguchi, Z. Deng, K. Luh, and J. Huang, "Robustness implies generalization via data-dependent generalization bounds," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 866–10 894.

[30] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and S. Y. Philip, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.

[31] S. Wang and H. Wang, "Distributional robustness bounds generalization errors," *arXiv preprint arXiv:2212.09962*, 2024.

[32] A. Van Wynsberghe, "Sustainable AI: AI for sustainability and the sustainability of AI," *AI and Ethics*, vol. 1, no. 3, pp. 213– 218, 2021.

[33] S. AbdulRahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5476–5497, 2020.

[34] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3579–3605, 2021.

[35] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3133–3143, 2020.

[36] H. Ye, G. Y. Li, and B.-H. Juang, "Deep learning based end-to-end wireless communication systems without pilots." *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 3, pp. 702–714, 2021.

[37] S. Manouchehri, J. Haghighat, M. Eslami, and W. Hamouda, "A delay-efficient deep learning approach for lossless turbo source coding," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6704–6709, 2022.

[38] H. Ye, L. Liang, and G. Y. Li, "Circular convolutional autoencoder for channel coding," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[39] Y. Zhang, H. Wu, and M. Coates, "On the design of channel coding autoencoders with arbitrary rates for ISI channels," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 426–430, 2021.

[40] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Deep joint source-channel coding for wireless image retrieval," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5070–5074.

[41] M. Yang, C. Bian, and H.-S. Kim, "Deep joint source-channel coding for wireless image transmission with OFDM," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[42] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114– 117, 2017.

[43] S. Mohammadzadeh, V. H. Nascimento, R. C. de Lamare, and N. Hajarolasvadi, "Robust beamforming based on complex-valued convolutional neural networks for sensor arrays," *IEEE Signal Processing Letters*, vol. 29, pp. 2108–2112, 2022.

[44] A. M. Elbir, K. V. Mishra, M. R. B. Shankar, and B. Ottersten, "A family of deep learning architectures for channel estimation and hybrid beamforming in multi-carrier mm-wave massive MIMO," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 642–656, 2022.

[45] D. d. S. Brilhante, J. C. Manjarres, R. Moreira, L. de Oliveira Veiga, J. F. de Rezende, F. Müller, A. Klautau, L. Leonel Mendes, and F. A. P. de Figueiredo, "A literature survey on AI-aided beamforming and beam management for 5G and 6G systems," *Sensors*, vol. 23, no. 9, p. 4359, 2023.

[46] N. Shlezinger, M. Ma, O. Lavi, N. T. Nguyen, Y. C. Eldar, and M. Juntti, "Artificial intelligence-empowered hybrid multiple-input/multiple-output beamforming: Learning to optimize for high-throughput scalable MIMO," *IEEE Vehicular Technology Magazine*, 2024.

[47] S. H. Lim, S. Kim, B. Shim, and J. W. Choi, "Deep learning-based beam tracking for millimeter-wave communications under mobility," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7458–7469, 2021.

[48] F. Sohrabi, T. Jiang, W. Cui, and W. Yu, "Active sensing for communications by learning," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1780–1794, 2022.

[49] Y. Wei, Z. Zhong, and V. Y. Tan, "Fast beam alignment via pure exploration in multi-armed bandits," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3264–3279, 2022.

[50] W. Yi, W. Zhiqing, and F. Zhiyong, "Beam training and tracking in mmWave communication: A survey," *China Communications*, 2024.

[51] Q. Hu, F. Gao, H. Zhang, S. Jin, and G. Y. Li, "Deep learning for channel estimation: Interpretation, performance, and comparison," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2398–2412, 2020.

[52] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2827–2840, 2020.

[53] B. Ozpoyraz, A. T. Dogukan, Y. Gevez, U. Altun, and E. Basar, "Deep learning-aided 6G wireless networks: A comprehensive survey of revolutionary PHY architectures," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1749– 1809, 2022.

[54] N. Ye, S. Miao, J. Pan, Q. Ouyang, X. Li, and X. Hou, "Artificial intelligence for wireless physical-layer technologies (AI4PHY): A comprehensive survey," *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[55] Y. Yang, S. Mandt, L. Theis et al., "An introduction to neural data compression," *Foundations and Trends® in Computer Graphics and Vision*, vol. 15, no. 2, pp. 113–200, 2023.

[56] S. Han, J. Oh, K. Oh, and J. Ha, "Deep-learning for breaking the trapping sets in low-density parity-check codes," *IEEE Transactions on Communications*, vol. 70, no. 5, pp. 2909– 2923, 2022.

[57] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Turbo autoencoder: Deep learning based channel codes for point-to-point communication channels," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[58] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1182– 1192.

[59] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[60] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2645–2657, 2023.

[61] E. Bobrov, D. Kropotov, H. Lu, and D. Zaev, "Massive MIMO adaptive modulation and coding using online deep learning algorithm," *IEEE Communications Letters*, vol. 26, no. 4, pp. 818–822, 2021.

[62] N. Van Huynh and G. Y. Li, "Transfer learning for signal detection in wireless networks," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2325–2329, 2022.

[63] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1581–1592, 2019.

[64] M. Chu, A. Liu, V. K. Lau, C. Jiang, and T. Yang, "Deep reinforcement learning based end-to-end multiuser channel prediction and beamforming," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 271–10 285, 2022.

[65] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast beamforming design via deep learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 1065–1069, 2019.

[66] P. Ramezanpour, M. J. Rezaei, and M. R. Mosavi, "Deep learning-based beamforming for rejecting interferences," *IET Signal Processing*, vol. 14, no. 7, pp. 467–473, 2020.

[67] K. Chen, C. Qi, C.-X. Wang, and G. Y. Li, "Beam training and tracking for extremely large-scale MIMO communications," *IEEE Transactions on Wireless Communications*, 2023.

[68] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8017– 8045, 2022.

[69] J. Guo, T. Chen, S. Jin, G. Y. Li, X. Wang, and X. Hou, "Deep learning for joint channel estimation and feedback in massive MIMO systems," *Digital Communications and Networks*, vol. 10, no. 1, pp. 83–93, 2024.

[70] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Dual CNN-based channel estimation for MIMO-OFDM systems," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5859– 5872, 2021.

[71] R. Shankar, "Bi-directional LSTM based channel estimation in 5G massive MIMO OFDM systems over TDL-C model with rayleigh fading distribution," *International Journal of Communication Systems*, vol. 36, no. 16, p. e5585, 2023.

[72] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2022.

[73] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.

[74] Z. Lu, R. Li, K. Lu, X. Chen, E. Hossain, Z. Zhao, and H. Zhang, "Semantics-empowered communications: A tutorialcum- survey," *IEEE Communications Surveys & Tutorials*, 2023.

[75] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," *IEEE Communications Surveys & Tutorials*, 2024.

[76] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.

[77] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of Artificial Intelligence*, pp. 603–649, 2020.

[78] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[79] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Nature, 2022.

[80] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 230–244, 2022.

[81] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1394–1398, 2022.

[82] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2020.

[83] W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Cao, and K. B. Letaief, "Semantic communication meets edge intelligence," *IEEE Wireless Communications*, vol. 29, no. 5, pp. 28–35, 2022.

[84] H. Tong, Z. Yang, S. Wang, Y. Hu, O. Semiari, W. Saad, and C. Yin, "Federated learning for audio semantic communication," *Frontiers in Communications and Networks*, vol. 2, p. 734402, 2021.

[85] Z. Han and K. R. Liu, *Resource Allocation for Wireless Networks: Basics, Techniques, and Applications*. Cambridge University Press, 2008.

[86] Y. Teng, M. Liu, F. R. Yu, V. C. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2134–2168, 2018.

[87] R. Zheng and C. Hua, *Sequential Learning and Decision- Making in Wireless Resource Management*. Springer, 2016.

[88] E. Hossain, M. Rasti, and L. B. Le, *Radio Resource Management in Wireless Networks: An Engineering Approach*. Cambridge University Press, 2017.

[89] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[90] F. Dong, F. Liu, Y. Cui, W. Wang, K. Han, and Z. Wang, "Sensing as a service in 6G perceptive networks: A unified framework for ISAC resource allocation," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3522–3536, 2022.

[91] A. Liu, Z. Huang, M. Li, Y. Wan, W. Li, T. X. Han, C. Liu, R. Du, D. K. P. Tan, J. Lu et al., "A survey on fundamental limits of integrated sensing and communication," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 994–1034, 2022.

[92] S. Wang, W. Dai, H. Wang, and G. Y. Li, "Robust waveform design for integrated sensing and communication," *IEEE Transactions on Signal Processing*, 2024.

[93] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[94] X. Ge, F. Yang, and Q.-L. Han, "Distributed networked control systems: A brief overview," *Information Sciences*, vol. 380, pp. 117–131, 2017.

[95] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2131–2165, 2021.

[96] H. Djigal, J. Xu, L. Liu, and Y. Zhang, "Machine and deep learning for resource allocation in multi-access edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2449–2494, 2022.

[97] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[98] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282–2292, 2019.

[99] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3163–3173, 2019.

[100] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141– 5152, 2019.

[101] X. Xiong, K. Zheng, L. Lei, and L. Hou, "Resource allocation based on deep reinforcement learning in IoT edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 6, pp. 1133–1146, 2020.

[102] K. Xu, N. Van Huynh, and G. Y. Li, "Distributed-training-and-execution multi-agent reinforcement learning for power control in HetNet," *IEEE Transactions on Communications*, 2023.

[103] M. Lee, G. Yu, and G. Y. Li, "Learning to branch: Accelerating resource allocation in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 958–970, 2019.

[104] ——, "Graph embedding-based wireless link scheduling with few training samples," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2282–2294, 2020.

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

[106] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 9–39, 2023.

[107] D. Wen, X. Li, Y. Zhou, Y. Shi, S. Wu, and C. Jiang, "Integrated sensing-communication-computation for edge artificial intelligence," *IEEE Internet of Things Magazine*, vol. 7, no. 4, pp. 14–20, 2024.

[108] B. Zhang and G. Y. Li, "White-box 3D-OMP-transformer for ISAC," *arXiv preprint arXiv:2407.02251*, 2024.

[109] B. Zhang, Z. Qin, and G. Y. Li, "Compression ratio learning and semantic communications for video imaging," *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[110] S. Zhou and G. Y. Li, "FedGiA: An efficient hybrid algorithm for federated learning," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1493–1508, 2023.

[111] ——, "Federated learning via inexact ADMM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9699–9708, 2023.

[112] U. Demirhan and A. Alkhateeb, "Integrated sensing and communication for 6G: Ten key machine learning roles," *IEEE Communications Magazine*, vol. 61, no. 5, pp. 113–119, 2023.

[113] S. H. Alsamhi, A. V. Shvetsov, S. Kumar, J. Hassan, M. A. Alhartomi, S. V. Shvetsova, R. Sahal, and A. Hawbani, "Computing in the sky: A survey on intelligent ubiquitous computing for UAV-assisted 6G networks and industry 4.0/5.0," *Drones*, vol. 6, no. 7, p. 177, 2022.

[114] V. A. Nugroho and B. M. Lee, "A survey of federated learning for mmWave massive MIMO," *IEEE Internet of Things Journal*, 2024.

[115] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 487–500, 2022.

[116] O. Wang, S. Zhou, and G. Y. Li, "Few-shot learning for new environment adaptation," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 351– 356.

[117] O. Wang, J. Gao, and G. Y. Li, "Learn to adapt to new environments from past experience and few pilot blocks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 2, pp. 373–385, 2022.

[118] Y. Liu, Z. Qin, and G. Y. Li, "Energy-efficient distributed spiking neural network for wireless edge intelligence," *IEEE Transactions on Wireless Communications*, 2024.

[119] O. Wang, S. Zhou, and G. Y. Li, "BADM: Batch ADMM for deep learning," *arXiv preprint arXiv:2407.01640*, 2024.

[120] H. Cai, C. Gan, L. Zhu, and S. Han, "Tinytl: Reduce memory, not parameters for efficient on-device learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 285– 11 297, 2020.