Revisiting Outage for Edge Inference Systems

Zhanwei Wang, *Graduate Student Member, IEEE*, Qunsong Zeng, *Member, IEEE*,
Haotian Zheng, *Graduate Student Member, IEEE*, and Kaibin Huang, *Fellow, IEEE*

***Abstract*—One of the key missions of sixth-generation (6G) mobile networks is to deploy large-scale artificial intelligence (AI) models at the network edge to provide remote-inference services for edge devices. The resultant platform, known as edge inference, will support a wide range of Internet-of-Things applications, such as autonomous driving, industrial automation, and augmented reality. Given the mission-critical and time-sensitive nature of these tasks, it is essential to design edge inference systems that are both reliable and capable of meeting stringent end-to-end (E2E) latency constraints. Existing studies, which primarily focus on communication reliability as characterized by channel outage probability, may fail to guarantee E2E performance, specifically in terms of E2E inference accuracy and latency. To address this limitation, we propose a theoretical framework that introduces and mathematically characterizes the inference outage (InfOut) probability, which quantifies the likelihood that the E2E inference accuracy falls below a target threshold. Under an E2E latency constraint, this framework establishes a fundamental tradeoff between communication overhead (i.e., uploading more feature elements) and inference reliability as quantified by the InfOut probability. To find a tractable way to optimize this tradeoff, we derive accurate surrogate functions for InfOut probability by applying a Gaussian approximation to the distribution of the received discriminant gain. Experimental results demonstrate the superiority of the proposed design over conventional communication-centric approaches in terms of E2E inference reliability.**

***Index Terms*—Edge inference, outage probability, feature selection, computation-communication tradeoff.**

I. Introduction

One mission of *sixth-generation* (6G) mobile networks is the widespread deployment of pre-trained *artificial intelligence* (AI) models at the network edge to support ubiquitous and real-time intelligent services [1]–[4]. This emerging paradigm, known as edge inference, will serve as a platform for deploying next-generation Internet-of-Things applications, ranging from autonomous driving to industrial automation to augmented reality [5], [6]. In such a system, features extracted from sensing data are transmitted from an edge device to an edge server for remote inference using a large-scale AI model. Given that many relevant tasks are mission-critical and time-sensitive [7], [8], it is essential to develop latency-constrained edge inference systems with guaranteed performance. A primary challenge in designing such systems is the unreliable wireless links connecting edge servers and devices, as their outage events can disrupt operations and degrade performance. To address this challenge, we propose a theoretical framework in which we introduce and mathematically characterize the new definition of *inference outage* (InfOut) probability. This framework establishes a fundamental *communication-computation* (C^2) tradeoff, which is then optimized to design novel feature transmission schemes that minimize the InfOut probability.

Since fading is a fundamental characteristic of wireless channels, ensuring reliability has been a primary concern in the design of wireless communication systems from their inception. Outage probability, defined as the likelihood of a wireless link disconnecting due to deep fades, serves as a basic metric of reliability [9]. One avenue of research in reliable communications involves mathematically characterizing link-level outage probability using abstract channel models, such as space diversity [10]–[12], and Nakagami fading channels [13], [14]. This research targets various systems and scenarios, e.g., multi-hop transmission [13] and inaccurate channel estimation [15]. From the perspective of reliable network design, researchers have introduced the concept of network outage probability, which generalizes outage probability to account for variations in links across a network [16]. Stochastic geometry has been adopted as a tractable tool for deriving analytical expressions for network outage probability, incorporating factors such as interference, fading, and network density [17], [18]. Another vein of research focuses on designing techniques to cope with fading. When *channel state information at the transmitter* (CSIT) is available, outages can be minimized by adapting power, modulation, and coding to the time-varying channels [19]–[22]. However, these adaptive approaches require accurate channel estimation and feedback, which incur additional communication overhead and latency, as well as require more complex transmitter hardware. They may not be feasible in fast-fading scenarios. When CSIT is unavailable, communication reliability can be ensured through repeated transmissions using the basic protocol of *Automatic Repeat reQuest* (ARQ).

The retransmission approach has notable drawbacks, including increased communication latency and higher channel usage. These limitations create challenges in supporting mission-critical tasks with stringent deadlines, prompting the development of new techniques in the 4G and 5G eras. In particular, the breakthroughs in *multiple-input multiple-output* (MIMO) communications have enabled the leveraging of space diversity to mitigate channel fading [23]. Researchers have explored the fundamental tradeoff between reducing outage probability

Z. Wang, Q. Zeng, H. Zheng, and K. Huang are with the Department of Electrical and Computer Engineering, The University of Hong Kong, Hong Kong SAR, China (Email: {zhanweiw, qszeng, htzheng, huangkb}@eee.hku.hk). Corresponding authors: K. Huang; Q. Zeng.

The work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515011747), Research Grants Council of the Hong Kong Special Administrative Region, China under a fellowship award (HKU RFS2122-7S04), NSFC/RGC Collaborative Research Scheme (CRS_HKU702/24), the Areas of Excellence scheme grant (AoE/E-601/22-R), Collaborative Research Fund (C1009-22G), and the Grants 17212423 & 17304925, and in part by the Croucher Senior Fellowship and Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) (SGDX20230821091559018).

through spatial diversity and increasing transmission rates via spatial multiplexing, a relationship known as the diversity-multiplexing tradeoff [24]. The demand for 5G systems to support *ultra-reliable and low-latency communication* (URLLC) has led to the adoption of *short packet transmission* (SPT). However, the inherent conflict between achieving URLLC and maintaining high data rates means that SPT is typically suited only for low-rate, mission-critical tasks, such as transmitting control commands and basic sensing data, including humidity, temperature, and pollution levels [25], [26]. In this context, outages, measured by packet decoding error probability, are addressed by developing advanced SPT techniques, such as non-coherent transmission [27], optimal framework structures [28], power control [29], and wireless power transfer [30]. Despite these advancements, approaches designed for low-rate tasks face difficulties in ensuring the reliability of 6G edge inference systems that require data-intensive communication, such as the transmission of high-dimensional features.

In 6G, edge inference systems are designed to provide a platform for delivering remote inference services to support AI-enabled mobile applications such as sensing, autonomous driving, and robotic control [31], [32]. As edge inference systems represent the natural convergence of AI and communication, evaluating their reliable performance requires the generalization of traditional channel outage probability to the likelihood of failing to achieve a target *end-to-end* (E2E) inference accuracy, termed InfOut probability. A mathematical study of this performance metric has not been extensively explored in the literature, as existing work has primarily focused on developing goal-oriented techniques aimed at improving the E2E performance of edge inference systems in the presence of channel distortion, as described shortly. One popular architecture for these systems, known as split inference, balances the computational load between devices and servers by flexibly dividing a pre-trained AI model into a low-complexity device sub-model for data-feature extraction and a server-side sub-model for remote inference [33]. Existing split-inference techniques can optimize the accuracy-latency tradeoff [34], support scalable over-the-air data aggregation [35], and employ progressive feature transmission to ensure high reliability [36]. Additionally, for latency-sensitive applications, ultra-low-latency edge inference systems have been developed based on short-packet feature transmission [37] and leveraging the robustness of AI models to cope with channel distortion [38]. Another popular architecture for edge inference systems is joint source and channel coding (JSCC), which exploits the auto-encoder architecture to jointly train models for inference and channel coding to overcome channel noise, thereby achieving high E2E inference accuracy [39]. A range of relevant research has been conducted, including task-specific analog-to-digital converters [40], deep learning based JSCC [41], and explainable JSCC utilizing semantic channel capacity bounds [42]. Despite extensive efforts on algorithm designs, there is a lack of in-depth mathematical studies on the reliability of edge inference systems despite its being a fundamental topic. Results from such studies can provide performance guarantees by quantifying the worst-case inference accuracy and guide new breakthroughs in reliable edge inference, which motivates the current work.

The proposed InfOut probability depends on the interplay of randomness in propagation and the performance of *deep neural networks* (DNNs) [43]. The latter is influenced by dynamic variations in device computing capacities, model parameters, data inputs, and other factors. The reliability of DNN performance, specifically, has been investigated in the field of computer science and is typically assessed using Monte Carlo sampling [44]. Based on this approach, the worst-case inference performance, measured by the *k-th percentile performance* (KPP), is quantified and subsequently enhanced through model training [45]. On the other hand, the reliability of DNN models under attacks has been studied by evaluating the proportion of adversarial examples that successfully induce incorrect model predictions [46]. The existing studies assume a stand-alone computing process within a device or server, where wireless propagation is hence irrelevant. In contrast, in this work, we consider latency-constrained edge inference systems over wireless links, where the reliability issue is exacerbated by additional factors such as fading and the imposition of E2E latency constraints on resource-constrained devices. Targeting such a system performing a remote classification task, the InfOut probability is defined as the probability that the E2E inference accuracy falls below a predefined threshold. Then this work presents a theoretical framework to characterize this probability. The key contributions and findings are summarized as follows.

- **Analysis of Inference Outage Probability:** For analytical tractability, we assume linear classification with feature vectors extracted by the device following a *Gaussian mixture model* (GMM) [47], [48]. The derived results are subsequently extended and validated to design outage-minimization schemes for the more complex case of *convolutional neural network* (CNN) classification with a general feature distribution. A key step in the analysis is to upper bound the InfOut probability by the probability that the *discriminant gain* (DG) of channel-distorted features, which are received at the server, falls below a threshold. Based on this bound, we derive a tractable surrogate function for characterizing the InfOut probability by showing that the received DG for a high-dimensional feature space can be suitably modeled as a Gaussian random variable according to the Lindeberg-Feller central limit theorem. The ensuing analysis of the said bound reveals a C^2 tradeoff. Specifically, consider two parameters: the number of input samples for a single inference operation and the number of uploaded features for each sample. Increasing one parameter while keeping the other fixed can incur higher E2E latency but reduce the InfOut probability, and vice versa. This finding motivates the following parametric optimization to balance this tradeoff.
- **Inference Outage Probability Minimization:** Directly optimizing the C^2 tradeoff to minimize the InfOut probability is intractable. We address this challenge by transforming the problem into one of maximizing a continuous and differentiable surrogate of the receive DG. This

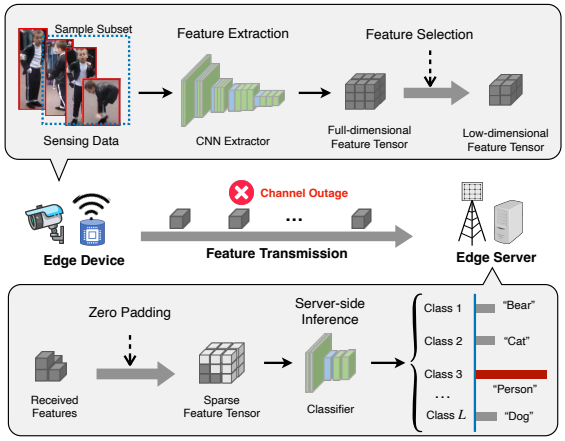


Fig. 1: The transceiver framework of the edge inference system.

surrogate is shown to be a concave function of the number of uploaded features, thereby ensuring a unique optimal solution. For the more complex case of CNN classification, we define the receive DG using the inverse Gaussian Q-function related to inference accuracy and approximate its distribution as Gaussian, a method validated with real datasets. Subsequently, we design a numerical algorithm to determine the optimal number of transmitted features per sample by learning on a training dataset and employing random masking of feature vectors.

- **Experiments:** The analytical results are validated using both synthetic (e.g., GMM) and real datasets (e.g., ModelNet [49]). The designs from optimizing the C^2 tradeoff closely match the optimal performance from a brute-force search and outperform conventional methods prioritizing target accuracy while neglecting the channel effects on E2E accuracy.

The remainder of this paper is organized as follows. Section II introduces the system model and defines performance metrics. In Section III, we present the InfOut probability analysis for the edge inference system. Outage minimization strategies for linear and CNN classification are presented in Sections IV and V, respectively. Section VI reports experimental results, while concluding remarks are provided in Section VII.

II. Models and Metrics

We consider an edge inference system, as shown in Fig. 1, where a device transmits feature vectors, extracted from observations, to an edge server for object classification. The associated models and performance metrics are discussed in the following subsections.

A. Sensing and Computation Model

In the considered edge inference system, sensing noise and potential obstructions can degrade inference performance [50]. To address this issue, we consider a sensing scenario that involves fusing feature vectors extracted from K observations[1].

[1]For instance, consider a scenario in which a device captures a series of images of a moving vehicle in an intelligent traffic monitoring system. Temporal fusion of these snapshots reduces motion blur and enhances resolution, enabling the device to transmit critical information to the server for inference [37], [51].

The fused feature vector, denoted as $\overline{\mathbf{x}}$, is obtained through average pooling:

$$\overline{\mathbf{x}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k, \tag{1}$$

where $\mathbf{x}_k \in \mathbb{R}^D$ denotes the feature vector extracted from the k-th observation. In streaming applications, feature extraction can be pipelined across consecutive observations. Let N_F denote the number of *floating-point operations* (FLOPs) required to process one observation and let f_c denote the local computation speed (in FLOPs/s). Accordingly, the per-observation feature-extraction latency is $T_{\mathsf{fea}} = N_F/f_c$. Let T_{sen} denote the sensing interval (e.g., the camera frame period). Then, under the sensing–computation pipeline, the time to obtain features from K observations is

$$T_{\mathsf{comp}} = T_{\mathsf{sen}} + T_{\mathsf{fea}} + (K-1)\max\{T_{\mathsf{sen}}, T_{\mathsf{fea}}\}. \tag{2}$$

We consider a representative high–frame-rate sensing regime in which the sensing interval T_{sen} is negligible (or observations are pre-buffered), so that computation for feature extraction becomes the bottleneck under a constrained power budget and limited computational speed. With such consideration, the computation latency in (2) reduces to

$$T_{\mathsf{comp}} = \frac{KN_F}{f_c}. \tag{3}$$

Note that this linear computation model is widely adopted for compute-limited edge devices with a single local processor [52], [53].

B. Data Distribution Model

While CNN based classification is designed for generic distributions, in the case of linear classification, we assume the following data distribution for tractability. We consider feature vectors are drawn from a *Gaussian mixture model* (GMM) [36], [48]. Let D and L denote the dimension of the feature vector and the number of classes, respectively. Specifically, each feature vector $\mathbf{x}_k$ is independently sampled from a Gaussian distribution with mean $\boldsymbol{\mu}_\ell \in \mathbb{R}^D$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{D\times D}$. The mean vector (or class centroid) varies across classes, while the covariance matrix is assumed to be identical for all classes [36]. Without loss of generality, we set $\mathbf{C} = \mathrm{diag}(C_{1,1}, C_{2,2}, \ldots, C_{D,D})$ as a diagonal matrix, which can be obtained via *principal component analysis* (PCA) [54]. The joint distribution of the K feature vectors is given by:

$$(\mathbf{x}_1, \ldots, \mathbf{x}_K) \sim \frac{1}{L}\sum_{\ell=1}^{L}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{x}_k \mid \boldsymbol{\mu}_\ell, \mathbf{C}\right), \tag{4}$$

where $\mathcal{N}\left(\mathbf{x}_k \mid \boldsymbol{\mu}_\ell, \mathbf{C}\right)$ denotes the Gaussian *probability density function* (PDF) with mean $\boldsymbol{\mu}_\ell$ and covariance matrix $\mathbf{C}$. Building on such model, the fused feature vector, defined in (1), subjects to:

$$\overline{\mathbf{x}} \sim \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{N}(\boldsymbol{\mu}_\ell, \overline{\mathbf{C}}), \tag{5}$$

where $\overline{\mathbf{C}} = \mathbf{C}/K$ with diagonal element being $\overline{C}_{d,d} = C_{d,d}/K$. Note that the fusion of K feature vectors suppresses the feature variance.

C. Communication Model

To mitigate the communication overhead from uploading high-dimensional features, we consider transmitting a subset of the fused features, denoted as $\mathcal{S} \subseteq \{1, 2, \ldots, D\}$, whose cardinality is $S = |\mathcal{S}|$. Each selected feature and its index are quantized into Q_B and Q_I bits, respectively[2], ensuring negligible quantization error. Transmitting these features occupies S time slots, each lasting T_Δ seconds. The resulting communication latency is provided as

$$T_{\mathsf{comm}} = T_\Delta S. \tag{6}$$

For each time slot $t \in \{1, 2, \ldots, S\}$, the channel outage probability, indicating the transmission failure, is expressed as

$$P_{\mathsf{out}} = \Pr(T_\Delta r_t < Q_B + Q_I). \tag{7}$$

Here, r_t denotes the transmission rate, given by

$$r_t = B_W \log_2 \left(1 + \frac{p|h_t|^2}{N_0 B_W}\right), \tag{8}$$

where p is the transmit power, B_W represents the system bandwidth, N_0 is the noise power spectrum density, and $h_t \sim \mathcal{CN}(0, \sigma^2)$ denotes the Rayleigh fading channel coefficient in the t-th slot. The channel is assumed i.i.d. varying over time slots but remaining constant throughout one time slot. For convenience, we then define the *activation probability* as $P_{\mathsf{act}} \triangleq 1 - P_{\mathsf{out}}$. Under the assumption of i.i.d. block-fading, each feature is successfully received with probability P_{act}. The successfully received feature set at the edge server is denoted as $\tilde{\mathcal{S}} \subseteq \mathcal{S}$.

D. Inference Model

We consider two classifier models based on the received feature set $\tilde{\mathcal{S}}$.

1) Linear Classification: We consider a *maximum likelihood* (ML) classifier for the distribution in (4), where the classification boundary between each pair of classes is a hyperplane in the feature space. Due to the uniform prior on the object classes, the ML classifier is equivalent to a *maximum a posteriori* (MAP) classifier. The label $\hat{\ell}$ is estimated as

$$\begin{aligned} \hat{\ell} &= \operatorname*{argmax}_{\ell} \log \Pr(\overline{\mathbf{x}}|\ell, \tilde{\mathcal{S}}) \\ &= \operatorname*{argmin}_{\ell} z_\ell(\tilde{\mathcal{S}}), \end{aligned} \tag{9}$$

where $z_\ell(\tilde{\mathcal{S}})$ is the squared Mahalanobis distance between the received features in $\tilde{\mathcal{S}}$ and the centroid of class-ℓ, given as

$$z_\ell(\tilde{\mathcal{S}}) = \sum_{d \in \tilde{\mathcal{S}}} \frac{(\overline{x}(d) - \mu_\ell(d))^2}{\overline{C}_{d,d}}. \tag{10}$$

[2]Given D dimensions to be indexed, the required bits per dimension are assumed to be fixed at $Q_I = \lceil \log_2(D) \rceil$.

Here, $\overline{x}(d)$, $\mu_\ell(d)$ and $\overline{C}_{d,d}$ represent the d-th feature of $\overline{\mathbf{x}}$, the centroid of class-ℓ and the corresponding auto-covariance, respectively. Hence, the linear classification problem reduces to finding the class label which can minimize the Mahalanobis distance.

2) CNN Classification: We also consider a more realistic but analytically intractable scenario where feature vectors are extracted from observations using a well-trained CNN model. The layers of CNN model are split into a device sub-model and a server sub-model, represented as functions $f_{\mathsf{sen}}(\cdot)$ and $f_{\mathsf{ser}}(\cdot)$, respectively. The feature vector of the k-th observation is constructed by passing the observation $\mathbf{M}_k$ of the common object through a pre-trained CNN, i.e., $\mathbf{x}_k = f_{\mathsf{sen}}(\mathbf{M}_k)$. The feature vectors are then aggregated to $\overline{\mathbf{x}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_k$ before feature selection and transmission. Upon receiving the feature elements and their associated indices, the edge server reconstructs the feature map into its original dimensionality D by zero-padding any unselected and channel-lost features. Let $\tilde{\mathbf{x}}_{\mathsf{cnn}} \in \mathbb{R}^D$ denote the output of this feature reconstruction. Subsequently, the edge server feeds the sparse feature map $\tilde{\mathbf{x}}_{\mathsf{cnn}}$ into the server sub-model, i.e., $\{c_1, \ldots, c_\ell, \ldots, c_L\} = f_{\mathsf{ser}}(\tilde{\mathbf{x}}_{\mathsf{cnn}})$, where c_ℓ represents the confidence score of the ℓ-th class. The CNN classifier then outputs the inferred label with the highest confidence score, i.e., $\hat{\ell} = \operatorname{argmax}_\ell c_\ell$.

E. Relevant Metrics

For linear and CNN classification, relevant metrics are characterized as follows.

1) Inference Outage Probability: For a classification task with K processed observations and a set of received features $\tilde{\mathcal{S}}$, the inference accuracy, denoted as $a(K, \tilde{\mathcal{S}})$, is commonly defined as the probability of correctly predicting the object label, expressed as

$$a(K, \tilde{\mathcal{S}}) = \frac{1}{L}\sum_{\ell=1}^{L} \Pr(\hat{\ell} = \ell \mid \ell, K, \tilde{\mathcal{S}}). \tag{11}$$

Due to the feature loss caused by block-fading channels, the set of received features $\tilde{\mathcal{S}}$ is random, making the inference accuracy a random variable over wireless channels. To capture the channel-induced randomness of the E2E performance, we assume that a follows the distribution $a \sim \mathcal{D}_\theta(K, \mathcal{S})$, where $\mathcal{D}$ denotes the distribution of inference accuracy, θ represents the distribution's parameter, K is the number of processed observations, and $\mathcal{S}$ is the selected feature set at the sensor. Given a target inference accuracy A_{th}, an inference outage occurs if the accuracy requirement is not met. In this context, the InfOut probability, which measures the reliability of the system, is denoted as

$$\begin{aligned} P_{\mathsf{out}}^{\mathsf{e2e}} &= \Pr(a \le A_{\mathsf{th}} \mid K, \mathcal{S}) \\ &= \sum_{\tilde{\mathcal{S}} \subseteq \mathcal{S}} \mathbb{I}(a(K, \tilde{\mathcal{S}}) \le A_{\mathsf{th}}) P(\tilde{\mathcal{S}}), \end{aligned} \tag{12}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and $P(\tilde{\mathcal{S}})$ is the *probability mass function* (PMF) of the received feature set

$\tilde{\mathcal{S}}$ at the edge server[3].

2) On-device Feature Importance: We consider two types of metrics to measure the on-device feature importance for linear and CNN classifications, respectively.

- *Discriminant Gain:* For linear classification, the pairwise DG quantifies the discernibility between two classes within a subspace of the feature space. Given the fused feature vector $\overline{\mathbf{x}}$, the DG between class ℓ and ℓ', denoted as $G_{\ell,\ell'}$, is defined as the symmetric *Kullback-Leibler* (KL) divergence [36]:

$$\begin{aligned} G_{\ell,\ell'} =& \mathsf{KL}(\mathcal{N}(\boldsymbol{\mu}_\ell, \overline{\mathbf{C}}) \,||\, \mathcal{N}(\boldsymbol{\mu}_{\ell'}, \overline{\mathbf{C}})) \\ &+ \mathsf{KL}(\mathcal{N}(\boldsymbol{\mu}_{\ell'}, \overline{\mathbf{C}}) \,||\, \mathcal{N}(\boldsymbol{\mu}_\ell, \overline{\mathbf{C}})) \\ =& (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^\mathsf{T} \overline{\mathbf{C}}^{-1} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}) \\ =& K \sum_{d=1}^{D} W_d(\ell, \ell'), \end{aligned} \tag{13}$$

  where $W_d(\ell, \ell')$ is the pair-wise DG of the d-th dimension, given as

$$W_d(\ell, \ell') = \frac{(\mu_\ell(d) - \mu_{\ell'}(d))^2}{C_{d,d}}. \tag{14}$$

  Using this metric, we quantify the importance of each feature dimension and enable the DG based feature selection scheme. Specifically, the importance of the d-th dimension is measured by the minimum DG among all class pairs, defined as

$$\hat{W}_d = \min_{\ell \neq \ell'} W_d(\ell, \ell'). \tag{15}$$

  The min-DG rule prioritizes features that strengthen the most confusable class pair, aligning with our objective of ensuring worst-case inference performance.
- *Feature Magnitude:* The DG defined in (15) for a linear classifier, which underpins the associated metric of feature importance, is not applicable to a CNN model. In this context, we adopt a magnitude based feature selection scheme, where the importance of each feature element is determined by its magnitude [55]. Given a target number of selected features, $S = |\mathcal{S}|$, the device selects the top-S features with the largest magnitudes for transmission. Since each selected feature is transmitted in a fixed-duration slot with a fixed power budget, the top-S selection does not change the per-feature transmit energy, but only the feature dimensions that are delivered.

3) E2E Inference Latency: We measure the task completion time by the E2E inference latency, which captures the latency accrued by the sequential inference pipeline. Given the computation and the feature-transmission latencies in (3) and (6), the E2E latency is given by

$$T_{\mathsf{e2e}} = T_{\mathsf{comp}} + T_{\mathsf{comm}}. \tag{16}$$

[3]While we focus on classification as a representative inference task, the considered metric readily extends to other inference tasks. Specifically, for regression, the InfOut probability can be defined as the probability that the *mean squared error* (MSE) exceeds a target threshold; for detection, it can be defined as the probability that the *intersection-over-union* (IoU) falls below a prescribed threshold.

For time-sensitive applications (e.g., collision avoidance in autonomous driving), we enforce a strict latency requirement $T_{\mathsf{e2e}} \leq T_{\max}$ where $T_{\max}$ denotes the target E2E task deadline.

III. Analysis of Inference Outage Probability

This section provides a theoretical analysis of the InfOut probability for linear classification. The derived insights are further applied to minimizing the InfOut probability in CNN based classification, as discussed in Sec. V.

A. Tractable Surrogate of Inference Accuracy

The computation of the InfOut probability in (12) requires an accurate characterization of the inference accuracy distribution. For tractability, we consider the lower bound of inference accuracy provided in Lemma 1 as its surrogate.

Lemma 1 ([37]). *The inference accuracy with K observations and received feature set $\tilde{\mathcal{S}}$, denoted as $a(K, \tilde{\mathcal{S}})$, is lower bounded by*

$$a(K, \tilde{\mathcal{S}}) \geq a_{\mathsf{low}}(K, G_{\mathsf{R}}) \triangleq 1 - (L-1) Q\left(\frac{\sqrt{K G_{\mathsf{R}}}}{2}\right), \tag{17}$$

where G_{R} is defined as the receive DG per observation:

$$G_{\mathsf{R}} = \sum_{d \in \tilde{\mathcal{S}}} \hat{W}_d. \tag{18}$$

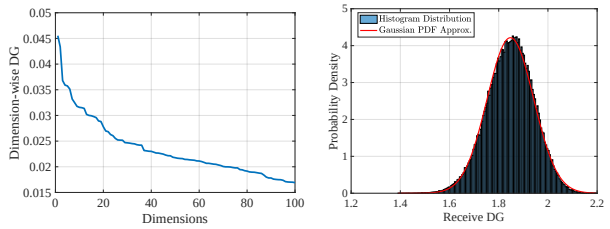
$\hat{W}_d$ is the minimum DG of the d-th feature dimension in (15).

Lemma 1 indicates that the lower bound of inference accuracy is a monotonically increasing function of the receive DG, as defined in (18). The value of G_{R} increases as the number of successfully received features, denoted as $\tilde{S} = |\tilde{\mathcal{S}}|$, grows due to the positive DG per dimension, i.e., $\hat{W}_d \geq 0$. However, feature loss caused by fading channels introduces randomness into $\tilde{\mathcal{S}}$, resulting in a distribution of the receive DG and variability in inference accuracy. This uncertainty raises reliability concerns for edge inference systems.

By leveraging the one-to-one mapping between the lower-bounded inference accuracy and the receive DG defined in (17), the InfOut probability in (12) can be upper-bounded as:

$$\begin{aligned} P_{\mathsf{out}}^{\mathsf{e2e}} &= \sum_{\tilde{\mathcal{S}} \subseteq \mathcal{S}} \mathbb{I}(a(K, \tilde{\mathcal{S}}) \leq A_{\mathsf{th}}) P(\tilde{\mathcal{S}}) \\ &= 1 - \sum_{\tilde{\mathcal{S}} \subseteq \mathcal{S}} \mathbb{I}(a(K, \tilde{\mathcal{S}}) > A_{\mathsf{th}}) P(\tilde{\mathcal{S}}) \\ &\leq 1 - \sum_{\tilde{\mathcal{S}} \subseteq \mathcal{S}} \mathbb{I}(a_{\mathsf{low}}(K, G_{\mathsf{R}}) > A_{\mathsf{th}}) P(\tilde{\mathcal{S}}) \\ &= \sum_{\tilde{\mathcal{S}} \subseteq \mathcal{S}} \mathbb{I}(K G_{\mathsf{R}} \leq G_{\mathsf{th}}) P(\tilde{\mathcal{S}}) \\ &= \Pr(K G_{\mathsf{R}} \leq G_{\mathsf{th}}), \end{aligned} \tag{19}$$

where $G_{\mathsf{th}} \triangleq 4\left(Q^{-1}\left(\frac{1-A_{\mathsf{th}}}{L-1}\right)\right)^2$ denotes the required DG threshold to achieve an inference accuracy of A_{th}. The result in (19) demonstrates that the receive DG, G_{R}, can serve as a tractable surrogate for inference accuracy.



(a) Discriminant Gain ($\hat{W}_d$) (b) Gaussian Approximation

Fig. 2: Considering dimension-wise DG $\hat{W}_d$ sorted in descending order in subfigure (a), the PDF comparison between the real distribution and the Gaussian approximation is presented in subfigure (b). The number of selected features is set at $S = 100$ under an activation probability of $P_{\mathsf{act}} = 0.8$.

B. Inference Outage Probability

Without loss of generality, we assume that the DG values are arranged in decreasing order after PCA, i.e., $\hat{W}_d \geq \hat{W}_{d+1}, d = 1, \cdots, D-1$. We consider a DG based feature selection scheme that selects the top-S features with the highest DG values. In such manner, the receive DG in (18) can be rewritten as

$$G_{\mathsf{R}} = \sum_{d \in \tilde{\mathcal{S}}} \hat{W}_d = \sum_{d=1}^{S} \hat{W}_d I_d, \tag{20}$$

where I_d is an indicator representing the successful transmission of the d-th feature dimension over fading channels. We assume i.i.d. fading across feature slots (e.g., via time–frequency interleaving beyond the channel coherence time and bandwidth), such that $\{I_d\}$ are i.i.d. Bernoulli:

$$I_d = \begin{cases} 1, & \text{with probability of} \quad P_{\mathsf{act}}, \\ 0, & \text{with probability of} \quad 1 - P_{\mathsf{act}}. \end{cases} \tag{21}$$

Notably, G_{R} is the weighted sum of i.i.d Bernoulli random variables, with its mean and variance given by

$$\begin{aligned} \mathbb{E}[G_{\mathsf{R}}] &= P_{\mathsf{act}} G_1(S), \\ \mathrm{Var}(G_{\mathsf{R}}) &= (1 - P_{\mathsf{act}}) P_{\mathsf{act}} G_2(S), \end{aligned} \tag{22}$$

where $G_1(S)$ and $G_2(S)$ are functions of the number of selected features S:

$$G_1(S) = \sum_{d=1}^{S} \hat{W}_d, \quad G_2(S) = \sum_{d=1}^{S} \hat{W}_d^2. \tag{23}$$

Here, we refer to $G_1(S)$ as *transmit DG*, which quantifies the DG of selected features at the sensor side. Meanwhile, $G_2(S)$ represents the sum of squared dimension-wise DGs, termed the *transmit DG power*.

To characterize the distribution of G_{R}, we show that the weighted sum of i.i.d. Bernoulli random variables in (20) satisfies Lindeberg's condition, as stated in Lemma 2.

Lemma 2 (Lindeberg's Condition [56])**.** *Let $X_d = \hat{W}_d I_d$, $d = 1, 2, \ldots, S$ be independent random variables with mean $\mu_{X_d} = \hat{W}_d P_{\mathsf{act}}$ and variance $\mathrm{Var}(X_d) = P_{\mathsf{act}}(1 - P_{\mathsf{act}})\hat{W}_d^2$. For any $\epsilon > 0$, the Lindeberg condition holds:*

$$\lim_{S \to \infty} \frac{1}{\sigma_{\mathsf{G}}^2(S)} \sum_{d=1}^{S} \mathbb{E}\left[(X_d - \mu_{X_d})^2 \mathbb{I}\left(|X_d - \mu_{X_d}| > \epsilon \sigma_{\mathsf{G}}(S) \right) \right] = 0, \tag{24}$$

where $\sigma_{\mathsf{G}}^2(S) = \sum_{d=1}^{S} Var(X_d) = P_{\mathsf{act}}(1 - P_{\mathsf{act}}) \sum_{d=1}^{S} \hat{W}_d^2$ denotes the aggregate variance.

The proof is provided in Appendix A.

Consequently, the receive DG can be approximated by a Gaussian distribution using the Lindeberg-Feller Central Limit Theorem, as provided in Lemma 3.

Lemma 3 (Distribution of Receive DG)**.** *If the Lindeberg condition in Lemma 2 holds, then for a sufficiently large number of selected features S (a typical scenario in DNNs), the distribution of receive DG in* (20) *can be approximated as*

$$G_{\mathsf{R}} \to \mathcal{N}\left(\mathbb{E}[G_{\mathsf{R}}], Var(G_{\mathsf{R}}) \right), \quad \text{weakly as} \quad S \to \infty, \tag{25}$$

where $\mathbb{E}[G_{\mathsf{R}}]$ and $Var(G_{\mathsf{R}})$ are the mean and variance of the distribution, provided in (22)*.*

To illustrate the approximation, Fig. 2 shows the statistics of the receive DG, computed using 100 dimensions with an activation probability of $P_{\mathsf{act}} = 0.8$. It can be observed that the approximation closely matches the empirical distribution. Using this approximation, the upper bound of the InfOut probability in (19) can be expressed as

$$P_{\mathsf{out}}^{\mathsf{e2e}} \leq \Pr(K G_{\mathsf{R}} \leq G_{\mathsf{th}}) \tag{26}$$

$$\approx Q\left(\frac{P_{\mathsf{act}} G_{\mathsf{f}}(S) - \frac{G_{\mathsf{th}}}{K\sqrt{G_2(S)}}}{\sqrt{P_{\mathsf{act}}(1 - P_{\mathsf{act}})}} \right), \tag{27}$$

where $Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$ denotes the Q-function, and $G_{\mathsf{f}}(S) = \frac{G_1(S)}{\sqrt{G_2(S)}}$ is a monotone-increasing function of the number of selected features S (see Appendix B).

Remark 1 (Computation-communication tradeoff)**.** *The result in* (27) *theoretically shows that the InfOut probability can be reduced by increasing the number of selected features and/or the number of processed observations. However, this reduction comes at the cost of increased communication and/or computation latency, respectively. Setting a strict deadline for E2E latency introduces a competition between computation and communication: allocating more time for processing more observations produces a higher quality feature vector, while fewer features can be uploaded in the reduced time available for communication, and vice versa. Thus, a fundamental* communication-computation *(C^2) tradeoff emerges.*

Fig. 3 validates the C^2 tradeoff controlled by the number of selected features under a latency constraint of 10 ms. The InfOut probability initially decreases and subsequently increases as the number of selected features grows. Additionally, a more reliable channel (with higher activation probability P_{act}) achieves a lower InfOut probability, highlighting the interplay between channel outage and inference outage.

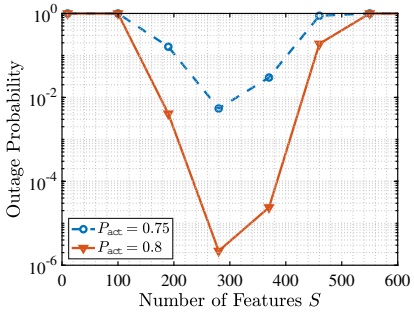


Fig. 3: InfOut probability under different channel outage probabilities. The numerical settings of a binary classifier are $A_{\text{th}} = 99\%$, $T_\Delta = 10\ \mu$s, $T_{\max} = 10$ ms, $f_c = 2.5$ GFLOPs/s, and $N_F = 2.3$ MFLOPs.

IV. Minimization of Inference Outage Probability

In this section, we enhance the reliability of latency-constrained edge inference systems by optimizing the C^2 tradeoff described in Remark 1. The identified C^2 tradeoff is governed by the number of processed observations and transmitted features under E2E latency constraints. To minimize the InfOut probability while satisfying the latency requirement, these control variables are jointly optimized. The resulting optimization problem is formulated as

$$\min_{S,K} \quad Q\left(\frac{P_{\text{act}}G_{\text{f}}(S) - \frac{G_{\text{th}}}{K\sqrt{G_2(S)}}}{\sqrt{P_{\text{act}}(1-P_{\text{act}})}}\right) \tag{28a}$$

$$\text{s.t.} \quad \frac{KN_F}{f_c} + T_\Delta S \le T_{\max}, \tag{28b}$$

$$S \in \{1, 2, \ldots, D\}, \tag{28c}$$

$$K \in \{1, 2, \ldots, K_{\max}\}, \tag{28d}$$

where $K_{\max}$ denotes the maximum number of observations of the object. Constraint (28b) accounts for the device-side feature extraction and uplink transmission latency. The server-side inference latency is assumed to be negligible due to sufficient computing resources at the edge server, or treated as a constant that can be absorbed into the E2E deadline. To solve problem (28), we approximate the objective using a surrogate function derived from the lower bounds on the transmit DG $G_1(S)$ and its associated power $G_2(S)$. Subsequently, the optimal number of selected features is determined under the DG based feature selection scheme.

A. Lower Bounds on Transmit DG

The impact of the number of selected features on the objective of Problem (28) is captured by two discrete functions, $G_1(S)$ and $G_2(S)$, where $S \in \{1, 2, \ldots, D\}$. To facilitate the analysis of these functions, we derive their lower bounds by leveraging integrals of the continuous and differentiable DG function, as defined in Definition 1.

Definition 1 (Discriminant Gain Function)**.** *The DG function is defined as a continuous and differentiable function of dimension index $t \in [0, D]$, given as*

$$g(t) = \frac{\hat{W}_d - \hat{W}_{d+1}}{2}\cos(\pi(t-d+1)) + \frac{\hat{W}_d + \hat{W}_{d+1}}{2}, \tag{29}$$

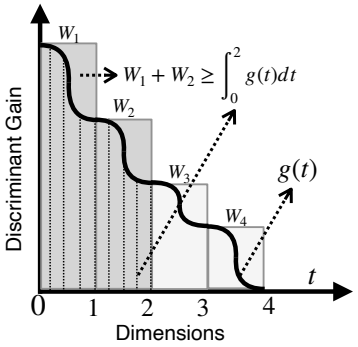


Fig. 4: An example of DG function with $D = 4$.

where $\hat{W}_d$ denotes the DG of the d-th dimension in (15)*, arranged in decreasing order such that $\hat{W}_d \ge \hat{W}_{d+1}, \forall d \in \{1, 2, \ldots, D-1\}$. Otherwise, $\forall t \notin [0, D]$, $g(t) = 0$.*

The defined DG function establishes an approximate relationship between the dimension index d and the corresponding dimension-wise DG. By leveraging Definition 1, the functions $G_1(S)$ and $G_2(S)$ can be lower bounded through the integration of the DG function $g(t)$, denoted as $\hat{G}_1(S)$ and $\hat{G}_2(S)$, respectively, expressed as:

$$\begin{aligned} G_1(S) &= \sum_{d=1}^{S} W_d \ge \int_0^S g(t)dt \triangleq \hat{G}_1(S), \\ G_2(S) &= \sum_{d=1}^{S} W_d^2 \ge \int_0^S g^2(t)dt \triangleq \hat{G}_2(S), \end{aligned} \tag{30}$$

where $S \in [1, D]$ is treated as a continuous variable, relaxing the discrete constraint on the number of selected features. An example of the DG function $g(t)$ and its associated lower bounds in (30) is illustrated in Fig. 4.

B. Optimal DG based Feature Selection

The lower bounds on transmit DG and its associated power enable the derivation of a surrogate function for the objective in Problem (28). Since increasing either the number of selected features, S, or the number of processed observations, K, reduces the InfOut probability, the constraint in (28b) must hold with equality. Accordingly, the maximum number of observations, denoted by $\hat{K}$, is derived from the constraint in (28b) and is given by:

$$\hat{K} = \lfloor -B_1 S + B_0 \rfloor, \tag{31}$$

where $B_1 = \frac{f_c T_\Delta}{N_F} > 0, B_0 = \frac{f_c T_{\max}}{N_F} > 0$.

We then relax $\hat{K}$ by allowing it to take non-integer values. Incorporating this relaxation and lower bounds in (30), we approximate the upper bound of the InfOut probability in (28a) as a function of the number of selected features S, given by

$$Q\left(\frac{P_{\text{act}}G_{\text{f}}(S) - \frac{G_{\text{th}}}{K\sqrt{G_2(S)}}}{\sqrt{P_{\text{act}}(1-P_{\text{act}})}}\right) \approx Q\left(\frac{f(S)}{\sqrt{P_{\text{act}}(1-P_{\text{act}})}}\right). \tag{32}$$

Here, $f(S)$ represents the surrogate function obtained by substituting $G_{\mathsf{f}} = \frac{G_1(S)}{\sqrt{G_2(S)}} \approx \frac{\hat{G}_1(S)}{\sqrt{\hat{G}_2(S)}}$ and $K \approx \hat{K}$ into the expression of the numerator in Q-function, given by

$$f(S) = \frac{P_{\mathsf{act}}\hat{G}_1(S)}{\sqrt{\hat{G}_2(S)}} - \frac{G_{\mathsf{th}}}{(B_0 - B_1 S)\sqrt{\hat{G}_2(S)}}. \tag{33}$$

It is obvious that InfOut probability is a monotonically decreasing function of $f(S)$. Consequently, the InfOut probability minimization problem can be reformulated as the maximization of the surrogate function, leading to the following problem:

$$\max_{S} \quad f(S) \tag{34a}$$

$$\text{s.t.} \quad S \in \{S_{\min}, S_{\min}+1, \ldots, S_{\max}\}, \tag{34b}$$

where $S_{\min} = \max\{1, \lceil \frac{B_0 - K_{\max}}{B_1} \rceil\}$ denotes the minimum number of selected features constrained by $K_{\max}$, and $S_{\max} = \min\{\lfloor \frac{B_0 - 1}{B_1} \rfloor, D\}$ represents the maximum value that guarantees at least one processed observation.

The surrogate $f(S)$ is found to be a concave function of S, exhibiting a unique maximum, established in Proposition 1.

Proposition 1 (Optimal Number of Selected Features)**.** *Let*

$$\nu(x) = P_{\mathsf{act}} g(x) \left(\hat{G}_2(x) - \frac{1}{2}\hat{G}_1(x) g(x) \right) + \frac{G_{\mathsf{th}}((B_0 - B_1 x) g^2(x) - 2\hat{G}_2(x))}{2(B_0 - B_1 x)^2}. \tag{35}$$

where $g(x)$ is the DG function defined in (29)*, and $\hat{G}_1(x), \hat{G}_2(x)$ are defined in* (30)*. The optimal number of selected features that solves Problem* (34) *is then*

$$S^* = \lfloor x^* \rceil_{f(\cdot)}, \tag{36}$$

where the rounding operator $\lfloor x \rceil_{f(\cdot)}$ is equal to $\lfloor x \rfloor$ if $f(\lfloor x \rfloor) \geq f(\lceil x \rceil)$, and is otherwise equal to $\lceil x \rceil$. The value x^* is given by

$$x^* = \{x | \nu(x) = 0, x \in [S_{\min}, S_{\max}]\}, \tag{37}$$

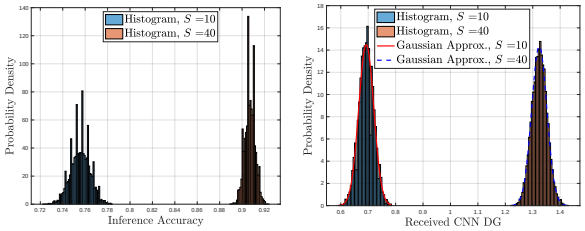
if $\nu(S_{\min}) \cdot \nu(S_{\max}) < 0$ holds, otherwise $S^* = \operatorname{argmax}_{x \in \{S_{\min}, S_{\max}\}} f(x)$.

The proof is provided in Appendix C.

Proposition 1 provides an optimal number of selected features for transmission which minimizes the InfOut probability for enhancing reliability. The optimal selection can be determined by finding the zero of the first derivative of $f(S)$, which is equivalent to solving $\nu(x) = 0$. The optimal solution can be obtained using a bisection search over the feasible range $S \in [S_{\min}, S_{\max}]$ with the complexity of $\mathcal{O}(\log \frac{S_{\max} - S_{\min}}{\varepsilon})$ and tolerance ε.

V. Extension to CNN Classification

In this section, we consider the case of CNN classification and analyze the associated InfOut probability by approximating the inference accuracy distribution using the corresponding receive DG. Subsequently, we address the InfOut probability minimization problem by estimating the distribution of the defined receive CNN DG.



(a) PDF of Inference Accuracy (b) PDF of Receive CNN DG

Fig. 5: Under the magnitude based feature selection scheme, the distribution comparison between inference accuracy and receive CNN DG using the VGG16 model [57] and the ModelNet dataset [49]. The settings are $D = 512, L = 20, \alpha = 1, \beta = 1, K = 12, P_{\mathsf{act}} = 0.7$.

A. Approximation for Receive CNN DG

Unlike linear classifiers, the nonlinearity of the CNN classifier makes it complicated to model the inference accuracy distribution. As shown in Fig. 5(a), the PDF of inference accuracy exhibits an intractable distribution that varies with the selected features, making InfOut probability computation challenging. Moreover, DG based feature selection is not applicable to the CNN case due to the unknown dimension-wise DG of generic CNN feature distributions. To address these challenges, we leverage the two insights from linear classification, as shown in Remark 2.

Remark 2 (Insights from Linear Classification)**.** *Two insights from the linear case guide the subsequent CNN extension. First, under i.i.d. feature losses, the receive DG can be written as a sum of many per-feature contributions, and hence admits an accurate Gaussian approximation when the number of retained features is sufficiently large. Second, under this approximation, the InfOut probability becomes a monotone function of the receive-DG surrogate, which motivates a DG-based feature-selection rule that yields a unique optimum to balance the C^2 tradeoff under E2E latency constraint.*

Building on these findings, we employ a magnitude based feature selection scheme that selects the top-S features with the largest magnitudes, and define the receive CNN DG as follows. This enables computing the InfOut probability using a tractable distribution.

Definition 2 (Receive CNN Discriminant Gain)**.** *Given the inference accuracy of CNN classifier, denoted as a_{cnn}, the corresponding receive CNN DG, G_{cnn}, is quantified as a monotone increasing function of a_{cnn}, given by*

$$G_{\mathsf{cnn}} = \alpha Q^{-1}(\beta(1 - a_{\mathsf{cnn}})), \tag{38}$$

where α and β are the fitting parameters.

Using this mapping and the magnitude based feature selection, we approximate the distribution of the receive CNN DG as a Gaussian random variable under a sufficiently large number of selected features (i.e., $S \geq S_{\min}$):

$$G_{\mathsf{cnn}} \sim \mathcal{N}\left(\mu_{(K,S,P_{\mathsf{act}})}, \sigma^2_{(K,S,P_{\mathsf{act}})}\right), \tag{39}$$

Algorithm 1: Receive CNN DG Distribution Estimation

Input: Sets of activation probability $\mathcal{P}_{\mathsf{act}}$, number of observations $\mathcal{K}$, and selected features $\mathcal{S}$

Initialization: Training datasets and well-trained model;
for Network Parameters: $P_{\mathsf{act}} \in \mathcal{P}_{\mathsf{act}}, K \in \mathcal{K}, S \in \mathcal{S}$ **do**
  for Number of trials: $n = \{1, 2, \ldots, N\}$ **do**
    for Data samples in training dataset **do**
      Extract feature vector $\mathbf{x} \in \mathbb{R}^D$ using the selected observation batch with the size of K;
      Compute the channel-effect feature vector $\hat{\mathbf{x}} = [\hat{x}(1), \ldots, \hat{x}(D)]$ by randomly masking Top-S features with P_{act} and setting others as zeros;
      Infer label using $\hat{\mathbf{x}}$;
    end for
    Compute the inference accuracy $a_{\mathsf{cnn}}(n)$;
    Compute the receive CNN DG $G_{\mathsf{cnn}}(n)$ with (38);
  end for
  Compute the estimated mean of DG $\hat{\mu}_{(K,S,P_{\mathsf{act}})} = \frac{1}{N}\sum_{n=1}^{N} G_{\mathsf{cnn}}(n)$;
  Compute the estimated variance $\hat{\sigma}^2_{(K,S,P_{\mathsf{act}})} = \frac{1}{N-1}\sum_{n=1}^{N}(G_{\mathsf{cnn}}(n) - \hat{\mu}_{(K,S,P_{\mathsf{act}})})^2$;
end for
return Lookup table of receive CNN DG distribution: $\{\hat{\mu}_{(K,S,P_{\mathsf{act}})}\}, \{\hat{\sigma}_{(K,S,P_{\mathsf{act}})}\}$

where $\mu_{(K,S,P_{\mathsf{act}})}$ and $\sigma^2_{(K,S,P_{\mathsf{act}})}$ are the mean and variance of the receive CNN DG, respectively. These coefficients are determined by the number of observations K, transmitted feature number S, and channel activation probability P_{act}.

In Fig. 5, we validate the Gaussian approximation of the receive CNN DG using a real dataset. Fig. 5(a) demonstrates an irregular and intractable PDF of inference accuracy. By using the defined receive CNN DG, Fig. 5(b) shows that the Gaussian approximation accurately captures the distribution across different settings of selected features.

B. Optimal Magnitude based Feature Selection

Building on the Gaussian approximation of receive CNN DG in (39), the InfOut probability of CNN cases under the accuracy threshold A_{th} can be expressed as

$$\begin{aligned} P^{\mathsf{e2e}}_{\mathsf{out}} &= \Pr(a_{\mathsf{cnn}} \leq A_{\mathsf{th}}) \\ &= \Pr\left(G_{\mathsf{cnn}} \leq G_{\mathsf{th}}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_{(K,S,P_{\mathsf{act}})}} \int_{-\infty}^{G_{\mathsf{th}}} \exp\left(-\frac{(x-\mu_{(K,S,P_{\mathsf{act}})})^2}{2\sigma^2_{(K,S,P_{\mathsf{act}})}}\right) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\Psi_{\mathsf{cnn}}} \exp\left(-\frac{y^2}{2}\right) dy \\ &= 1 - Q\left(-\Psi_{\mathsf{cnn}}\right) \\ &= Q\left(\Psi_{\mathsf{cnn}}\right) \end{aligned} \tag{40}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(-\frac{t^2}{2}\right) dt$ is the Gaussian Q-function and Ψ_{cnn} is the surrogate function that minimizes the InfOut probability of CNN classification, given by

$$\Psi_{\mathsf{cnn}} = \frac{\mu_{(K,S,P_{\mathsf{act}})} - G_{\mathsf{th}}}{\sigma_{(K,S,P_{\mathsf{act}})}} \tag{41}$$

with $G_{\mathsf{th}} = \alpha Q^{-1}(\beta(1 - A_{\mathsf{th}}))$ being the threshold of the required receive CNN DG. It follows that the InfOut probability in the CNN case is a monotonically decreasing function of the surrogate function Ψ_{cnn}.

However, maximizing Ψ_{cnn} depends on the unknown parameters $\mu_{(K,S,P_{\mathsf{act}})}$ and $\sigma_{(K,S,P_{\mathsf{act}})}$. To estimate these parameters, we develop an algorithm that emulates the effects of random feature loss on the inference process using training datasets, as outlined in Algorithm 1. Using the estimated parameters, the optimization problem for CNN classification is reformulated as

$$\begin{aligned} \max_{S} \quad & \hat{\Psi}_{\mathsf{cnn}}(S) \\ \text{s.t.} \quad & (34\text{b}), (31), \end{aligned} \tag{42}$$

where $\hat{\Psi}_{\mathsf{cnn}}(S)$ is the estimated surrogate expressed in terms of the number of selected features S, given by

$$\hat{\Psi}_{\mathsf{cnn}}(S) = \frac{\hat{\mu}_{(\hat{K},S,P_{\mathsf{act}})} - G_{\mathsf{th}}}{\hat{\sigma}_{(\hat{K},S,P_{\mathsf{act}})}}, \tag{43}$$

Here, $\hat{K}$ is the maximum achievable number of processed observations in (31). $\hat{\mu}_{(\hat{K},S,P_{\mathsf{act}})} \approx \mu_{(\hat{K},S,P_{\mathsf{act}})}$ and $\hat{\sigma}_{(\hat{K},S,P_{\mathsf{act}})} \approx \sigma_{(\hat{K},S,P_{\mathsf{act}})}$ are the estimated parameters using Algorithm 1. With knowledge of the long-term CSI (i.e., the distribution of channel gain) at the transmitter, the channel activation probability can be computed using (7). Conditioned on P_{act}, the optimal number of selected features is obtained by identifying the solution $S \in \{S_{\min}, \ldots, S_{\max}\}$ that maximizes the estimated surrogate function $\hat{\Psi}_{\mathsf{cnn}}(S)$. This solution can be efficiently found using a bisection search, with the complexity of $\mathcal{O}(\log(S_{\max} - S_{\min} + 1))$.

VI. Experimental Results

A. Experimental Setup

Unless specified otherwise, the default experimental settings are as follows:

1) Computation and communication configuration: We present an edge inference framework consisting of an edge device and an edge server, operating under a $T_{\max} = 10$ ms E2E latency constraint that encompasses both on-device computation and feature transmission. For the computation settings, the edge device randomly selects K observations of the target object for feature extraction using the VGG16 model, which contains 14.7 million network parameters [57]. With this feature extractor, the computation workload for extracting features from a single observation is 936.2 MFLOPs. The edge device is equipped with an NVIDIA Jetson TX2 Series, which provides a computation speed of $f_c = 1$ TFLOPs/s [58]. For the communication settings, each feature is quantized to $Q_B = 16$ bits, with the index quantized by $Q_I = 9$ bits, and is assumed to be transmitted within $T_\Delta = 0.3$ ms. Thus, the resulting computation latency is $T_{\mathsf{comp}}(K) = 0.936K$ ms, while transmission latency is comparable and modeled

as $T_{\mathsf{comm}}(S) = 0.3S$ ms. The system bandwidth is $B_W = 5$ MHz, and the noise power spectral density at the receiver is $N_0 = 10^{-9}$ W/Hz [59]. The Rayleigh fading channel gain is modeled as $h \sim \mathcal{CN}(0,1)$. The resulting channel outage probability is given by

$$P_{\mathsf{out}} = 1 - \exp\left(-\frac{N_0 B_W}{P_{\max}}\left(2^{\frac{Q_B+Q_I}{T_\Delta B_W}} - 1\right)\right), \tag{44}$$

which adapts to the transmit power constraint $P_{\max}$. The accuracy requirements are set at 97% of the maximum achievable accuracy, resulting in $A_{\mathsf{th}} = 96.8\%$ for linear classification and $A_{\mathsf{th}} = 87.3\%$ for CNN based classification.

2) Classifier settings: The two classifiers and their corresponding datasets are detailed as follows.

- **Linear classification on synthetic GMM data:** For the linear classifier, feature vectors are generated according to GMM defined in (4). The feature vectors have a dimensionality of $D = 30$. The centroid of one cluster is a vector with all elements equal to $+1$, while the centroid of the other cluster is a vector with all elements equal to -1. The covariance matrix is given by $\mathbf{C} = \mathrm{diag}\{\frac{2}{3}d + 10\}, d \in \{1, 2, \ldots, 30\}$, modeling the decreasing DG across dimensions. Building on the dimension-wise DG computed by (15), the top-S features are selected for transmission. The inferred label is obtained by feeding the received features into the classifier given in (9).
- **CNN based classification on real-world data:** For the CNN classifier, we utilize the well-known ModelNet dataset [49], which contains multi-view object observations (e.g., a person or a plant), and implement the CNN architecture using the VGG16 model [57]. The VGG16 model is partitioned into a feature extractor and a classifier network, where the feature extractor runs on the device and the classifier operates on the server, following the approach in [35]. The resulting CNN architecture is trained for average pooling and targets a subset of ModelNet dataset containing $L = 20$ popular object classes. To perform feature extraction, the device randomly selects K observations of the same class from the dataset and processes them through the feature extractor. Specifically, each ModelNet image is resized from $3 \times 224 \times 224$ to $3 \times 56 \times 56$ before being processed by the on-device feature extractor, producing a $512 \times 1 \times 1$ tensor [37]. The top-S elements with the highest amplitudes in the feature tensor are selected for transmission. Finally, the received feature tensor is reconstructed and passed to the server-side classifier to generate the inferred label.

3) Benchmarking schemes: To evaluate the performance of the proposed optimal C^2 tradeoff, we consider the following benchmark schemes.

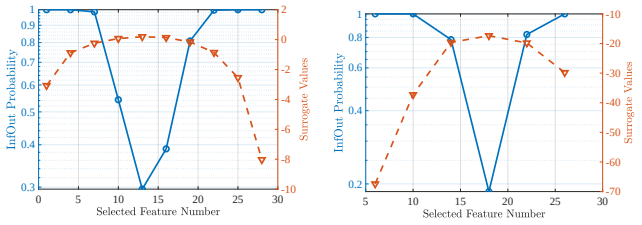
- **Brute-force search:** Given the channel outage probability P_{out}, the feasible solution set is determined by identifying all pairs of the number of observations K and selected features S that satisfy the E2E latency constraint. The optimal solution is then obtained through an exhaustive search over all feasible solutions to minimize the InfOut probability.
- **Maximal features (MaxFeat):** This communication-dominant approach allocates most of the latency budget to feature transmission. Among the feasible solutions that satisfy the latency constraint, this scheme prioritizes the one with the maximum number of features. Once the solution with the maximum features is identified, the number of observations is maximized.
- **Maximal observations (MaxObs):** Unlike MaxFeat, this scheme focuses on incorporating as many observations as possible by extending the computation latency. After identifying feasible solutions with the maximum number of observations, the number of features is maximized.
- **Accuracy-threshold based maximal observations/features (ATB-MaxObs/ATB-MaxFeat):** Unlike the previous baselines, which only maximize the number of observations or selected features, this scheme enforces an accuracy requirement on feasible solutions—a common practice in conventional edge inference under the assumption of reliable channels [34], [36], [60]. Specifically, it uses one-shot inference on the training dataset to filter out solutions whose predicted accuracy falls below a prescribed threshold, and then selects the final solution by maximizing either the number of observations (ATB-MaxObs) or the number of selected features (ATB-MaxFeat), as described above. Mathematically, let $a_{\mathsf{pred}}(K,S)$ denote the profiled one-shot accuracy under reliable channels ($P_{\mathsf{act}} = 1$). The predicted-feasible set is $\mathcal{F}_{\mathsf{pred}} = \left\{(K,S) \;\middle|\; a_{\mathsf{pred}}(K,S) \geq A_{\mathsf{th}},\ \frac{KN_F}{f_c} + T_\Delta S \leq T_{\max}\right\}$. Under fading ($P_{\mathsf{act}} < 1$), the realized accuracy becomes random; denote it by $a(K,S;P_{\mathsf{act}})$. We quantify the miscalibration of the one-shot accuracy prediction by the *mismatch probability*, given by

$$\varepsilon_b \triangleq \Pr\left(a(K_b,S_b;P_{\mathsf{act}}) \leq A_{\mathsf{th}}\right), (K_b,S_b) \in \mathcal{F}_{\mathsf{pred}}, \tag{45}$$

  which measures how often the baseline decision (K_b,S_b) violates the target accuracy under fading. Since the event $a(K_b,S_b;P_{\mathsf{act}}) \leq A_{\mathsf{th}}$ coincides with the InfOut event for the fixed design (K_b,S_b), the mismatch probability can be computed via (27) for linear classification and via (40) for CNN classification. This implies that ε_b is monotone increasing with the channel-outage probability (in the practical regime $P_{\mathsf{out}} \leq 0.5$).

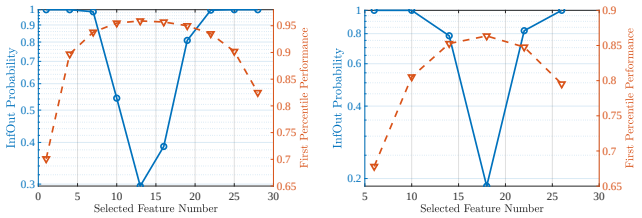
B. Computation-communication Tradeoff

The C^2 tradeoff is illustrated in Fig. 6 using the criteria of InfOut probability and surrogate values. First, as the number of selected features increases, the InfOut probability initially decreases before increasing again, resulting in a unique minimum. This behavior illustrates the tradeoff between communication and computation in latency-constrained edge inference systems. Transmitting more features, which increases communication latency, improves the system's ability to withstand channel outages, thereby reducing the InfOut probability. However, the resulting decrease in the number of processed observations eventually degrades feature quality, causing the



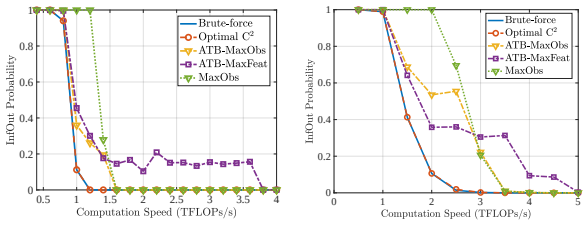
(a) Linear Classification (b) CNN Classification

Fig. 6: The illustration of the C^2 tradeoff under E2E latency constraint, where the relationship between the number of features and observations is modeled by (31).



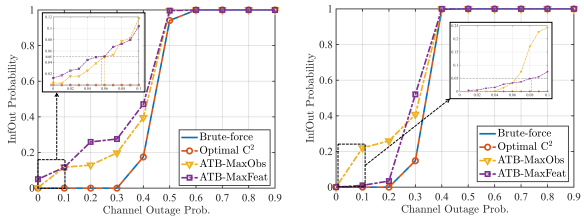
(a) Linear Classification (b) CNN Classification

Fig. 7: The comparison between InfOut probability and first percentile performance adopted by [45].



(a) Linear Classification (b) CNN Classification

Fig. 8: Comparison between optimal C^2 scheme and benchmarks for different settings of computation speed.



(a) Linear Classification (b) CNN Classification

Fig. 9: Comparison between optimal C^2 scheme and benchmarks for different settings of channel outage probability.

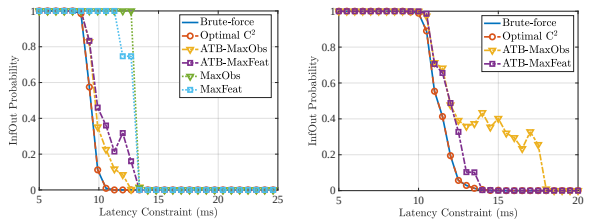
InfOut probability to rise. The unimodal nature of the InfOut probability with respect to the number of transmitted features confirms the existence of a unique minimum, as established in Proposition 1. Second, the surrogate values exhibit a trend opposite to that of the InfOut probability. This observation validates the findings in (32) and (40) for both linear and CNN based classification models. Specifically, both of them reach their extrema at the same point corresponding to the number of selected features.

In Fig. 7, we compare the defined InfOut probability with first percentile performance, which is defined as the value that separates the lowest 1% of samples from the highest 99% of the samples in the accuracy distribution [45]. As the number of transmitted features increases, first percentile performance exhibits an inverse trend to the InfOut probability, indicating that the defined InfOut probability effectively captures the reliability of edge inference systems. These observations further validate the tractability and accuracy of the proposed analytical framework, which is derived based on the surrogate function.

C. Inference Outage Performance

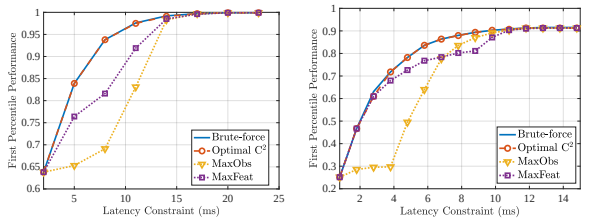
In this subsection, we evaluate inference outage performance by comparing the proposed approach with benchmarks for both linear and CNN based classification. As shown in Fig. 8, the InfOut probability is evaluated with different computation speeds. Specifically, the InfOut probability decreases with increasing computation speed across all schemes. This behavior is attributed to faster computation, which either enables the processing of more observations during feature extraction (improving feature quality) or allows additional latency to be allocated for transmitting more features (enhancing feature quantity). Both factors work together to reduce the InfOut probability. Furthermore, the proposed scheme consistently outperforms benchmarks such as ATB-MaxFeat, ATB-MaxObs, and MaxObs. This superior performance arises from the precise optimization of the C^2 tradeoff. In contrast, ATB-MaxFeat and ATB-MaxObs benchmarks that assume reliable channels and enforce a target accuracy exhibit degraded performance, as they fail to account for the inference accuracy distribution affected by fading channels. The MaxObs scheme prioritizes the number of observations, resulting in significant performance improvements with increased computation speed. At low computation speed, MaxObs spends most of the latency budget on computation and leaves too few features to transmit. Consequently, the normalized receive DG enters the steep region of the Q-function, causing the InfOut probability to increase sharply. However, it suffers from severe performance degradation under low computation capacity. Moreover, the proposed scheme closely approximates the brute-force solution, demonstrating its near-optimal performance.

Then, we evaluate the effects of channel outage probability on system performance, as shown in Fig. 9. As the channel outage probability increases, the InfOut probability correspondingly rises. A higher channel outage probability results in fewer features being received by the edge server, thereby compromising inference system reliability. Within the channel outage probability range of $[0, 0.5]$ for the linear classifier and $[0, 0.4]$ for CNN based classification, the proposed scheme maintains a remarkably low InfOut probability, consistently outperforming the benchmark methods. This underscores the advantage of optimizing the C^2 tradeoff in enhancing the reliability of latency-constrained edge inference systems. However, when the channel outage probability exceeds 0.5, none of the schemes can guarantee reliable inference. The reason is that, when nearly half of the transmitted features are erased, meeting the target accuracy becomes unlikely, pushing the InfOut probability close to one. In addition, to operate in a rare-outage (high-reliability) regime with $P_{\text{out}}^{\text{e2e}} \leq 0.05$, the



(a) Linear Classification (b) CNN Classification

Fig. 10: Comparison between optimal C^2 scheme and benchmarks for different settings of latency constraints.



(a) Linear Classification (b) CNN Classification

Fig. 11: The first percentile performance comparison between optimal C^2 scheme and benchmarks for different settings of latency constraints.

channel-outage probability of the baselines must be limited to $P_{\mathsf{out}} \leq 0.06$ for linear classification and $P_{\mathsf{out}} \leq 0.08$ for CNN classification. By contrast, the proposed approach maintains $P_{\mathsf{out}}^{\mathsf{e2e}} \leq 0.05$ up to $P_{\mathsf{out}} \leq 0.3$ for the linear case and $P_{\mathsf{out}} \leq 0.2$ for CNN classification. This improved tolerance to channel outages stems from the optimized C^2 tradeoff with respect to the channel state.

Next, we compare the proposed scheme with benchmarks under various latency constraints in Fig. 10. The results show that a more relaxed latency constraint leads to a lower InfOut probability. This is because a larger E2E latency allows for processing more observations and transmitting additional features, thereby improving both feature quality and quantity, which in turn reduces the InfOut probability. Similar to the findings in Fig. 8, the proposed scheme consistently demonstrates its optimality across different latency constraints, outperforming the benchmark methods.

Finally, Fig. 11 evaluates the worst-case performance of the optimal C^2 approach using first percentile performance, as applied in [45]. The results show that first percentile performance remains low across all schemes under stringent latency constraints. As the latency constraint is relaxed, performance improves and eventually saturates as more features are transmitted and more observations are processed, thereby enhancing the reliability of the inference. Furthermore, the proposed C^2 scheme outperforms the benchmarks and closely approaches the brute-force solution across various latency constraint settings.

VII. Conclusion

We have investigated the reliability of latency-constrained edge inference systems by introducing the concept called InfOut probability. A fundamental C^2 tradeoff was revealed and quantified: mitigating inference outage requires the edge device to process more observations and upload more features, which increases both computation and communication overhead. However, these requirements conflict under a constraint on latency. To optimize this tradeoff, we have derived a unimodal surrogate function for linear classification and utilized the insights to achieve near-optimal performance in the case of CNN based classification.

This work represents arguably the first theoretical framework for analyzing the reliability of an edge inference system from the perspective of outage probability. The revisiting of outage in the new context opens new directions in communication theory, particularly for edge-AI applications that are both latency-sensitive and mission-critical. Future research could further investigate the optimal balance among latency, reliability, and inference accuracy by exploring diverse communication techniques such as short-packet transmission, MIMO beamforming, and multi-user resource allocation.

Appendix

A. Proof of Lemma 2

To confirm the satisfaction of Lindeberg's condition, we verify the *uniform asymptotic negligibility* (UAN) condition as follows:

$$\lim_{S\to\infty} \max_{1\leq d\leq S} \frac{\mathrm{Var}(X_d)}{\sigma_{\mathsf{G}}^2(S)} = \lim_{S\to\infty} \frac{P_{\mathsf{act}}(1-P_{\mathsf{act}})\hat{W}_1^2}{\sigma_{\mathsf{G}}^2(S)} = 0. \quad (46)$$

Next, we analyze the Lindeberg term. The upper bound of $\forall\, |X_d - \mathbb{E}[X_d]|$ is given by:

$$\begin{aligned} |X_d - \mathbb{E}[X_d]| &\leq \max\{1-P_{\mathsf{act}}, P_{\mathsf{act}}\}\hat{W}_d \\ &\leq C \triangleq \max\{1-P_{\mathsf{act}}, P_{\mathsf{act}}\}\hat{W}_1, \end{aligned} \quad (47)$$

where C denotes the maximum deviation of X_d from its mean. As $S \to \infty$, we have $\epsilon\sigma_{\mathsf{G}}^2(S) = \epsilon P_{\mathsf{act}}(1-P_{\mathsf{act}})\sum_{d=1}^{S}\hat{W}_d^2 > C$. Consequently, the indicator function $\mathbb{I}(|X_d - \mathbb{E}[X_d]| > \epsilon\sigma_{\mathsf{G}}(S))$ equals zero, ensuring that the Lindeberg term vanishes. Thus, Lindeberg's condition in (24) is satisfied, completing the proof.

B. Proof of the Monotonicity of $G_{\mathsf{f}}(S)$

Since the dimension-wise DG follows a decreasing order, i.e., $\hat{W}_d \geq \hat{W}_{d+1}$, the monotonicity of $G_{\mathsf{f}}(S)$ is determined by the sign of the difference $G_{\mathsf{f}}^2(S+1) - G_{\mathsf{f}}^2(S)$, given by:

$$\begin{aligned} G_{\mathsf{f}}^2(S+1) - G_{\mathsf{f}}^2(S) &= \frac{(G_1(S)+\hat{W}_{S+1})^2}{G_2(S)+\hat{W}_{S+1}^2} - \frac{G_1^2(S)}{G_2(S)} \\ &= \frac{\hat{W}_{S+1}G_{\Delta}(S)}{(G_2(S)+\hat{W}_{S+1}^2)G_2(S)}, \end{aligned} \quad (48)$$

$$y''(x)=\hat{G}_2^{-\frac{3}{2}}(x)\left(\underbrace{(-B_1g^2(x)+(B_0-B_1x)g(x)g'(x))\hat{G}_2(x)}_{\leq 0}\underbrace{-\frac{(B_0-B_1x)g^4(x)}{4}}_{\leq 0}\right)\leq 0. \tag{50}$$

where $G_\Delta(S)=G_2(S)\hat{W}_{S+1}+2G_1(S)G_2(S)-G_1^2(S)\hat{W}_{S+1}$ is proven to be positive, given by

$$\begin{aligned}
G_\Delta(S)=&G_2(S)\hat{W}_{S+1}+2G_1(S)G_2(S)-G_1^2(S)\hat{W}_{S+1}\\
=&\hat{W}_{S+1}(G_2(S)-G_1^2(S))+2G_1(S)G_2(S)\\
=&-\hat{W}_{S+1}\sum_{d_1\neq d_2}\hat{W}_{d_1}\hat{W}_{d_2}+2\sum_{d_1=1}^{S}\sum_{d_2=1}^{S}\hat{W}_{d_1}\hat{W}_{d_2}^2\\
\geq&-\hat{W}_{S+1}\sum_{d_1\neq d_2}\hat{W}_{d_1}\hat{W}_{d_2}+\sum_{d_1\neq d_2}\hat{W}_{d_1}\hat{W}_{d_2}^2+\sum_{d=1}^{S}\hat{W}_d^3\\
\geq&\underbrace{-\hat{W}_{S+1}\sum_{d_1\neq d_2}\hat{W}_{d_1}\hat{W}_{d_2}+\hat{W}_{S+1}\sum_{d_1\neq d_2}\hat{W}_{d_1}\hat{W}_{d_2}}_{=0}\\
&+\sum_{d=1}^{S}\hat{W}_d^3\\
\geq&0.
\end{aligned}\tag{49}$$

Since $G_\Delta(S)\geq 0$, it follows that $G_{\mathsf{f}}^2(S+1)-G_{\mathsf{f}}^2(S)\geq 0$ for all $S\in 1,2,\ldots,D$, proving that $G_{\mathsf{f}}(S)$ is a monotonic increasing function. This completes the proof.

C. Proof of Proposition 1

To simplify notation, we express the surrogate function as a linear combination of two continuous functions over $x\in[0,D]$, given by

$$f(x)=P_{\mathsf{act}}f_1(x)+f_2(x), \tag{51}$$

where:

$$f_1(x)=\frac{\hat{G}_1(x)}{\sqrt{\hat{G}_2(x)}},\quad f_2(x)=-\frac{G_{\mathsf{th}}}{(B_0-B_1x)\sqrt{\hat{G}_2(x)}}. \tag{52}$$

Here, the DG based feature selection scheme ensures several properties of $\hat{G}_1(x)$ and $\hat{G}_2(x)$, given by

$$\begin{aligned}
&\hat{G}_1'(x)=g(x)\geq 0,\quad \hat{G}_2'(x)=g^2(x)\geq 0,\\
&\hat{G}_1''(x)=g'(x)\leq 0,\quad \hat{G}_2''(x)=2g(x)g'(x)\leq 0.
\end{aligned}\tag{53}$$

Based on (53), we show the concavity of $f(x)$ by separately showing $f_1(x)$ and $f_2(x)$ are concave functions. First, we prove that the $f_1(x)$ is a concave function. The first derivative of $f_1(x)$ is given by

$$\begin{aligned}
f_1'(x)=&\frac{\hat{G}_1'(x)}{\sqrt{\hat{G}_2(x)}}-\frac{\hat{G}_1(x)\hat{G}_2'(x)}{2\hat{G}_2^{\frac{3}{2}}(x)}\\
=&\frac{\hat{G}_2(x)g(x)-\frac{1}{2}\hat{G}_1(x)g^2(x)}{\hat{G}_2^{\frac{3}{2}}(x)}.
\end{aligned}\tag{54}$$

The second derivative of $f_1(x)$ is upper bounded by

$$\begin{aligned}
f_1''(x)=&\hat{G}_2^{-\frac{5}{2}}(x)\left(g'(x)\hat{G}_2^2(x)-\hat{G}_1(x)\hat{G}_2(x)g(x)g'(x)\right.\\
&\left.+\frac{3}{4}\hat{G}_1(x)g^4(x)-\hat{G}_2(x)g^3(x)\right)\\
=&\hat{G}_2^{-\frac{5}{2}}(x)\left\{g'(x)\hat{G}_2(x)[\hat{G}_2(x)-\hat{G}_1(x)g(x)]-g^3(x)\right.\\
&\left.\times\left[\hat{G}_2(x)-\frac{3}{4}\hat{G}_1(x)g(x)\right]\right\}\\
\leq&\hat{G}_2^{-\frac{5}{2}}(x)\left\{g'(x)\hat{G}_2(x)[\hat{G}_2(x)-\hat{G}_1(x)g(x)]\right.\\
&\left.-g^3(x)\left[\hat{G}_2(x)-\hat{G}_1(x)g(x)\right]\right\}\\
=&\underbrace{\hat{G}_2^{-\frac{5}{2}}(x)}_{\geq 0}\underbrace{[g'(x)\hat{G}_2(x)-g^3(x)]}_{\leq 0}\underbrace{[\hat{G}_2(x)-\hat{G}_1(x)g(x)]}_{\triangleq\zeta(x)\geq 0}\\
\leq&0,
\end{aligned}\tag{55}$$

where $\zeta(x)=\hat{G}_2(x)-\hat{G}_1(x)g(x)$ can be proven to be a non-negative function over $x\in[0,D]$. This follows from the fact that its derivative satisfies

$$\zeta'(x)=-g'(x)\hat{G}_1(x)\geq 0. \tag{56}$$

Since $\zeta(x)$ is non-decreasing, its minimum occurs at $x=0$, where

$$\zeta(x)\geq\min_x\zeta(x)=\zeta(0)=\hat{G}_2(0)-\hat{G}_1(0)g(0)=0, \tag{57}$$

where $\hat{G}_2(0)=\hat{G}_1(0)=0$. Thus, $\zeta(x)\geq 0$ for all x, confirming that $f_1(x)$ is concave as $f_1''(x)\leq 0$.

Next, we show that $f_2(x)=-\frac{G_{\mathsf{th}}}{y(x)}$ is concave. Let

$$y(x)=(B_0-B_1x)\sqrt{\hat{G}_2(x)}. \tag{58}$$

The second derivative of $f_2(x)$ is given by:

$$f_2''(x)=\frac{G_{\mathsf{th}}}{y^3(x)}(y''(x)y^2(x)-2(y'(x))^2). \tag{59}$$

From (59), it suggests that $y''(x)\leq 0$ is a sufficient condition that $f_2(x)$ is a concave function (i.e., $f_2''(x)\leq 0$), which can be proven by (50). In the end, $f(x)$ is a concave function of x due to the sum of two concave functions.

Based on the analysis above, the resulting optimal solution ensuring the maximum of $f(x)$ can be obtained by finding the zero of $f'(x)=0$, given as

$$x^*=\{x\mid f'(x)=0, x\in[S_{\min},S_{\max}]\}, \tag{60}$$

where the derivative $f'(x)$ is given by:

$$f'(x)=\hat{G}_2^{-\frac{3}{2}}(x)\nu(x), \tag{61}$$

with

$$\nu(x) = P_{\mathsf{act}} g(x) \left(\hat{G}_2(x) - \frac{1}{2} \hat{G}_1(x) g(x) \right) + \frac{G_{\mathsf{th}}((B_0 - B_1 x) g^2(x) - 2\hat{G}_2(x))}{2(B_0 - B_1 x)^2}.$$

Since $f(x)$ is concave for $x \in [S_{\min}, S_{\max}]$, the optimal number of selected features in problem (34), denoted as S^*, can be determined by evaluating the nearest (feasible) integers below and above x^*. The value that maximizes the objective function $f(x)$ is then selected, provided that $\nu(S_{\min}) \cdot \nu(S_{\max}) \leq 0$. Otherwise, if $\nu(S_{\min}) \cdot \nu(S_{\max}) \geq 0$, the optimal number of selected features is one of the endpoints, i.e., $S^* = \operatorname{argmax}_{S \in \{S_{\min}, S_{\max}\}} f(x)$. This completes the proof.

REFERENCES

[1] Z. Liu, X. Chen, H. Wu, Z. Wang, X. Chen, D. Niyato, and K. Huang, "Integrated sensing and edge ai: Realizing intelligent perception in 6G," *IEEE Commun. Surveys Tuts.*, vol. 28, pp. 2725–2770, 2026.

[2] Q. Chen, Z. Wang, X. Chen, J. Wen, D. Zhou, S. Ji, M. Sheng, and K. Huang, "Space–ground fluid AI for 6G edge intelligence," *Engineering*, vol. 54, pp. 14–19, 2025.

[3] Z. Wang, K. Huang, and Y. C. Eldar, "Spectrum breathing: Protecting over-the-air federated learning against interference," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 10 058–10 071, 2024.

[4] Z. Wang, M. Cui, H. Yang, Q. Zeng, M. Sheng, and K. Huang, "Airbreath sensing: Protecting over-the-air distributed sensing against interference," arXiv:2508.11267, 2025.

[5] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, 2020.

[6] D. Wen *et al.*, "Task-oriented sensing, computation, and communication integration for multi-device edge AI," *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2486–2502, 2023.

[7] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige *et al.*, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *IEEE Int. Conf. Rob. Autom. (ICRA)*, Brisbane, Australia, 2018, pp. 4243–4250.

[8] D. Wen, X. Jiao, P. Liu, G. Zhu, Y. Shi, and K. Huang, "Task-oriented over-the-air computation for multi-device edge AI," *IEEE Trans. Wireless Commun.*, vol. 23, no. 3, pp. 2039–2053, 2023.

[9] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[10] M. K. Simon and M.-S. Alouini, *Digital communication over fading channels*. John Wiley & Sons, 2004.

[11] M.-S. Alouini and A. J. Goldsmith, "Capacity of rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. Veh. Technol.*, vol. 48, no. 4, pp. 1165–1181, 1999.

[12] Q. Chen, W. Meng, S. Han, C. Li, and T. Q. S. Quek, "Coverage analysis of SAGIN with sectorized beam pattern under shadowed-rician fading channels," *IEEE Trans. Commun.*, vol. 71, no. 8, pp. 4988–5004, Aug. 2023.

[13] M. O. Hasna and M.-S. Alouini, "Outage probability of multihop transmission over nakagami fading channels," *IEEE Commun. Lett.*, vol. 7, no. 5, pp. 216–218, 2003.

[14] M.-S. Alouini and A. J. Goldsmith, "Adaptive modulation over nakagami fading channels," *Wireless Pers. Commun.*, vol. 13, pp. 119–143, 2000.

[15] X. Tang, M.-S. Alouini, and A. J. Goldsmith, "Effect of channel estimation error on M-QAM BER performance in rayleigh fading," *IEEE Trans. Commun.*, vol. 47, no. 12, pp. 1856–1864, 1999.

[16] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[17] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.

[18] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-Tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, April 2012.

[19] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1986–1992, 1997.

[20] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, 1997.

[21] ——, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 595–602, 1998.

[22] S. Vishwanath and A. Goldsmith, "Adaptive turbo-coded modulation for flat-fading channels," *IEEE Trans. Commun.*, vol. 51, no. 6, pp. 964–972, 2003.

[23] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[24] L. Zheng and D. N. C. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple-antenna channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1073–1096, 2003.

[25] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint power and blocklength optimization for URLLC in a factory automation scenario," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1786–1801, 2019.

[26] Y. Zhao, J. Hu, K. Yang, and X. Wei, "A joint communication and control system for URLLC in industrial IoT," *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 15 074–15 079, 2023.

[27] J. Östman, G. Durisi, E. G. Ström, M. C. Coşkun, and G. Liva, "Short packets over block-memoryless fading channels: Pilot-assisted or noncoherent transmission?" *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1521–1536, 2018.

[28] K. F. Trillingsgaard and P. Popovski, "Downlink transmission of short packets: framing and control information revisited," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2048–2061, 2017.

[29] C. She, C. Yang, and T. Q. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2266–2280, 2018.

[30] Y. Hu, Y. Zhu, M. C. Gursoy, and A. Schmeink, "SWIPT-enabled relaying in IoT networks operating with finite blocklength codes," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 74–88, 2018.

[31] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, and K. Huang, "Pushing large language models to the 6g edge: Vision, challenges, and opportunities," *IEEE Commun. Mag.*, vol. 63, no. 9, pp. 52–59, 2025.

[32] H. Yang, Z. Wang, and K. Huang, "Optimal batch-size control for low-latency federated learning with device heterogeneity," *IEEE Trans. Commun.*, vol. 74, pp. 5232–5247, 2026.

[33] J. Shao and J. Zhang, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 20–26, 2020.

[34] Z. Liu, Q. Lan, and K. Huang, "Resource allocation for multiuser edge inference with batching and early exiting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1186–1200, 2023.

[35] Z. Liu, Q. Lan, A. E. Kalør, P. Popovski, and K. Huang, "Over-the-air multi-view pooling for distributed sensing," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7652–7667, 2024.

[36] Q. Lan, Q. Zeng, P. Popovski, D. Gündüz, and K. Huang, "Progressive feature transmission for split classification at the wireless edge," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3837–3852, 2023.

[37] Z. Wang, A. E. Kalør, Y. Zhou, P. Popovski, and K. Huang, "Ultra-low-latency edge inference for distributed sensing," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 1908–1922, 2026.

[38] Q. Zeng, Z. Wang, Y. Zhou, H. Wu, L. Yang, and K. Huang, "Knowledge-based ultra-low-latency semantic communications for robotic edge intelligence," *IEEE Trans. Commun.*, 2024.

[39] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6G: Vision, principles, and technologies," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 78–85, 2023.

[40] P. Neuhaus, N. Shlezinger, M. Dörpinghaus, Y. C. Eldar, and G. Fettweis, "Task-based analog-to-digital converters," *IEEE Trans. Signal Process.*, vol. 69, pp. 5403–5418, 2021.

[41] Y. E. Sagduyu, S. Ulukus, and A. Yener, "Task-oriented communications for NextG: End-to-end deep learning and AI security aspects," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 52–60, 2023.

[42] S. Ma, W. Qiao, Y. Wu, H. Li, G. Shi, D. Gao, Y. Shi, S. Li, and N. Al-Dhahir, "Task-oriented explainable semantic communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9248–9262, 2023.

[43] Y. Peng, J. Guo, and C. Yang, "Learning resource allocation policy: Vertex-gnn or edge-gnn?" *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 190–209, 2024.

[44] X. Zhang *et al.*, "Robustness of neural networks: A probabilistic and practical approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2540–2554, 2019.

[45] Z. Yan, Y. Qin, W. Wen, X. S. Hu, and Y. Shi, "Improving realistic worst-case performance of NVCiM DNN accelerators through training with right-censored gaussian noise," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD)*, San Francisco, CA, USA, 2023.

[46] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symp. Secur. Privacy (SP)*, San Jose, CA, USA, 2017, pp. 39–57.

[47] M. Ye and R. Yang, "Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 23–28 2014.

[48] J. A. Figueroa, "Semi-supervised learning using deep generative models and auxiliary tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Workshop*, Vancouver, Canada, Dec. 13–14 2019.

[49] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, Santiago, Chile, Dec. 7-13 2015.

[50] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, 2020, pp. 6876–6883.

[51] S. P. Biswas, P. Roy, N. Patra, A. Mukherjee, and N. Dey, "Intelligent traffic monitoring system," in *Proc. Int. Conf. Comput. Commun. Technol. (IC3T)*, vol. 2, Allahabad, India, 2016, pp. 535–545.

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

[53] C. You, K. Huang, H. Chae, and B. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017.

[54] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev.: Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.

[55] Y. Zhu, C. Li, B. Luo, J. Tang, and X. Wang, "Dense feature aggregation and pruning for RGBT tracking," in *Proc. 27th ACM Int. Conf. Multimedia (ACM MM)*, Nice, France, 2019, pp. 465–472.

[56] W. Feller, *An Introduction to Probability Theory and Its Applications, vol. 2*. John Wiley & Sons, 1991.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 7-9 2015.

[58] NVIDIA Corp., "NVIDIA Jetson TX2 Series Module Datasheet," [Online]. Available: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-tx2/.

[59] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7947–7962, 2021.

[60] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, 2019.

Zhanwei Wang (Member, IEEE) received the B.Eng. degree in Information Engineering and the M.Eng. degree in Information and Communication Engineering from Xidian University, Xi'an, China, in 2018 and 2021, respectively. He received the Ph.D. degree in Electrical and Computer Engineering from The University of Hong Kong (HKU), Hong Kong SAR, China. He is currently a Research Assistant Professor with the Department of Electrical and Electronic Engineering, HKU. His research interests include wireless communications, edge intelligence, atomic receivers, and space artificial intelligence (Space AI).

Qunsong Zeng (Member, IEEE) received the B.S. degree in physics (Yan Jici Talent Students Program) from the University of Science and Technology of China (USTC) and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong (HKU). He is currently a Research Assistant Professor with the Department of Electrical and Computer Engineering, The University of Hong Kong, Hong Kong. He was a recipient of HKU Foundation Award, P.K. Yu Memorial Scholarship, Y S and Christabel Lung Scholarship, etc. His research interests include edge intelligence, in-memory baseband processing and quantum receivers. He served as the co-chair for IEEE PIMRC'26 (WS-12), IEEE PIMRC'25 (WS-10) and IEEE/CIC ICCC'25 (WS-22).

Haotian Zheng (Graduate Student Member, IEEE) received the B.Eng. degree in Electronic Information Engineering from Xidian University, Xi'an, China, in 2024. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of Hong Kong (HKU), Hong Kong SAR, China. His research interests include edge intelligence and space artificial intelligence (Space AI).

Kaibin Huang (Fellow, IEEE) received the B.Eng. and M.Eng. degrees from the National University of Singapore and the Ph.D. degree from The University of Texas at Austin, all in electrical engineering. He is the Philip K H Wong Wilson K L Wong Professor in Electrical Engineering and the Department Head at the Dept. of Electrical and Computer Engineering, The University of Hong Kong (HKU), Hong Kong. His work was recognized with seven Best Paper awards from the IEEE Communication Society. He has served on the editorial boards of five major journals in the area of wireless communications and co-edited twelve journal special issues. He has been named as a Highly Cited Researcher by Clarivate in the last seven years (2019-2025) and an AI 2000 Most Influential Scholar (Top 30 in Internet of Things) in 2023-2025. He received the 2025 IEEE Wireless Communications Technical Committee Recognition Award. He is a member of the Engineering Panel of Hong Kong Research Grants Council. He was an IEEE Distinguished Lecturer (2020-2022). He is a Croucher Senior Research Fellow (2026), Fellow of the IEEE (2021), and Fellow of the U.S. National Academy of Inventors (2024).