Impact of Calibration in ML-Aided Wireless Resource Allocation

Rashika Raina*, Nidhi Simmons*, David E. Simmons†, Michel Daoud Yacoub‡, and Trung Q. Duong§

*Centre for Wireless Innovation, School of Electronics, Electrical Engineering and Computer Science,
Queen's University Belfast, BT3 9DT, UK (e-mail: rraina01@qub.ac.uk, nidhi.simmons@qub.ac.uk)
†Dhali Holdings Ltd., Belfast BT5 7HW, UK, (e-mail: dr.desimmons@gmail.com)
‡Wireless Technology Laboratory, School of Electrical and Computer Engineering,
University of Campinas, Campinas 13083-970, Brazil, (e-mail: mdyacoub@unicamp.br)
§Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1C 5S7, Canada, and
School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, U.K.
(e-mail: tduong@mun.ca.)

***Abstract*—Machine learning (ML) is becoming a core tool for enabling accurate and reliable decision-making in 6G wireless networks. This paper analyzes the calibration behavior of an ML-driven outage predictor trained using a system-specific outage loss function. The study is conducted within a single-user, multi-resource allocation system under Rician fading conditions. We present the outage probability expressions for this system under perfect calibration. Through extensive Monte Carlo simulations, we demonstrate that, under perfect calibration and with a large number of resources, the outage probability equals the expected predicted probability, conditioned on it being below the threshold used for classification. Contrary to this, when only one resource is available, the outage probability corresponds to the expected predicted probability across all predictions. These results offer practical guidance to system designers in selecting the threshold used for classification to meet reliability targets. Also, for finite Rician factor, channel uncertainty makes calibration meaningful; however, as it tends to infinity, the randomness in channel behavior diminishes, and the predictor's outputs reliably mirror actual outcomes, making calibration redundant. Finally, we show that post-processing calibration cannot reduce the minimum achievable outage probability, as it does not provide the predictor with any additional knowledge about future channel states.**

***Index Terms*—Machine learning calibration, Outage prediction, Outage loss function, Post-processing calibration, Resource allocation, Trustworthy AI.**

I. Introduction

As wireless networks evolve toward 6G, machine learning (ML) is increasingly used to generate accurate predictions alongside calibrated confidence estimates, enabling robust decision-making under uncertainty. Key 6G use cases, such as ultra-reliable low-latency communications, autonomous vehicles, and smart factory automation, demand not just accurate predictions but also well-calibrated confidence levels. In these critical scenarios, reliable confidence estimates are vital.

Among various ML models, neural networks (NNs) are widely adopted due to their high predictive accuracy. Despite their strong performance, NNs often suffer from poor calibration, which undermines their reliability in real-world applications, particularly in wireless systems where accurate probability estimates are critical [1]. Calibration is commonly assessed using reliability diagrams, which compare predicted confidence levels with observed accuracies. However, traditional linear binning used in these diagrams fails to capture model behavior at low probability levels. These low-probability regions are particularly relevant in wireless systems, where infrequent events can significantly impact system performance. To better capture this behavior, [2] proposed logarithmic binning that provides a detailed view across several orders of magnitude and better highlights rare events.

ML has already proven effective in a variety of wireless communication tasks including blockage prediction [3], power allocation in massive multiple-input multiple-output (MIMO) systems [4], energy efficiency improvements in heterogeneous networks [5], and resource allocation [6], [7]. These models typically rely on standard loss functions such as binary cross-entropy (BCE) and mean squared error, to optimize model performance [8]. Although widely used in training ML models, such loss functions often fail to reflect the specific objectives of wireless systems. To overcome the limitations of standard loss functions, tailored alternatives have been developed that incorporate domain-specific knowledge and more closely align with the objectives of wireless systems [9]. These have been successfully applied to problems such as hybrid beamforming for MIMO systems [10], optimizing reflection coefficients in reconfigurable intelligent surfaces [11], resource allocation in cell-free massive MIMO networks [12] and ultra-reliable low-latency communications [13]. Of particular relevance to this

A more theoretically oriented work related to this paper but focusing on a simpler Rayleigh fading scenario is currently under review for publication in the IEEE Transactions on Network Science and Engineering.

This work was supported by the Royal Academy of Engineering (grant ref RF\201920\19\191).

work is the outage loss function (OLF), introduced in [14], a custom loss function specifically designed to directly minimize outage probability.

Nevertheless, calibration remains a significant challenge in ML-aided communication systems, irrespective of the loss function used. Techniques such as Bayesian learning [15] aim to model predictive uncertainty, while conformal prediction [16] provides formal, distribution-free calibration guarantees. Post-processing calibration, in contrast, adjusts model confidence scores using a held-out validation set and uses techniques such as Platt scaling, isotonic regression, or beta calibration for classification, and Dirichlet calibration and temperature scaling for multiclass settings [17]–[20].

Our contribution focuses on the effect of applying post-processing calibration methods to the outage predictor proposed in [14], specifically evaluating their impact on outage probability in an ML-assisted resource allocation framework over Rician fading channels. Using Monte Carlo simulations, we show that in the infinite-resource limit, the outage probability equals the conditional expectation of the predictor's output, provided it is less than the threshold used for classification. In contrast, in the single-resource setting, it is determined by the predictor's overall expected output. We also show that post-processing calibration methods cannot reduce the system's minimum achievable outage probability.

The paper proceeds as follows: Section II revisits the resource allocation strategy from [14] along with a discussion of key calibration concepts. Section III presents the outage probability expressions for a calibrated predictor from [21]. Section IV describes the data generation process, the calibration techniques used in this work, and the numerical results demonstrating the behavior of the outage probability expressions presented in Section III. Finally, Section V offers concluding observations and reflects on the overall contributions of the paper.

II. System Model

Revisiting the system model in [14] which considered a single-user multi-resource system for a Rayleigh channel, we extend the analysis to account for Rician fading channels.

A. Channel Model

Here, each resource $j \in \mathtt{R}$ has a time-varying channel state $h_j(t) \in \mathbb{C}$, where the channel states are considered to be independent and identically distributed (i.i.d.) across distinct resources. Additionally, the correlation between $h_j(t)$ and $h_j(t+l)$ diminishes as $l \to \infty$. To model this time variation, the channel states $h_j(t)$ across resources are generated from the fast Fourier Transform (FFT) of a continually evolving tapped channel response: $\boldsymbol{g}(t) = [g_1(t), \ldots, g_v(t)]$, where each tap $g_i(t)$ is an independent complex stochastic process indexed by i. In accordance with [22], the Rician channel model used here elects $g_1(t)$ to include a deterministic line-of-sight (LoS) component with its corresponding Doppler shift, whereas all the remaining taps constitute independent zero-mean complex Gaussian components representing scattered (non-LoS) paths. The time evolution of the taps is defined as:

$$g_1(t) = \sqrt{\frac{\mathrm{K}}{\mathrm{K}+1}} e^{j(2\pi f_D t + \phi)} + \sqrt{\frac{1}{\mathrm{K}+1}} \tilde{g}_1(t),$$
$$g_i(t) = \tilde{g}_i(t), \forall i > 1, \qquad (1)$$

where f_D is the Doppler frequency and ϕ ∼Unif$[0, 2\pi]$ accounts for the fixed initial phase offset. Each $\tilde{g}_i(t) \sim \mathcal{CN}(0,\sigma^2)$ evolves over time with small, random phase shifts as:

$$\tilde{g}_i(t+l) = \tilde{g}_i(t) \cdot e^{jl\theta_{it}}, \, \theta_{it} \sim \text{Unif}[-\xi, \xi] \qquad (2)$$

This model captures both deterministic and stochastic fading: the LoS component imparts a Doppler-induced phase rotation, while the non-LoS evolution reflects small, random phase variations consistent with Clarke's 3D scattering model [23], where the channel's autocorrelation exhibits a sinc-shaped decay.

The user leverages an ML-based outage predictor to select a suitable resource by learning temporal channel correlations. The channel evolution for resource j at time t is described using two vectors: $\overleftarrow{\boldsymbol{h}}_j^k(t) \triangleq [h_j(t-k+1), \ldots, h_j(t)]^T$ and $\overrightarrow{\boldsymbol{h}}_j^l(t) \triangleq [h_j(t+1), \ldots, h_j(t+l)]^T$, where $k, l \in \mathbb{N}$.

The capacity, $C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right) \in \mathbb{R}^+$, represents the maximum rate at which communication can be supported by the channel from time $t+1$ to $t+l$, with arbitrarily low errors. In the case of a quasi-static Gaussian channel, this capacity is characterized by the expression provided in [24, eq. (5.80)]:

$$C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right) = \sum_{i=1}^{l} \log_2 \left(1 + \mathtt{SNR} \left|h_j\left(t+i\right)\right|^2\right) \quad \text{bits/s/Hz}, \qquad (3)$$

where $\mathtt{SNR}$ is the average signal-to-noise ratio per sample. A resource is considered to be in outage if the user's required communication rate, denoted by $\gamma_{\mathtt{th}}$, exceeds the available capacity; otherwise, it is regarded as sufficient. The outage probability for a single resource j can be expressed as

$$\mathtt{P}_j(\gamma_{\mathtt{th}}) = \mathbb{P}\left[C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right) < \gamma_{\mathtt{th}}\right]. \qquad (4)$$

B. ML-based Resource Allocation

This work considers a calibrated ML-based outage predictor, in contrast to [14], which did not account for calibration. This predictor outputs a confidence score $\mathtt{Q}\left(\overleftarrow{\boldsymbol{h}}_j^k(t); \Theta\right) \in [0,1]$, estimating the conditional probability $\mathbb{P}\left[C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right) < \gamma_{\mathtt{th}} \,\middle|\, \overleftarrow{\boldsymbol{h}}_j^k(t)\right]$, based on the past k channel samples for resource j. An outage is predicted when the output exceeds a threshold $\mathtt{q}_{\mathtt{th}}$. To allocate resources, the user sequentially evaluates each $j \in 1, \ldots, |\mathtt{R}|$ using $\mathtt{Q}\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right)$. Each time the predictor identifies an outage event, the corresponding index is incremented by one. If no resource satisfies this condition, the model defaults to selecting the final resource.

To evaluate the calibration performance of a predictor, reliability diagrams are used, which plots the observed accuracy as a function of the predicted confidence level [25]. In this

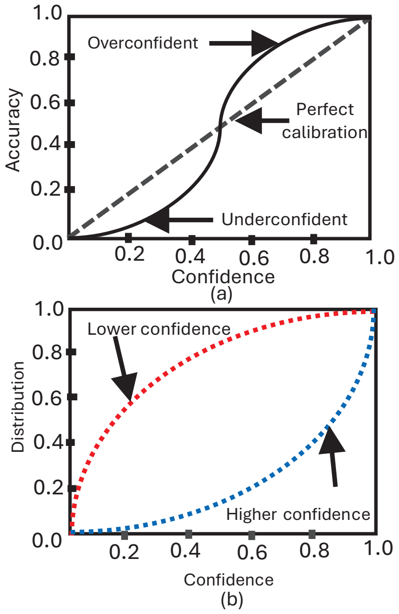

Fig. 1. (a) Accuracy-confidence curve $A_j(\mathtt{q})$ (top), and (b) confidence distribution $F_j(\mathtt{q})$.

framework, the accuracy-confidence function $A_j(\mathtt{q})$ for our outage predictor can be expressed as:

$$A_j(\mathtt{q}) \triangleq \mathbb{P}\left[C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right) < \gamma_{\mathtt{th}} \middle| \mathtt{Q}\left(\overleftarrow{\boldsymbol{h}}_j^k(t);\Theta\right) = \mathtt{q}\right], \tag{5}$$

representing the true outage probability conditioned on the predictor assigning an outage confidence of $\mathtt{q}$.

Fig. 1 (a) shows $A_j(\mathtt{q})$ curve for perfect calibration ($A_j(\mathtt{q}) = \mathtt{q}$), under-confidence ($A_j(\mathtt{q}) > \mathtt{q}$), and over-confidence ($A_j(\mathtt{q}) < \mathtt{q}$). Fig. 1 (b) shows the confidence distribution $F_j(\mathtt{q})$, defined as

$$F_j(\mathtt{q}) = \mathbb{P}\left[\mathtt{Q}\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right) \leq \mathtt{q}\right], \tag{6}$$

which indicates the proportion of predictions with confidence at most $\mathtt{q}$. A lower $F_j(\mathtt{q})$ corresponds to a model that more frequently produces high-confidence predictions.

In practice, predictors are often incorrectly calibrated (i.e., $A_j(\mathtt{q}) \neq \mathtt{q}$). Calibration aims to adjust $\mathtt{q}$ so that $A_j(\mathtt{q}) \approx \mathtt{q}$. Let $\mathtt{Q}^c$ denote the calibrated predictor. Following [25], calibration is a deterministic transformation:

$$\mathtt{Q}^c = \mathcal{C}(\mathtt{Q};\mathtt{D}) = r \circ \mathtt{Q}, \tag{7}$$

where $r : [0,1] \to [0,1]$ is a recalibration function learned on held-out data set $\mathtt{D}$. Since $\mathtt{Q}^c$ is obtained via post-processing, it cannot add new predictive information. By the data processing inequality, we have

$$I\left(C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right);\mathtt{Q}^c\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right)\right) \leq I\left(C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right);\mathtt{Q}\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right)\right), \tag{8}$$

where $I(\cdot,\cdot)$ denotes mutual information [26].

C. Resource Allocation Framework

We follow the resource selection strategy from [14], where the user sequentially scans the resources $j \in \mathtt{R} = \{1,2,\ldots,|\mathtt{R}|\}$ and uses the outage predictor $\mathtt{Q}^\star\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right)$ to determine whether communication is likely to succeed. The user selects the first resource for which the predictor indicates success. If none of the resources satisfy this condition, the predictor selects the final resource. Formally, let

$$\tilde{\mathtt{R}}(\mathtt{q}_{\mathtt{th}}) = \left\{j \in \mathtt{R} \text{ s.t. } \mathtt{Q}^\star\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right);\Theta) \leq \mathtt{q}_{\mathtt{th}}\right\}, \tag{9}$$

represent the subset of resources where the outage predictor predicts no outages, using $\mathtt{q}_{\mathtt{th}}$ as the threshold for classification. The greedy allocation scheme then selects

$$j'(\mathtt{q}_{\mathtt{th}}) = \begin{cases} \min_{j\in\tilde{\mathtt{R}}(\mathtt{q}_{\mathtt{th}})} j & \text{if } \tilde{\mathtt{R}}(\mathtt{q}_{\mathtt{th}}) \neq \emptyset, \\ |\mathtt{R}| & \text{otherwise.} \end{cases} \tag{10}$$

Let $\mathtt{P}^\star_{|\mathtt{R}|}(\gamma_{\mathtt{th}},\mathtt{q}_{\mathtt{th}}) \in \{\mathtt{P}_{|\mathtt{R}|}(\gamma_{\mathtt{th}},\mathtt{q}_{\mathtt{th}}),\ \mathtt{P}^c_{|\mathtt{R}|}(\gamma_{\mathtt{th}},\mathtt{q}_{\mathtt{th}})\}$ denote the system's outage probability when using either the uncalibrated predictor $\mathtt{Q}$ or the calibrated predictor $\mathtt{Q}^c$, respectively.

The outage probability expressed in terms of $A_j(\mathtt{q})$ and $F_j(\mathtt{q})$ is given by [21, eq. (14)]:

$$\begin{aligned} \mathtt{P}^\star_{|\mathtt{R}|}(\gamma_{\mathtt{th}},\mathtt{q}_{\mathtt{th}}) = {} & \mathtt{P}_j(\gamma_{\mathtt{th}})\left(1-F_j(\mathtt{q}_{\mathtt{th}})\right)^{|\mathtt{R}|-1} \\ & +\mathbb{E}\left[A_j\left(\mathtt{Q}^\star\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right)\right) | \mathtt{Q}^\star\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right) \leq \mathtt{q}_{\mathtt{th}}\right] \\ & \times\left(1-\left(1-F_j(\mathtt{q}_{\mathtt{th}})\right)^{|\mathtt{R}|-1}\right). \end{aligned} \tag{11}$$

For $|\mathtt{R}|\to\infty$, the outage probability, as derived in [14], is given by:

$$\mathtt{P}^\star_\infty(\gamma_{\mathtt{th}},\mathtt{q}_{\mathtt{th}}) = \mathbb{P}\left[C\left(\overrightarrow{\boldsymbol{h}}_j^l(t)\right) < \gamma_{\mathtt{th}} \middle| \mathtt{Q}^\star\left(\overleftarrow{\boldsymbol{h}}_j^k(t)\right) \leq \mathtt{q}_{\mathtt{th}}\right], \tag{12}$$

with $\mathtt{Q}^\star \in \{\mathtt{Q},\mathtt{Q}^c\}$ as before.

The OLF, tailored to this resource allocation setting, is defined in [14, eq. (29)] and is used to train the outage predictor, enabling the generation of the results presented in Section IV. Going forward, we omit the dependence on t for the sequences $\overleftarrow{\boldsymbol{h}}_j^k(t)$ and $\overrightarrow{\boldsymbol{h}}_j^l(t)$, as well as the dependence on j for $\mathtt{P} \triangleq \mathtt{P}_j$, $A(\mathtt{q}) \triangleq A_j(\mathtt{q})$, and $F(\mathtt{q}) \triangleq F_j(\mathtt{q})$. Where appropriate, we also drop the dependence of the predictor output $\mathtt{Q}^\star(\cdot)$ and of $C(\cdot)$ on $\overleftarrow{\boldsymbol{h}}_j^k(t)$ and $\overrightarrow{\boldsymbol{h}}_j^l(t)$, respectively. Additionally, we denote by $\mathtt{Q}^\star$ a random variable with the same distribution as any confidence level $\mathtt{Q}^\star$ assigned by the predictor to a resource j.

III. Outage Expressions for Calibrated Predictor

This section presents outage probability expressions and analyzes how the output of a calibrated predictor determines system-level outage probability. Please note that, this work does not focus on deriving these expressions here (see [21] for appropriate derivations).

For the system described in Section II, the outage probability for a perfectly calibrated predictor can be written as [21, eq. (16) and (17)]:

$$\mathtt{P} = \mathbb{E}[\mathtt{Q}^c], \tag{13}$$

$$\text{and} \qquad \mathtt{P}^c_\infty(\gamma_{\mathtt{th}},\mathtt{q}_{\mathtt{th}}) = \mathbb{E}[\mathtt{Q}^c \mid \mathtt{Q}^c \leq \mathtt{q}_{\mathtt{th}}]. \tag{14}$$

The outage probability for an $|\mathtt{R}|$ resource system can then be expressed as [21, eq. (18)]:

$$\mathtt{P}^c_{|\mathtt{R}|}(\gamma_{\mathtt{th}}, \mathtt{q}_{\mathtt{th}}) = \mathbb{E}[\mathtt{Q}^c](1-F(\mathtt{q}_{\mathtt{th}}))^{|\mathtt{R}|-1} + \mathbb{E}[\mathtt{Q}^c \mid \mathtt{Q}^c \leq \mathtt{q}_{\mathtt{th}}]\Big(1-(1-F(\mathtt{q}_{\mathtt{th}}))^{|\mathtt{R}|-1}\Big). \quad (15)$$

These expressions indicate that, for a perfectly calibrated predictor, the outage probability in the single-resource case equals the overall expected confidence score, as shown in (13), whereas (14) shows that for a sufficiently large number of resources, $\mathtt{q}_{\mathtt{th}}$ should be set so that the expected confidence, conditioned on $\mathtt{Q}^c \leq \mathtt{q}_{\mathtt{th}}$, matches the target outage probability. Furthermore, the following inequalities from [21, eq. (19) and (20)] offer practical criteria for choosing appropriate $\mathtt{q}_{\mathtt{th}}$:

$$\mathtt{P}^c_{\infty}(\gamma_{\mathtt{th}}, \mathtt{q}_{\mathtt{th}}) \leq \mathtt{q}_{\mathtt{th}}, \quad (16)$$

$$\text{and } \mathtt{P}^c_{|\mathtt{R}|}(\gamma_{\mathtt{th}}, \mathtt{q}_{\mathtt{th}}) \leq \mathbb{E}[\mathtt{Q}^c](1-F(\mathtt{q}_{\mathtt{th}}))^{|\mathtt{R}|-1} + \mathtt{q}_{\mathtt{th}}\left(1-(1-F(\mathtt{q}_{\mathtt{th}}))^{|\mathtt{R}|-1}\right). \quad (17)$$

Equation (16) shows that, for perfectly calibrated predictors and a large number of resources, setting $\mathtt{q}_{\mathtt{th}}$ equal to the desired outage probability results in a system outage probability that is better than their desired outage probability.

While calibration enhances the alignment between predicted confidence and true accuracy, it does not increase the underlying information used for prediction. As a post-processing procedure, it cannot improve informativeness or lower the minimum achievable outage probability. Formally, for any calibration method satisfying (7), from [21, Theorem 2], we have:

$$\min_{\mathtt{q}_{\mathtt{th}}} \mathtt{P}^c_{|\mathtt{R}|}(\gamma_{\mathtt{th}}, \mathtt{q}_{\mathtt{th}}) \geq \min_{\mathtt{q}_{\mathtt{th}}} \mathtt{P}_{|\mathtt{R}|}(\gamma_{\mathtt{th}}, \mathtt{q}_{\mathtt{th}}). \quad (18)$$

IV. Simulation Results

This section describes the data generation process, details the calibration methods applied, and presents the results showing how calibrated predictions affect outage performance and support the selection of $\mathtt{q}_{\mathtt{th}}$ for meeting reliability targets.

A. Data Generation

The dataset used for training and testing is synthetically generated based on the Rician fading model described in Section II. Data generation proceeds as follows:

1) At time $t = 0$, generate a tapped-delay line with $v = 1024$ complex entries, where each tap is independently drawn from a zero-mean complex Gaussian distribution with variance $\frac{\Omega}{v(\mathrm{K}+1)}$.
2) For the first tap, add a deterministic LoS component $\sqrt{\frac{\Omega\mathrm{K}}{\mathrm{K}+1}}e^{j(2\pi f_D t+\phi)}$, where Ω is the total power per tap, $\phi \sim \mathrm{Unif}[0, 2\pi]$ and $f_D = 0.01$ is the normalized Doppler frequency. This f_D corresponds to a physical Doppler shift of 10–100 Hz at sampling intervals of 0.1–1 ms (5.8 GHz carrier), consistent with user velocities of 0.5–5 m/s (typical in pedestrian and low-mobility scenarios.) Each scattered component evolves as $e^{jt\theta_i}$, with θ_i independently drawn from ± 0.1 radians.
3) Perform a FFT to the time-domain vector to obtain a frequency-domain vector of length 1024.
4) Extract $|\mathtt{R}| \leq v$ equally spaced frequency-domain samples that represent resources at a fixed point in time.

DQN-LSTM Model	
Architecture	**Hyperparameters**
Layer 1: LSTM layer (32 hidden units)	Hidden units : 32 Epochs : 30 Epoch size : 150 Input sequence length (k) : 100 Output sequence length (l) : 10 Output dimension : 1 Learning rate : 0.001 Discount factor : 0.9 Epsilon : Decaying strategy Rewards : 1 (correct prediction); -1(otherwise)
Layer 2: Dense layer (10 units with PReLU activation)	
Layer 3: Dense layer (10 units with PReLU activation)	

Fig. 2. DQN-LSTM Model Architecture and Hyperparameters.

5) Repeat steps 2 – 4 for $k+l$ time steps to generate k input and l output samples per resource.

B. Training and Evaluation

We train a Deep Q-Network (DQN) with an integrated long short-term memory (LSTM) layer, following the architecture in [27], using both the OLF [14] (with $\alpha = 10$) and BCE. The model is implemented in TensorFlow (Keras) and optimized using ADAM. Architectural and training details are listed in Fig 2. Each experiment involves generating sequences of $k+l$ frequency-domain samples per resource. The first k samples are processed by the DQN-LSTM model, while the remaining l samples, denoted by $\vec{\boldsymbol{h}}^l_j(t)$, are used to evaluate whether the target communication rate is achievable. To demonstrate the outage probability behavior established under perfect calibration in Section III, we apply two standard post-processing calibration methods: beta scaling and isotonic regression. Isotonic regression is implemented using the `IsotonicRegression` class from *scikit-learn*, beta calibration is performed using the `BetaCalibration` class from the *betacal* library, and *SciPy* is used to support numerical computations.

1) ***Beta Calibration:*** This parametric method models predicted probabilities using a beta distribution, making it effective for handling skewed or extreme outputs. The calibrated probability $\hat{q}_i$ is computed as:
$$\hat{q}_i = \frac{1}{1+\exp\left(c - a\ln(s) - b\ln(1-s)\right)}, \quad (19)$$
where s is the uncalibrated probability, and a, b, and c are parameters learned via maximum likelihood on a validation set [18]. Beta calibration is implemented using the `BetaCalibration` class from the *betacal* library. The model is trained using the `fit` method and applied via `predict` to convert uncalibrated outputs to calibrated probabilities.
2) ***Isotonic Regression:*** This non-parametric calibration method fits a non-decreasing, piecewise-constant function $f(\cdot)$ that maps uncalibrated probabilities $\hat{p}_i$ to calibrated outputs $\hat{q}_i = f(\hat{p}_i)$. The function f is learned by minimizing the empirical squared error:

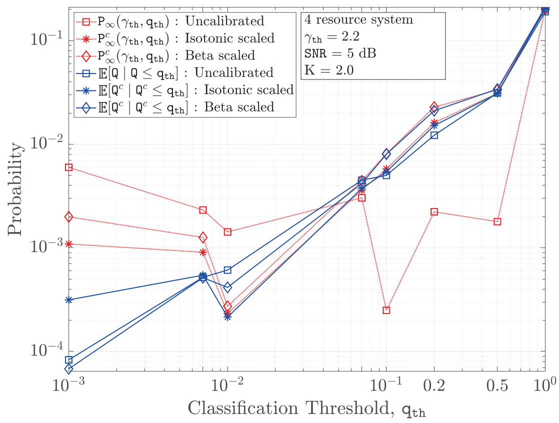

Fig. 3. Comparison of $P^\star_\infty(\gamma_{th}, q_{th})$ and $\mathbb{E}[Q^\star \mid Q^\star \leq q_{th}]$ over varying q_{th} values, in a 4-resource system using OLF with $\gamma_{th} = 2.2$, SNR = 5 dB, and K = 2.0.

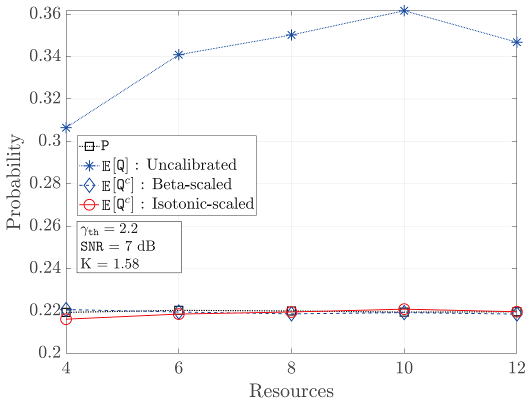

Fig. 4. Comparison of P and $\mathbb{E}[Q^\star]$ across different numbers of resources R, using OLF with $\gamma_{th} = 2.2$, SNR = 7 dB, and K = 1.58.

$$\sum_{i=1}^{n} \left(f(\hat{p}_i) - y_i\right)^2 , \qquad (20)$$

where y_i denotes the true label. This is implemented using the IsotonicRegression class in *scikit-learn*. Once trained on a validation set, the learned function f is used to recalibrate new predictions.

C. Results

Figs. 3 and 4 showcase the core findings of this work, illustrating how a perfectly calibrated predictor behaves, as discussed in Section III. These plots do not compare the outage probability achieved by BCE-trained predictors with those trained using OLF, as q_{th} is not a hyperparameter in BCE-based training. In particular, Fig. 3 plots the outage probability for an infinite resource system, $P^\star_\infty(\gamma_{th}, q_{th})$ along with the conditional expected confidence $\mathbb{E}[Q^\star \mid Q^\star \leq q_{th}]$ for varying q_{th} values. Results are shown for a system with 4 resources, $\gamma_{th} = 2.2$, SNR = 5 dB, and K = 2.0. It can be observed that for calibrated predictors, the two curves align closely, in agreement with (14).

This result guides the selection of q_{th} to meet a target outage probability. For instance, a service provider with a service level agreement requiring outage probability below 0.03 (i.e., fewer than 2 outages per 60 transmissions) can use the calibrated predictor to select an appropriate q_{th}. Given $\gamma_{th} = 2.2$, SNR = 5 dB, K = 2.0, and a sufficiently large number of resources, Fig. 3 indicates that $q_{th} = 0.5$ satisfies this service level agreement. This example holds only for calibrated predictors. For uncalibrated predictors, q_{th} is chosen empirically, since (14) no longer holds. A key observation is that, both calibrated and uncalibrated OLF-trained predictors achieve the same minimum outage probability, but at different q_{th} values. This is because calibration shifts the value of q_{th} by aligning confidence outputs with actual outage likelihood.

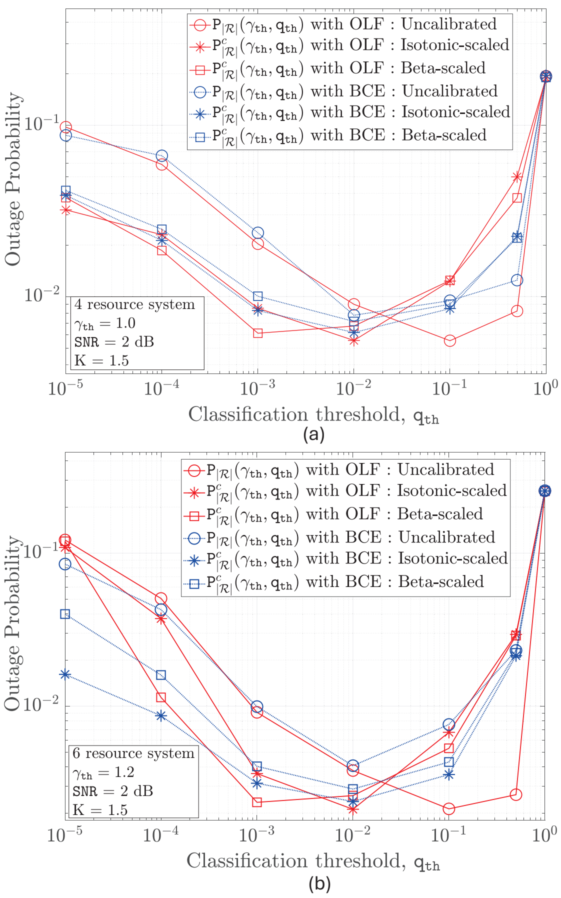

Fig. 5. $P^\star_{|R|}(\gamma_{th}, q_{th})$ vs q_{th} evaluated using OLF and BCE for K = 1.5 and SNR = 2 dB, with (a) $\gamma_{th} = 1.0$ in a 4-resource system, and (b) $\gamma_{th} = 1.2$ in a 6-resource system.

Fig. 4 presents a comparison between the outage probability observed in a single-resource setting, denoted by P, and the average predicted confidence $\mathbb{E}[Q^\star]$. The results are shown for varying numbers of resources, with system configuration fixed at $\gamma_{th} = 2.2$, SNR = 7 dB, and K = 1.58. As shown, the calibrated predictor yields $\mathbb{E}[Q^c]$ values that closely match the observed outage probability for a single resource, consistent with the behavior suggested by (13).

Note that for finite K, the channel exhibits stochastic variation, so calibration helps align the predictor's confidence with actual outcomes. As K→∞, the channel becomes fully deterministic, and the predictor's confidence inherently reflects the ground truth perfectly, without additional adjustment. This behavior can be empirically validated by reliability diagrams, which would exhibit perfect alignment between predicted confidences and observed outcomes.

Fig. 5 (a) and (b) presents the outage probability obtained using OLF and BCE, for varying $\mathtt{q_{th}}$ values, with $\mathtt{SNR} = 2$ dB, $\mathrm{K} = 1.5$ for (a) $\gamma_{\mathtt{th}} = 1.0$ in a 4-resource system and (b) $\gamma_{\mathtt{th}} = 1.2$ in a 6-resource system. Both subfigures in Fig. 5 exhibit a characteristic 'U'-shaped outage curve, indicating that very low or very high thresholds degrade performance. Very small $\mathtt{q_{th}}$ values tend to result in the user almost always being allocated the final resource, whereas very large values lead to allocation to the first resource. Additionally, it can be observed that the minimum outage probability achieved by the OLF-trained predictors, both calibrated and uncalibrated, is the same but is obtained at different values of $\mathtt{q_{th}}$, aligning with the behavior suggested in (18).

V. Conclusion

We explored the calibration accuracy of an ML-based outage predictor in the context of resource allocation in Rician fading environments. The analysis began by presenting outage probability expressions formulated under the assumption of perfect calibration. Through Monte Carlo simulations on a Rician fading channel, we verified that, in scenarios with sufficiently large resources, the outage probability corresponds to the expected value of the predictor's output, provided it was below the threshold used for classification. Conversely, when only a single resource is available, the outage probability simplifies to the predictor's overall expected confidence. These insights can support service providers in selecting appropriate threshold values that satisfy reliability targets defined in service-level agreements. Notably, for finite K, channel uncertainty makes calibration effective; however as $\mathrm{K} \to \infty$, the predictor's confidence aligns with outcomes by design and calibration becomes ineffective. Lastly, we demonstrated that post-processing calibration methods do not improve the minimum achievable outage probability, as they offered no additional information about future channel conditions.

References

[1] K. M. Cohen, S. Park, O. Simeone, and S. Shamai Shitz, "Calibrating AI Models for Few-Shot Demodulation via Conformal Prediction," in *ICASSP 2023 - 2023 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.

[2] R. Raina, N. Simmons, D. E. Simmons, and M. D. Yacoub, "Beyond Linear Binning: Logarithmic Insights for Calibrated Machine Learning in Wireless Systems," in *2024 IEEE 100th Veh. Technol. Conf. (VTC2024-Fall)*, Washington, DC, USA, 2024, pp. 1–6.

[3] S. Moon, H. Kim, Y.-H. You, C. H. Kim, and I. Hwang, "Online Learning-Based Beam and Blockage Prediction for Indoor Millimeter-Wave Communications," *ICT Express*, vol. 8, no. 1, pp. 1–6, Mar. 2022.

[4] L. Sanguinetti, A. Zappone, and M. Debbah, "Deep Learning Power Allocation in Massive MIMO," in *52nd Asilomar Conf. on Signals, Syst., and Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 1257–1261.

[5] T. Zhang, K. Zhu, and J. Wang, "Energy-Efficient Mode Selection and Resource Allocation for D2D-Enabled Heterogeneous Networks: A Deep Reinforcement Learning Approach," *IEEE Trans. on Wireless Commun.*, vol. 20, no. 2, pp. 1175–1187, Feb. 2021.

[6] R. Raina, D. E. Simmons, N. Simmons, and M. D. Yacoub, "Optimal Classifier for an ML-Assisted Resource Allocation in Wireless Communications," *IEEE Netw. Lett.*, vol. 6, no. 3, pp. 158–162, 2024.

[7] Y. Sun, M. Peng, and S. Mao, "Deep Reinforcement Learning-Based Mode Selection and Resource Management for Green Fog Radio Access Networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1960–1971, 2019.

[8] N. Simmons, D. E. Simmons, R. Raina, and M. D. Yacoub, "A Custom Loss Function for Machine Learning-based Resource Allocation Policies," in *2024 IEEE Int. Conf. on Mach. Learn. for Commun. and Netw. (ICMLCN)*, 2024, pp. 19–24.

[9] C. She, C. Sun, Z. Gu, Y. Li, C. Yang, H. V. Poor, and B. Vucetic, "A Tutorial on Ultrareliable and Low-latency Communications in 6G: Integrating Domain Knowledge Into Deep Learning," *Proc. of the IEEE*, vol. 109, no. 3, pp. 204–246, March 2021.

[10] T. Peken, S. Adiga, R. Tandon, and T. Bose, "Deep Learning for SVD and Hybrid Beamforming," *IEEE Trans. on Wireless Commun.*, vol. 19, no. 10, pp. 6621–6642, June 2020.

[11] K. Utkarsh, Ashish, and P. Kumar, "Transmit Power Reduction in an IRS Aided Wireless Communication System using DNN," in *2023 Int. Conf. on Microw., Opt., and Commun. Eng. (ICMOCE)*, Bhubaneswar, India, May 2023, pp. 1–5.

[12] N. Rajapaksha, K. B. S. Manosha, N. Rajatheva, and M. Latva-aho, "Unsupervised Learning-based Joint Power Control and Fronthaul Capacity Allocation in Cell-Free Massive MIMO with Hardware Impairments," *IEEE Wireless Commun. Lett.*, vol. 12, no. 7, pp. 1159–1163, April 2023.

[13] C. Sun and C. Yang, "Unsupervised Deep Learning for Ultra-Reliable and Low-Latency Communications," in *IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[14] N. Simmons, D. E. Simmons, and M. D. Yacoub, "Outage Performance and Novel Loss Function for an ML-Assisted Resource Allocation: An Exact Analytical Framework," *IEEE Trans. on Mach. Learn. in Commun. and Netw.*, vol. 2, pp. 335–350, 2024.

[15] E. Angelino, M. J. Johnson, and R. P. Adams, "Patterns of Scalable Bayesian Inference," *Foundations and Trends in Mach. Learn.*, vol. 9, no. 2-3, pp. 119–247, 2016.

[16] K. M. Cohen, S. Park, O. Simeone, and S. Shamai Shitz, "Calibrating AI Models for Wireless Communications via Conformal Prediction," *IEEE Trans. on Mach. Learn. in Commun. and Netw.*, vol. 1, pp. 296–312, 2023.

[17] B. Zadrozny and C. Elkan, "Transforming Classifier Scores into Accurate Multiclass Probability Estimates," in *Proc. of the eighth ACM SIGKDD Int. Conf. on Knowl. Discovery and Data Mining*, 2002, pp. 694–699.

[18] M. Kull, T. Silva Filho, and P. Flach, "Beta Calibration: A Well-Founded and Easily Implemented Improvement on Logistic Calibration for Binary Classifiers," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 623–631.

[19] M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach, "Beyond Temperature Scaling: Obtaining Well-Calibrated Multi-Class Probabilities with Dirichlet Calibration," *Advances in Neural Inf. Process. Syst.*, vol. 32, 2019.

[20] R. Raina, N. Simmons, D. E. Simmons, M. D. Yacoub, and S. L. Cotton, "Impact of Platt Scaling on Calibration in ML-based Wireless Resource Allocation," Invited paper, presented at the *2025 IEEE Int. Conf. on Mach. Learn. for Commun. and Netw. (ICMLCN),* Barcelona, Spain, May 26-29, 2025.

[21] R. Raina, N. Simmons, D. E. Simmons, M. D. Yacoub, and T. Q. Duong, "To Trust or Not to Trust: On Calibration in ML-based Resource Allocation for Wireless Networks," 2025. [Online]. Available: https://arxiv.org/abs/2507.17494

[22] T. ETSI, "Study on channel model for frequencies from 0.5 to 100 ghz," *138 901 v16. 1.0, 5G*, 2020.

[23] R. Clarke and W. L. Khoo, "3-D Mobile Radio Channel Statistics," *IEEE Trans. on Veh. Technol.*, vol. 46, no. 3, pp. 798–799, Aug. 1997.

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

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. of the 34th Int. Conf. on Mach. Learn.*, 2017, pp. 1321–1330.

[26] T. M. Cover, *Elements of Information Theory*. John Wiley & Sons, 1999.

[27] R. Raina, N. Simmons, D. E. Simmons, and M. D. Yacoub, "ML-Assisted Resource Allocation Outage Probability: Simple, Closed-Form Approximations," in *2023 IEEE Int. Conf. on Adv. Netw. and Telecommun. Syst. (ANTS)*, Jaipur, India, Dec. 2023, pp. 1–6.