CapAware: Capacity-Aware Uplink Bandwidth Prediction for Cellular Networks

Birkan Denizer[1], Olaf Landsiedel[1,2]
[1]*Kiel University, Germany*
[2]*Hamburg University of Technology, Germany*
birkan.denizer@cs.uni-kiel.de, olaf.landsiedel@tuhh.de

***Abstract*—As remotely controlled and autonomous vehicles become widely available, the demand for high Quality of Service over cellular networks for their remote control and monitoring is becoming increasingly important. Accurate prediction of available uplink bandwidth is essential to mitigate bandwidth fluctuations and avoid impacting real-time applications, ensuring reliable and low-latency video streams. In particular, bandwidth overpredictions lead to packet losses, retransmissions, and significant latency increases, especially during network handovers, as network buffers fill up. Prior bandwidth prediction approaches lower absolute or relative errors but fail to address the impacts of overpredictions and the associated latency spikes.**

This paper introduces CapAware, a bandwidth prediction approach explicitly designed to minimize capacity violations (i.e., overpredictions) and reduce latency spikes during network handovers for uplink streams. It utilizes an efficient neural network architecture with an integrated handover prediction mechanism and a learnable capacity-aware loss function. CapAware predicts network handovers with a 92.4% F1 score and improves efficiency by 24.4% using its custom loss function with predicted handover information. Compared to deep-learning baselines, CapAware improves network efficiency (i.e., utilization-to-capacity violation ratio) by 4.7% and 34.9% on 5G SA datasets.

***Index Terms*—Bandwidth prediction, handover prediction, capacity-aware, utilization, overprediction, 5G**

I. Introduction

Robust cellular connectivity is increasingly becoming vital for remote-controlled and autonomous vehicles [1]–[3]. Remote control and monitoring systems rely primarily on the uplink, for example, for video streams and data collection from edge locations, while the downlink is less commonly used. Reliable uplink throughput is crucial to the Quality of Service (QoS) that these applications require [1], [4]. However, existing bandwidth prediction approaches suffer from several key limitations: the use of symmetric loss functions, a lack of adaptation against network handovers, and a reliance on metrics that require modifications for low-level access.

First, existing prediction models mainly use symmetric loss functions, which are designed to minimize absolute or relative errors. While intuitive, such approaches inadequately represent real-world conditions, as they treat overpredictions and underpredictions equally. In practice, full buffer utilization and overprediction of available bandwidth are significantly more damaging, resulting in increased latency and packet loss due to overloaded buffers [5], [6]. Second, existing works on uplink bandwidth prediction largely overlook the significant impact of network handovers, events that often disrupt bandwidth continuity. For example, during handovers, the average latency increases by a factor of 2.26x and there is a 14% drop in bandwidth [7]. Therefore, predictions that fail to account for upcoming handovers are particularly prone to bursty errors due to drastic changes in bandwidth, worsening QoS. Third, many existing prediction methods rely on metrics that require either modifications for low-level access or specialized hardware support, rendering them impractical for general adoption [8], [9]. Beyond these limitations, existing downlink bandwidth prediction models increase bandwidth usage to react quickly to growing capacity, especially following network handovers [7]. However, on the uplink side, issues arise from excessive bandwidth allocation during both non-handover and handover scenarios; thus, existing downlink models are not well-suited for uplink predictions.

In this paper, we present CapAware, a bandwidth prediction approach designed to minimize capacity violations while maximizing network utilization. Central to our approach are (1) a capacity-aware loss function, created to minimize capacity violations by penalizing overpredictions heavily than underpredictions, a departure from standard symmetric loss functions used in prior bandwidth prediction models, and (2) a proactive handover prediction mechanism, which mitigates the negative impacts of handovers on bandwidth prediction. We collect a new 5G Standalone (SA) dataset that provides commonly available features and analyze the dataset regarding bandwidth and the impacts of network handovers. CapAware uses only network measurements widely available through standard APIs (e.g., Android).

Overall, in this paper, we make the following contributions:

1) We design CapAware[1], an approach that predicts uplink bandwidth using an integrated handover prediction model and a custom loss function, specifically designed to minimize capacity violations while reducing latency spikes, especially experienced during network handovers.
2) We collect a new dataset on a public 5G SA network with commonly available features. We analyze our 5G SA dataset alongside a publicly available 5G SA dataset, examining the effects of handover events on bandwidth and latency, see Section III.

[1]https://github.com/ds-kiel/CapAware

This is the author's accepted manuscript of an article published by IEEE.
The final authenticated version is available online at: https://doi.org/10.1109/LCN65610.2025.11146351
DOI: 10.1109/LCN65610.2025.11146351

3) We evaluate our handover prediction model and demonstrate that it achieves a 92.4% F1 score, see Section V-C. Our custom loss function, which utilizes predicted handover information, achieves 24.4% and 11.6% higher network efficiency than the Quantile and MSE loss functions, respectively.
4) We evaluate CapAware against SOTA models and show CapAware improves network efficiency up to 34.9% and 5.93% on 5G SA datasets, respectively, see Section V-D.

The remainder of this paper is structured as follows: Section II presents related work in the fields of handover prediction and the use of loss functions, Section III introduces the measurement setup and analysis of collected and public datasets, Section IV introduces the design and system model of CapAware, Section V evaluates CapAware on collected and public datasets, and Section VI concludes the paper.

II. Related Work

In this section, we first discuss related work on handover prediction and then loss functions for bandwidth prediction.

Handover Prediction Prior works commonly utilize network and localization information to predict network handovers [7], [10]–[14]. Lima et al. [13] uses only reference signal received power (RSRP) as input feature to a Long Short-Term Memory (LSTM) model as the first stage to predict future RSRP values and then uses a Random Forest (RF). Langolf et al. [14] predict handovers on 5G NSA networks using LSTM and Convolutional Neural Network (CNN) models on real-world data using reference signal (RS) information and localization information such as latitude and longitude. Prognos [7] predicts handovers on LTE and 5G networks by using low-level Radio Resource Control (RRC) messages and RS information, improving downlink bandwidth prediction during network handovers.

Unlike previous works, CapAware does not require localization features or RRC messages; therefore, it is compatible with commonly available APIs. CapAware relies solely on RS information to predict network handovers.

Symmetric vs. Custom Loss Functions for Predictions Today, most works on bandwidth prediction use symmetric loss functions (i.e., MSE, RMSE, MAE, etc.) that compare prediction quality in terms of absolute or relative errors [8], [9], [15]–[19]. PERCEIVE [8] uses an LSTM design for bandwidth predictions on LTE networks using transport block size (TBS), the number of resource blocks (RB), and RSRP information. PERCEIVE utilizes the Pinball Loss function to target a 0.45 quantile in its design, aiming for a slightly lower bandwidth with a negative offset while reducing over-predictions. SURE [9] utilizes a Transformer-based model with the MSE loss function for predicting the average TBS on 5G NSA networks by utilizing prior TBS, RB, RSRP, and the transmission power (Tx-Power) information from a rooted phone with special access to hardware. UplinkNet [19] utilizes ConvLSTM and LSTM with the MSE loss function to predict uplink bandwidth in 5G SA networks by leveraging commonly available features.

TABLE I: Comparison of prediction approaches: Prior works mostly do not predict network handovers and mainly use symmetric loss functions. However, CapAware predicts network handovers and uplink bandwidth using commonly available features and an asymmetric learnable custom loss function.

Model	Uplink Prediction	Handover Prediction	Loss Function	Input Features
Prognos [7]	No	Yes	NA	Proprietary
Perceive [8]	Yes	No	Asymmetric	Proprietary
SURE [9]	Yes	No	Symmetric	Proprietary
UplinkNet [19]	Yes	No	Symmetric	Common
CapAware	**Yes**	**Yes**	**Learnable**	**Common**

TABLE II: Overview of captured features

Data Category	Features
Position	Latitude, Longitude, Speed
Network deployment	Cell ID, Band (e.g., n78 with 80Mhz)
PHY metrics	CQI, RSRP (dBm), RSRQ (dB), SINR (dB)
Rx/Tx information	Rx/Tx bits per second

Unlike previous works, CapAware employs an asymmetric learnable custom loss function to reduce capacity violations. CapAware predicts uplink bandwidth using predicted handover information and commonly available features, and thus does not require modifications to software or hardware, see Table I for comparison of selected models.

III. Datasets

In this section, we first introduce our data measurement campaign. Next, we analyze the collected and publicly available datasets, examining handover events. As an overview, we address the following challenges:

- There is a limited number of 5G SA datasets with high feature correlations based on standard APIs. In Section III-A, we collect measurements on a 5G SA network and then analyze the collected data as well as a prior publicly available 5G SA dataset, in Sections III-B and III-C.
- Prior literature gives limited insights into the effects of network handovers on bandwidths. In Section III-D, we present an initial analysis of the impact of network handovers, followed by a detailed description of our handover prediction model in Section IV-B.

A. 5G SA Measurement Campaign

We collect GPS-labeled 5G SA network measurements from a mobile network operator (MNO) for over 6 months at a 1 Hz sampling interval in Kiel, Germany. We base our measurement campaign on a similar architecture outlined in the Fjord5G dataset [20]. We use a MikroTik Chateau 5G R16[2] router, which relies on Quectel RG520F-EU 5G sub-6GHz modules with 3GPP Release 16 support. We record the network status every second, including GPS position and baseband information on a public Vodafone 5G SA network, see Table II.

[2]https://mikrotik.com/product/chateau5gr16

TABLE III: Statistics of available bands

Metrics	Band n28	Band n3	Band n78
Duplex	FDD	FDD	TDD
Channel BW (MHz)	10	25	80
Percentage of data (%)	15	38	46
Mean (Mbps)	29	61	53
Max (Mbps)	68	121	178
% of Handovers	9	22	16

TABLE IV: Statistics of selected datasets

Metrics	Our 5G SA	UplinkNet 5G SA
Duration (seconds)	174505	120431
Mean (Mbps)	52.84	15.71
Min (Mbps)	0	0
Max (Mbps)	178.5	109.08
SD (Mbps)	29.04	16.02
Coefficient of Variation	54.95%	101.99%
# of Handovers	29777	NA

B. Dataset Overview

In this section, we give an overview of selected datasets. To ensure a practical, real-world applicability, we focus on features accessible through commonly available APIs such as CQI, SINR, RSRP, RSRQ, and ARFCN metrics. We select data points with valid bandwidth measurements and associated network data from selected datasets. We use forward-fill imputation to fill in the missing data points.

Our 5G SA dataset comprises traces with over 20 features from multiple ferries. We select 171,000 data points with bandwidth measurements in the uplink direction corresponding to slightly more than 47 hours. The handover dataset comprises over 9 million data points and 30,000 handover events for moving and stationary scenarios. We present a detailed analysis of our dataset in the next sections, III-C and III-D.

The UplinkNet dataset comprises traces from driving, walking, and train rides, featuring 10 recorded features in Japan and Thailand [19]. The dataset comprises 120,000 data points, equivalent to roughly 33 hours. The UplinkNet dataset does not include the CQI or handovers. We do not present a detailed analysis of the UplinkNet dataset as it exists in its publication.

C. Bandwidth Analysis

In this section, we analyze the uplink bandwidth in our dataset. Our dataset includes frequency bands n28, n3, and n78, with channel widths ranging from 10 MHz to 80 MHz, resulting in the use of both Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD). While FDD channels provide equal upload and download bandwidth, TDD channels do not offer equal upload and download bandwidth, instead providing an asymmetric 3:1 downlink-to-uplink ratio, as indicated in our data.

Based on our analysis, we recognize that channel bandwidths (i.e., 10 MHz (FDD), 25 MHz (FDD), and 80 MHz (TDD)) do not scale linearly with respect to achievable uplink bandwidth; see Section IV-C for the implications of this.

In Table III, we give statistics about available bands in our dataset. In terms of percentage, Band n78 dominates compared to n3 and n28. The peak uplink throughput of band n78 is 50% higher than band n3 and nearly triple that of band n28. However, band n3 offers more stable uplink bandwidth than band n78 and almost double that of band n28.

Figure 1 shows traces of the selected datasets. We plot the mean, standard deviation (SD), and the coefficient of variation (CV), i.e., the ratio of the SD to the mean. A large coefficient of variation indicates a high degree of randomness in the dataset, while a small one shows a low degree of randomness. We show that our dataset offers moderate relative variability, with a CV of 54%. In contrast, the UplinkNet 5G dataset exhibits high relative variability with a CV of 101%, suggesting a higher degree of randomness in the data, see Table IV.

D. Handover Analysis

In this section, we evaluate the effects of network handovers on latency and bandwidth. Figure 2 illustrates an example network trace from our bandwidth dataset that displays network handovers overlayed on top of uplink bandwidth. Due to network handovers, the achievable uplink bandwidth fluctuates significantly, for example, between 20 and 60 Mbps, especially in the later stages of the shown trace. Meanwhile, the inbound latency remains relatively low due to minimal usage. In contrast, the outbound latency and round-trip time (RTT) increase to 600 ms in some cases in the shown trace, resulting from network congestion and network handovers.

When we examine the frequency of handovers on the same band, n3 experiences double the handovers compared to n28 and almost 35% more handovers than n78, see Table III for details. When we analyze the band transitions, there is only 0.1% of the time a transition to a different band (e.g., from n28 to n3), meaning the connection stays anchored to the same band across base stations as long as possible.

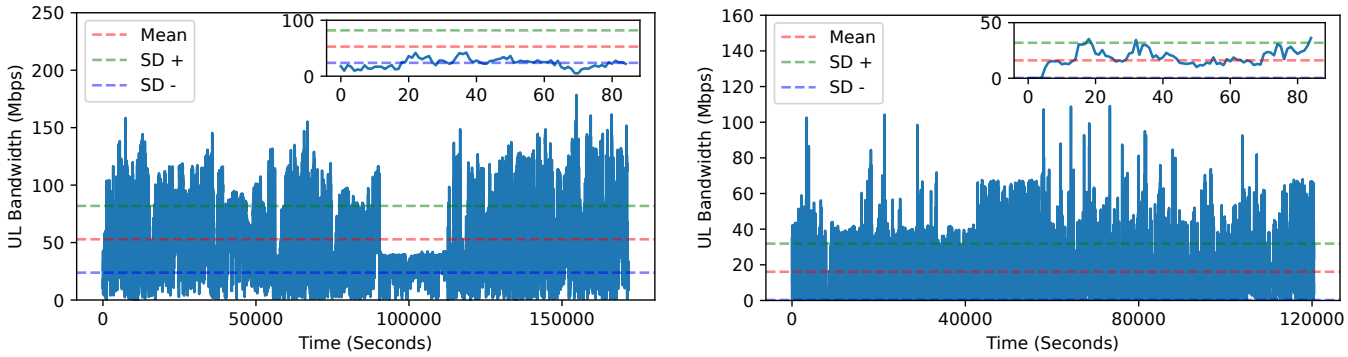
IV. Design

In this section, we introduce the design of CapAware. We begin by providing a high-level overview of the design of CapAware, followed by a detailed description of its pipeline stages. In CapAware, we address the following design challenges:

- Prior works mainly use localization features, RRC messages, or large LSTM models for handover prediction. In Section IV-B, we present an efficient handover prediction model without localization or RRC features.
- Standard symmetric loss functions mainly minimize absolute or relative errors, but do not account for the higher cost of overpredictions (i.e., packet loss and packet delay) compared to underpredictions. In Section IV-C, we propose a domain-specific learnable custom loss function for our bandwidth prediction model.

A. Design Overview

Next, we give an overview of CapAware. It employs a lightweight LSTM-based handover prediction model with



(a) Our 5G SA dataset

(b) UplinkNet 5G SA dataset

Fig. 1: Traces of selected datasets: Our 5G dataset in Fig. a indicates moderate relative variability in the data, as indicated by the coefficient of variation (i.e., the ratio of the standard deviation to the mean), 54.95%. On the other hand, UplinkNet 5G in Fig. b exhibits high relative variability, with a mean of 15.71 Mbps and a standard deviation of 16.02 Mbps, resulting in a coefficient of variation of 101.99%, which suggests higher randomness in the data.

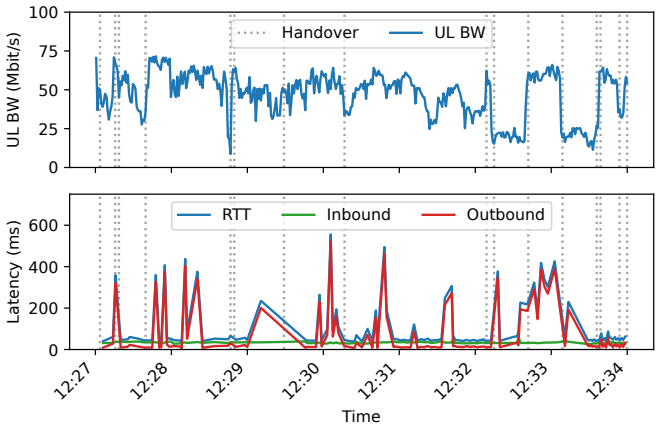


Fig. 2: We show the effects of handover on bandwidth and latency: Handovers cause bandwidth fluctuations, especially towards the later stages of the selected trace. While inbound latency stays very stable, outbound latency and, therefore, the RTT increase during network congestion and handovers.

three input features: RSRP, SINR, and CQI, see Figure 3. Our bandwidth prediction model utilizes a larger LSTM configuration than our handover prediction model, with a custom loss function that takes into account predicted handover information and its prior history, channel band information, and the same input features as our handover predictor. Using the predicted handover information, CapAware avoids bursty bandwidth fluctuations that occur during network handovers, as described in Section I, see Figure 3.

By keeping our prediction models separate, we achieve an efficient architecture for each task while maintaining a simple and flexible design. Our pipeline includes two key stages:

1) **Handover Prediction:** We design a binary classifier with probabilities for handover prediction based on commonly available network features in Section IV-B.
2) **Bandwidth Prediction:** We design CapAware to predict bandwidth based on network features and predicted handover information. CapAware uses a domain-specific custom loss function to avoid capacity violations while maintaining high network utilization in Section IV-C.

We choose an LSTM-based design as its memory cells and gates retain information over long sequences while being computationally efficient [21]. For example, prior studies show that LSTMs capture interactions between multivariate features that simpler models miss, especially in fluctuating environments such as cellular networks [18], [22].

B. Handover Prediction

Feature Selection for Handover Prediction: Prior work on handover prediction utilizes cellular features, such as RSRP and base station identity, as well as localization features, including speed and location, and additional special RRC messages for handover prediction, see Section II.

In contrast, we only use commonly available RSRP, SINR, and CQI metrics in CapAware, without relying on localization or RRC information, to keep our model generic for wide deployment. We present an ablation study of feature selection for handover prediction in Figure 5, see Section V-C.

Handover Prediction Model: We utilize a multi-layer LSTM model for binary classification of the upcoming handovers. Our model predicts a binary output (i.e., handover or no handover) along with the corresponding probability.

Based on an extensive hyperparameter search, we configure our handover prediction model with four LSTM layers, each containing 16 hidden units, followed by a single linear layer and the Binary Cross Entropy loss (BCELoss). We use the last 32 seconds of input as our sequence length. This results in a very lightweight model with around 8k parameters in total. We evaluate our handover prediction model in Section V-C.

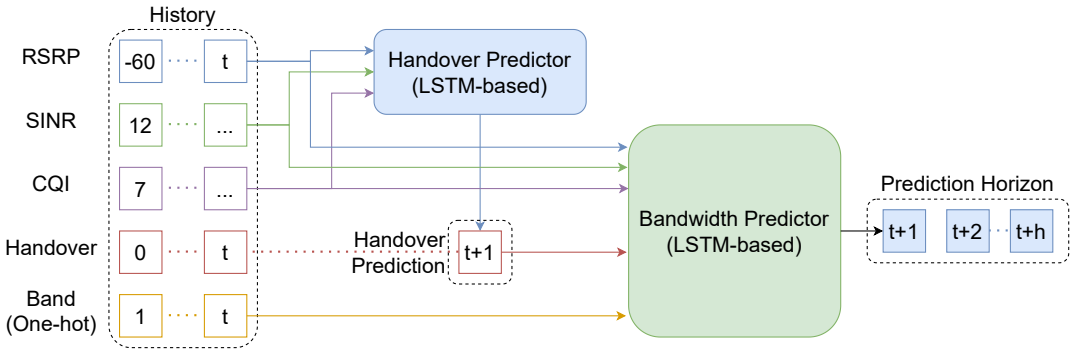


Fig. 3: CapAware (1) predicts handovers using a history of reference signal (RS) information and then (2) utilizes predicted handover information in its custom loss function, along with prior handover history, RS information, and channel band information, to predict uplink bandwidth of a connection for a specified prediction horizon.

TABLE V: Feature correlation results in a range between -1 and 1 with respect to UL bandwidth across datasets: ✓ shows selected features for training.

Features	Our 5G SA dataset	UplinkNet 5G SA
CQI	0.37 ✓	NA
RSRP	0.31 ✓	0.49 ✓
RSRQ	0.30	0.25
SINR	0.37 ✓	0.47 ✓
Band	Variable ✓	Variable ✓

C. Bandwidth Prediction

Feature Selection for Bandwidth Prediction: We analyze the correlation using Pearson, Spearman, and Kendall methods to identify important features for bandwidth prediction [23].

For our dataset, we observe a moderate correlation between features and uplink bandwidth. We choose CQI, SINR, and RSRP, see Table V. We additionally one-hot encode the channel band information decoded from ARFCN instead of using the channel bandwidth (i.e., 25 MHz, 80 MHz) as a numerical range. The primary reason is that FDD and TDD mechanisms do not guarantee a linearly increasing upload channel capacity as the channel bandwidth increases, see Section III-A. We drop one of the one-hot encoded columns to avoid multicollinearity.

On the UplinkNet 5G dataset, we see a moderate correlation. While RSRP and SINR exhibit a medium correlation, RSRQ shows a lower correlation. Thus, we select RSRP and SINR in addition to one-hot encoded band information. The UplinkNet 5G dataset does not provide CQI.

Bandwidth Prediction Model: We use a multi-layer LSTM model for bandwidth prediction. Based on an extensive hyperparameter search, we configure CapAware with three layers of 64 units of LSTM, two linear layers with dropout layers, and our custom loss function. We utilize the Rectified Linear Unit (ReLU) as the activation function of the neural network. We use the last 15 seconds of input as our sequence length. Our selection yields a lightweight model with 89,000 parameters.

Asymmetric Relative Utilization (ARU) Loss Function: CapAware builds on a custom loss function for bandwidth prediction. The key insight is that overpredictions and underpredictions do not affect the quality of the bandwidth prediction in equal proportions. On the one hand, underpredictions are undesirable because they result in the less-than-ideal use of available network resources. However, on the other hand, overpredictions result in highly probable packet loss because the predicted bandwidth exceeds the network capacity. This also has the secondary effect of significantly increasing the resulting latency as the network becomes increasingly congested, and the network buffers fill up. Therefore, we propose an asymmetric relative utilization (ARU) loss function inspired by DeepCog [24] for our bandwidth prediction model.

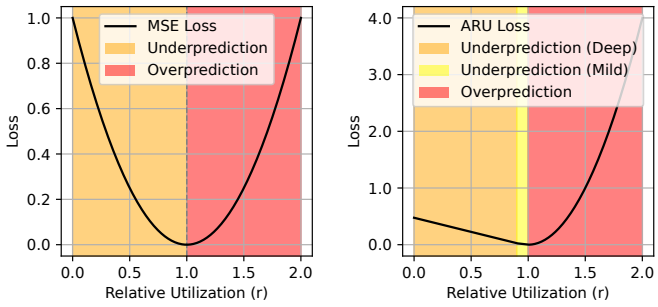
We use relative utilization (i.e., the ratio of prediction to the actual capacity) to create individual regions, such as overprediction, mild underprediction, and deep underprediction, rather than an absolute offset compared to DeepCog. It also uses the predicted handover information to increase the penalty of the overprediction region and lower utilization when a network handover is expected. Thus, ARU loss is independent of absolute network capacity and operates based on the network utilization rate, see Figure 4. Equation 1 with regions (1a), (1b), and (1c) shows how ARU loss works in detail:

$$r_i = \frac{\hat{y}_i}{y_i + \varepsilon}, \qquad \Delta_i = \hat{y}_i - y_i,$$
$$[\Delta_i]^+ = \max(\Delta_i, 0), \qquad [\Delta_i]^- = \max(-\Delta_i, 0).$$

$$\mathcal{L}_{\text{ARU}}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \lambda_{\text{over}} \left([\Delta_i]^+\right)^2, & r_i > 1 \quad (1a) \\ \lambda_{\text{mild}} \, [\Delta_i]^-, & \tau \le r_i \le 1 \quad (1b) \\ \lambda_{\text{deep}} \, [\Delta_i]^-, & r_i < \tau \quad (1c) \end{cases}$$

For the ARU loss, we have the following considerations:

1) **Overprediction** (1a)**:** When predictions exceed capacity ($r_i > 1$), excess traffic causes packet loss. A quadratic penalty, weighted by λ_{over}, discourages this sharply. The predicted handover probability increases the penalty.
2) **Mild Underprediction** (1b)**:** For predictions between the utilization threshold (i.e., 90%) and capacity ($\tau \le$



(a) MSE Loss (b) ARU Loss

Fig. 4: We compare a symmetric loss function and our custom loss function: On the left, the Mean Squared Error (MSE) loss produces equivalent loss for overpredictions and underpredictions. On the right, our asymmetric relative utilization (ARU) loss gives quadratic loss for overpredictions and unique linear losses for each deep and mild underprediction region.

$r_i \leq 1$), we accept minor variations in utilization as tolerable and impose a modest linear penalty λ_{mild}.

3) **Deep Underprediction** (1c)**:** If underprediction stays beneath the threshold τ, resources remain mainly idle. To mitigate this, we apply a stronger linear penalty λ_{deep}.

By defining each region individually, the ARU loss separately handles each condition based on its characteristics. When the prediction exceeds the capacity, it avoids overpredictions with a higher quadratic penalty than the linear penalties used for underpredictions. It utilizes a threshold parameter to distinguish between mild and deep underprediction ranges. The threshold parameter enables targeting, for example, 90% to 100% or 95% to 100% capacity, depending on QoS targets. It penalizes the predictions that fall between the threshold and capacity with a smaller linear penalty than those that fall under the threshold. Therefore, depending on the predicted bandwidth compared to capacity, the ARU loss applies the desired penalty to target the desired region. The overprediction penalty, mild underprediction penalty, deep underprediction penalty, and utilization threshold are fully customizable during training. It is possible to make these four parameters learnable, which we leave for future work.

V. Evaluation

In this section, we evaluate CapAware. First, we introduce baselines and then detail error metrics. Next, we discuss our evaluation results for handover and bandwidth predictions.

A. Baselines and Configuration

We select the following baselines:

1) *PERCEIVE: Deep learning-based Uplink Prediction* [8]
2) *SURE: Self-Attention-Based Uplink Prediction* [9]
3) *UplinkNet: Practical Uplink Prediction* [19]

PERCEIVE (200,000 parameters), SURE (140,000 parameters), and UplinkNet (4,000 parameters) are state-of-the-art models for uplink bandwidth prediction. While PERCEIVE and UplinkNet utilize LSTM layers, SURE employs a Transformer layer to predict the uplink bandwidth, see Section II. We label PERCEIVE configurations as PERCEIVE5, PERCEIVE15, and PERCEIVE50 to reflect the length of the input history.

TABLE VI: We use the following hyperparameters for training from PERCEIVE [8], SURE [9], and UplinkNet [19].

Parameter	CapAware	PERCEIVE	SURE	UplinkNet
Hidden sizes	64	150-100	N/A	16
# of enc layers	3	2	1	2
# of dec layers	2	2	1	2
# of heads	N/A	N/A	1	N/A
Dimension Model	N/A	N/A	128	N/A
Dimension FF	N/A	N/A	256	N/A
Activation	ReLU	ELU	GELU	ReLU
Optimizer	Adam	Adam	RAdam	Adam
Learning rate	0.001	0.001	0.00025	0.0001
Loss function	ARU	Quantile	MSE	MSE
Dropout	0.1	0.5	0.03	0.03
Batch size	32	32	128	256
Sequence length	15	5-15-50	50	5

We select the following configuration for CapAware (89k parameters) after an extensive hyperparameter search. We use Adam as the optimizer with a learning rate of 0.001 and the MinMaxScaler. For the loss function, we utilize the custom ARU loss defined in Section IV-C. We select a batch size of 32, a maximum epoch count of 1000, and an early stopping threshold of 10. For PERCEIVE, we use the Quantile loss function with the quantile set to 0.45. We split selected datasets into 60% training, 20% validation, and 20% test sets for evaluation. Table VI summarizes the parameters for CapAware and baselines.

B. Error Metrics

To show the impact of bandwidth overpredictions and underpredictions, it is not possible to use symmetric metrics such as MSE or MAPE. Thus, we use the following metrics:

1) **Utilization** (%): The share of the available capacity used by the prediction model (higher is better, until 100%).
2) **Area of Violation (AoV)** ($Mbit \cdot s$): The total area of the overpredicted bandwidth (e.g., area under the curve) that cannot be delivered (lower is better).
3) **Volume of Violation (VoV)** ($Mbit^2 \cdot s$): Squared overprediction error to penalize short and sharp violations higher than long and steady ones (e.g., 5 Mbps error for 1 second and 1 Mbps error for 5 seconds give the same AoV but different VoV) (lower is better).
4) **Efficiency Index (EI)** (%): EI score shows how efficiently a loss function or model balances high utilization with low VoV (Normalized to between 0 and 1 using MinMax scale) (higher is better).

C. Handover Prediction Evaluation

In this section, we evaluate our handover prediction model. We show the results of our ablation study for the handover prediction in Figure 5. When we rely solely on RSRP as our

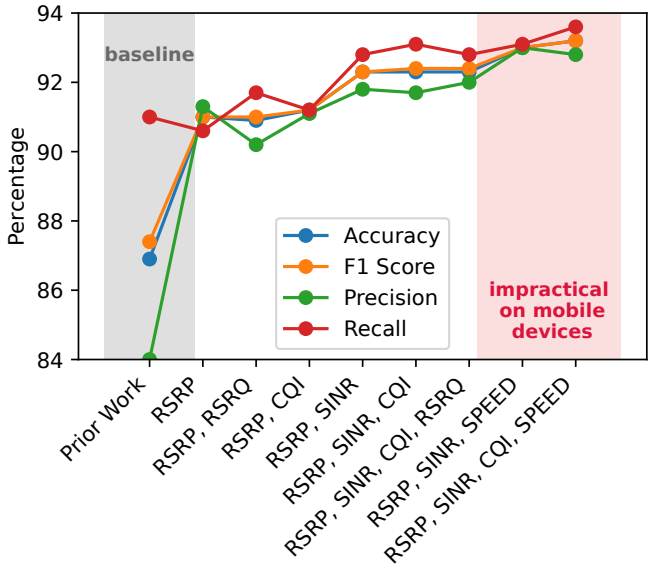


Fig. 5: We demonstrate the impact of feature selection on handover prediction quality in our dataset, showing that our work with RSRP, SINR, and CQI yields the best F1 scores, without the need for localization features. Using the speed metric improves the F1 score at the expense of increased energy consumption, making it impractical for mobile devices.

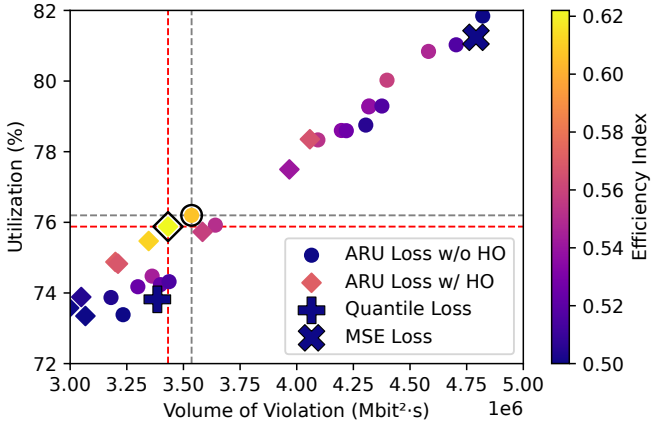


Fig. 6: We compare loss functions on utilization (higher is better) and VoV (lower is better) in our dataset. The ARU loss function without handovers at 76.2% network utilization (the highlighted circle) yields a 21.3% higher EI score (i.e., normalized utilization and violation ratio) compared to the Quantile loss function at 73.8% utilization. With the predicted handover information, the ARU loss function yields 75.9% network utilization (the highlighted diamond) and a 24.4% higher EI score compared to the Quantile loss. The legend color represents the mean EI for its respective loss function.

only feature, we achieve an F1 score of 91%. Adding both SINR and CQI improves results to an F1 score of 92.4%. Unlike SINR and CQI, adding RSRQ does not yield a clear advantage in terms of precision and recall. While adding speed from localization improves F1 scores by 0.8, this requires constant localization usage and significantly increases energy consumption, making it impractical on mobile devices. Thus, we rely on only RSRP, SINR, and CQI as our selected features, which yields a final score of 92.3% for accuracy, 92.4% for F1, 91.7% for precision, and 93.1% for recall.

Compared to prior work [13], which relies solely on the RSRP feature with a larger model having approximately 134,000 parameters, we utilize additional SINR and CQI features with a smaller and more efficient model, which has approximately 8,000 parameters, achieving a better F1 score. Additionally, our model only requires the last 32 steps of the handover history, compared to prior work that requires the last 100 steps.

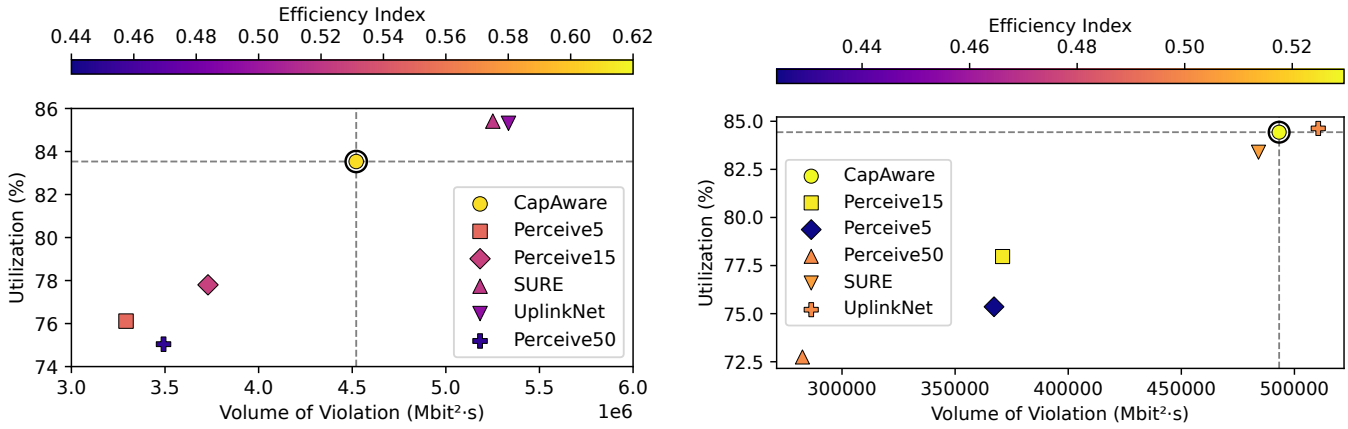
D. Bandwidth Prediction Evaluations

In this section, we evaluate our bandwidth prediction model. We plot utilization and VoV to analyze how an increase in utilization affects overprediction error in Figure 6. Depending on the configuration and loss function, utilization ranges from 72% to 82%. Next, we evaluate the effectiveness of ARU loss against MSE and Quantile loss functions. The symmetric MSE loss function gives an upper bound that does not differentiate between overpredictions and underpredictions. The Quantile loss function provides a simple lower bound with a quantile value of 0.45, which reduces overall utilization while minimizing prediction errors.

We observe that ARU loss configurations yield a better EI score compared to baselines, such as MSE loss (81.2% utilization) and Quantile loss (73.8% utilization) functions. The best configuration of ARU loss without handover information, at 76.2% network utilization, yields a 21.3% higher EI score compared to the Quantile loss functions. The second-best ARU loss configuration without handover achieves 80% network utilization, with an 11.6% higher EI score compared to the MSE loss function. The ARU loss offers the highest efficiency or throughput at either end of the range, with a customizable quadratic overprediction penalty and linear underprediction penalties. Moreover, the addition of handover information improves efficiency to 24.4% at 75.9% network utilization, compared to the Quantile loss function. Thus, the ARU loss mitigates bursty errors by dynamically adjusting utilization during network handovers, see Section IV-C.

Finally, we compare the complete CapAware architecture against selected baselines in Figure 7. We choose the throughput-optimized version of the ARU loss to be competitive against baselines with higher utilization ratios.

In our 5G SA dataset, CapAware offers 22.9% higher efficiency than UplinkNet, 16.9% higher efficiency than SURE, and 34.9% higher efficiency compared to the worst-performing PERCEIVE configuration, achieving 10.3% better efficiency than the best PERCEIVE configuration. An interesting observation is that CapAware offers 15.7% better efficiency even though CapAware and PERCEIVE15 use the same history



(a) Our 5G SA dataset

(b) UplinkNet 5G SA dataset

Fig. 7: We evaluate CapAware against baselines on 5G datasets: In Fig. a, CapAware gives 34.9% and 10.3% higher efficiency than the worst-performing (Perceive50) and the best-performing (Perceive5) baselines. The lack of handover information and CQI metric in the UplinkNet dataset reduces the efficiency advantage of CapAware in Fig. b. CapAware offers similar efficiency to the PERCEIVE15 model with higher utilization and achieves 5.9% and 4.7% better efficiency than the UplinkNet and SURE models, respectively. The legend color for the baselines represents the EI for its respective model.

length. CapAware also reduces VoV by 14% and 15.2% compared to the SURE and UplinkNet models, respectively.

In the UplinkNet 5G SA dataset, Capaware achieves 5.9% better efficiency than UplinkNet, 4.7% better than SURE, and 24.9% better than the worst-performing PERCEIVE configuration, while being indistinguishable from the best-performing PERCEIVE configuration. Only CapAware, SURE, and UplinkNet offer high utilization. Compared to UplinkNet, CapAware reduces VoV by 3.4% at the same utilization level. CapAware offers 1% higher utilization compared to SURE at the same VoV. The lack of handover events reduces the performance advantage of CapAware in this dataset, since our ARU loss function cannot dynamically adjust network utilization. Missing CQI information and high randomness in the data lower the prediction quality, see Section III-C.

VI. Conclusion

This paper introduces CapAware, an approach to predict uplink bandwidth designed to minimize overpredictions while maintaining high network utilization. CapAware predicts upcoming network handovers using an integrated handover prediction model that leverages commonly available network features. It utilizes our learnable ARU loss function, which applies a strong quadratic penalty for overpredictions and unique linear penalties for underprediction regions, see Section IV-C. Using the predicted handover probability, CapAware with its custom loss function, lowers the utilization to avoid high bursty losses during network handovers.

Our evaluation reveals that CapAware predicts network handovers with a 92.4% F1 score. The ARU loss function increases efficiency by 21.3% without handover information and 24.4% with handover information compared to the Quantile loss function. Throughput-optimized configuration of ARU loss achieves nearly the same network utilization as the MSE loss, with 11.6% higher efficiency. Therefore, configurable penalties of the ARU loss allow the user to target either the highest efficiency or the highest throughput.

Overall, the CapAware architecture offers between 4.7% and 34.9% higher efficiency in uplink bandwidth prediction compared to prior deep-learning models. It achieves higher efficiency on 5G datasets where network handover information and standard API metrics are available.

Acknowledgment

This research has been partially funded by the German Ministry of Transport and Digital Infrastructure within the project CAPTN Förde Areal II (45DTWV08D).

References

[1] Ericsson, "Remote operation of vehicles with 5g," extract from the ericsson mobility report, Ericsson AB, Stockholm, Sweden, June 2017. Accessed: 07-May-2025.

[2] C. R. Storck and F. Duarte-Figueiredo, "A survey of 5g technology evolution, standards, and infrastructure associated with vehicle-to-everything communications by internet of vehicles," *IEEE Access*, vol. 8, pp. 117593–117614, 2020.

[3] S. Hakak, T. R. Gadekallu, P. K. R. Maddikunta, S. P. Ramu, C. De Alwis, M. Liyanage, *et al.*, "Autonomous vehicles in 5g and beyond: A survey," *Vehicular Communications*, vol. 39, p. 100551, 2023.

[4] D. Chmieliauskas and S. Paulikas, "Evaluation of uplink video streaming qoe in 4g and 5g cellular networks using real-world measurements," *IEEE Access*, vol. 13, pp. 53996–54018, 2025.

[5] J. Gettys and K. Nichols, "Bufferbloat: Dark buffers in the internet: Networks without effective aqm may again be vulnerable to congestion collapse.," *Queue*, vol. 9, p. 40–54, Nov. 2011.

[6] L. Ruan, M. P. I. Dias, and E. Wong, "Enhancing latency performance through intelligent bandwidth allocation decisions: a survey and comparative study of machine learning techniques," *Journal of Optical Communications and Networking*, vol. 12, no. 4, pp. B20–B32, 2020.

[7] A. Hassan, A. Narayanan, A. Zhang, W. Ye, R. Zhu, S. Jin, J. Carpenter, Z. M. Mao, F. Qian, and Z.-L. Zhang, "Vivisecting mobility management in 5g cellular networks," in *Proceedings of the ACM SIGCOMM 2022 Conference*, SIGCOMM '22, (New York, NY, USA), p. 86–100, Association for Computing Machinery, 2022.

[8] J. Lee, S. Lee, J. Lee, S. D. Sathyanarayana, H. Lim, J. Lee, X. Zhu, S. Ramakrishnan, D. Grunwald, K. Lee, and S. Ha, "Perceive: deep learning-based cellular uplink prediction using real-time scheduling patterns," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, MobiSys '20, (New York, NY, USA), p. 377–390, Association for Computing Machinery, 2020.

[9] J. Jung, S. Lee, J. Shin, and Y. Kim, "Self-attention-based uplink radio resource prediction in 5g dual connectivity," *IEEE Internet of Things Journal*, vol. 10, no. 22, pp. 19925–19936, 2023.

[10] H. Ge, X. Wen, W. Zheng, Z. Lu, and B. Wang, "A history-based handover prediction for lte systems," in *2009 International Symposium on Computer Network and Multimedia Technology*, pp. 1–4, 2009.

[11] M. Ozturk, M. Gogate, O. Onireti, A. Adeel, A. Hussain, and M. A. Imran, "A novel deep learning driven, low-cost mobility prediction approach for 5g cellular networks: The case of the control/data separation architecture (cdsa)," *Neurocomputing*, vol. 358, pp. 479–489, 2019.

[12] L. Mei, J. Gou, Y. Cai, H. Cao, and Y. Liu, "Realtime mobile bandwidth and handoff predictions in 4g/5g networks," *Computer Networks*, vol. 204, p. 108736, 2022.

[13] J. P. Lima, Á. A. de Medeiros, E. P. de Aguiar, E. F. Silva, V. A. de Sousa, M. L. Nunes, and A. L. Reis, "Deep learning-based handover prediction for 5g and beyond networks," in *ICC 2023-IEEE International Conference on Communications*, pp. 3468–3473, IEEE, 2023.

[14] A. Langolf and S. Pachnicke, "Handover prediction for nsa 5g systems in maritime environments using machine learning," in *Mobile Communication - Technologies and Applications; 27th ITG-Symposium*, 2023.

[15] Y. Lin, Y. Gao, and W. Dong, "Bandwidth prediction for 5g cellular networks," in *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, pp. 1–10, 2022.

[16] M. Boban, C. Jiao, and M. Gharba, "Measurement-based evaluation of uplink throughput prediction," in *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, pp. 1–6, 2022.

[17] O. Basit, P. Dinh, I. Khan, Z. J. Kong, Y. C. Hu, D. Koutsonikolas, M. Lee, and C. Liu, "On the predictability of fine-grained cellular network throughput using machine learning models," in *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, pp. 47–56, 2024.

[18] B. Denizer and O. Landsiedel, "Bandseer: Bandwidth prediction for cellular networks," in *2024 IEEE 49th Conference on Local Computer Networks (LCN)*, pp. 1–8, 2024.

[19] K. Arunruangsirilert and J. Katto, "Uplinknet: Practical commercial 5g standalone (sa) uplink throughput prediction," in *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, 2024.

[20] B. Denizer and O. Landsiedel, "Fjord5g: A comprehensive 5g dataset for coastal maritime connectivity," in *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)*, pp. 1–5, 2025.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] Y. Lin, Y. Gao, and W. Dong, "Bandwidth prediction for 5g cellular networks," in *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, pp. 1–10, IEEE, 2022.

[23] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.

[24] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "Deepcog: Cognitive network management in sliced 5g networks with deep learning," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 280–288, 2019.