

Communication-Efficient Parking Recommendation via Federated Learning with Knowledge Distillation in Internet of Vehicles

Kuan-Yi Ke and Chuan-Chi Lai

Department of Communications Engineering, National Chung Cheng University, Taiwan, R.O.C.
Email: sdes730634@alum.ccu.edu.tw; chuanclai@ccu.edu.tw

Abstract—Deploying real-time smart parking prediction in Vehicle-to-Everything (V2X) environments faces significant communication and computational bottlenecks. Conventional Federated Learning (FL) often yields parameter-heavy models unsuitable for resource-constrained vehicular devices. This paper proposes a communication-efficient collaborative learning framework integrating FL with Knowledge Distillation (KD). In this framework, Roadside Units (RSUs) collaboratively train a high-capacity teacher model via FL, which is then distilled into a lightweight student model for efficient inference on On-Board Units (OBUs). A localized multi-criteria recommendation mechanism is further introduced at the edge to provide actionable guidance. Experimental results using real-world datasets show the framework reduces downlink communication payload by 91.60% (from 70.00 KB to 5.88 KB) and reduces transmission latency to 183 ms. Monte Carlo simulations confirm that the localized algorithm accommodates diverse driving preferences and supports real-time decision-making in dynamic Internet of Vehicles (IoV) environments.

Index Terms—Internet of Vehicles (IoV), Smart Parking, Recommendation System, Federated Learning, Knowledge Distillation, Edge Computing, Model Compression.

I. INTRODUCTION

As urbanization continues to intensify, parking congestion has become a persistent challenge in modern cities. Although recent deep learning models, such as LSTM and hybrid CNN-LSTM architectures, have demonstrated strong capability in time-series parking availability prediction [1], [2], [3], [4], their deployment usually requires continuous raw data transmission to central servers, which increases bandwidth consumption, introduces latency, and raises privacy concerns in vehicular environments.

To alleviate these limitations, Federated Learning (FL) [5] has emerged as a promising distributed learning paradigm for Internet of Vehicles (IoV) scenarios [6], [7], [8]. However, conventional FL methods such as FedAvg still require repeated synchronization of full model parameters, which may incur considerable communication overhead and transmission delay in bandwidth-constrained V2X networks [9], [10].

To address this issue, Knowledge Distillation (KD) [11] provides an effective mechanism for transferring the predictive capability of a large model to a compact model with substantially lower computational cost. Integrating FL with KD offers a practical way to preserve the privacy advantage of federated

training while improving deployment efficiency at the network edge.

Motivated by these observations, this study proposes a communication-efficient smart parking recommendation framework that combines FL with KD. In the proposed framework, Roadside Units (RSUs) collaboratively train a high-capacity teacher model through federated learning, and the aggregated teacher model is then distilled into a lightweight student model for deployment on OBUs. Furthermore, a localized multi-criteria recommendation mechanism is incorporated on the OBU, jointly considering predicted vacancy, travel distance, and estimated waiting time to support diverse driving preferences. Experimental results using real-world parking data [12] demonstrate that the proposed framework significantly reduces communication overhead while maintaining competitive prediction performance and enabling personalized parking recommendation.

II. PROBLEM DEFINITION

Deploying real-time smart parking prediction in Internet of Vehicles (IoV) environments requires timely model updates and low-latency inference. However, applying conventional Federated Learning (FL) to the Roadside Unit (RSU) and On-Board Unit (OBU) architecture faces major challenges in both communication and computation.

A. Communication Overhead in FL-based IoV Systems

In conventional FL schemes such as FedAvg, the central server and participating edge nodes repeatedly exchange model parameters during each communication round. Let K denote the number of participating nodes and R the number of communication rounds. The total communication cost can be expressed as

$$\Omega_{\text{FL}} = \sum_{r=1}^R \sum_{k=1}^K \left(|W_k^{(r,\text{up})}| + |W_k^{(r,\text{down})}| \right), \quad (1)$$

where $|W_k^{(r,\text{up})}|$ and $|W_k^{(r,\text{down})}|$ represent the uplink and downlink payloads, respectively. Under the common assumption that each node exchanges the full model with identical size $|W|$ in every round, the above expression can be simplified as

$$\Omega_{\text{FL}} = 2RK|W|. \quad (2)$$

In highly dynamic V2X environments, repeatedly transmitting full model parameters may lead to substantial Radio Access Network (RAN) load and increased transmission delay.

B. Computational Heterogeneity of OBUs

In addition to communication overhead, OBUs are typically resource-constrained and heterogeneous in processing capability. Let f_{OBU} denote the available computing capability of an OBU, and let $C(M)$ denote the computational complexity of model M . The inference latency can be approximated as

$$T_{\text{inf}} = \frac{C(M)}{f_{\text{OBU}}}. \quad (3)$$

For large models, the resulting latency may exceed the timing requirements of real-time parking assistance, especially in fast-moving vehicular scenarios.

C. Latency in Centralized Recommendation

Beyond parking occupancy prediction, recommendation generation in conventional centralized architectures often requires vehicles to continuously upload query-related information to the cloud for decision making. As discussed in [13], frequent uplink transmission increases bandwidth consumption and end-to-end latency. Therefore, enabling recommendation at the OBU side is desirable for delay-sensitive IoV services. However, such a design requires both the prediction model and the decision mechanism to be lightweight enough for on-device execution.

D. Optimization Problem Formulation

Based on the above analysis, the proposed FL+KD framework aims to minimize the downlink dissemination payload and edge inference latency while maintaining prediction accuracy. Note that during the federated training stage, RSUs and the central server still exchange the full teacher model parameters with cost Ω_{FL} as defined in Eq. (2). The key benefit of the proposed framework lies in the *model distribution* phase: instead of delivering the large teacher model to OBUs, only the distilled student model parameters W_S are broadcast. Let K' denote the number of receiving OBUs. The downlink dissemination cost under the proposed framework is

$$\Omega_{\text{down}}^{\text{KD}} = K' |W_S| \quad (4)$$

compared with $\Omega_{\text{down}}^{\text{FL}} = K' |W|$ in the conventional case. We formulate the optimization problem over the student model as:

$$\min_{W_S} \mathbb{E} [\Omega_{\text{down}}^{\text{KD}} + \lambda \cdot T_{\text{inf}}] \quad (5)$$

$$\text{s.t.} \quad \frac{|W_S|}{|W_T|} \leq \rho_{\text{max}} \quad (6)$$

$$R^2(f_S(x, W_S)) \geq R_{\text{req}}^2 \quad (7)$$

$$C(M_S) \leq C_{\text{max}} \quad (8)$$

where λ is a weighting coefficient balancing the communication payload and edge inference time. The first constraint

enforces that the compression ratio does not exceed a prescribed upper bound ρ_{max} . The second constraint guarantees that the predictive fidelity of the student model satisfies the minimum requirement R_{req}^2 . The third constraint bounds the computational complexity to the maximum hardware capacity C_{max} of the edge OBUs.

III. SYSTEM ARCHITECTURE

To address the communication and computational challenges of real-time parking prediction in Vehicle-to-Everything (V2X) environments, this study proposes a three-tier collaborative edge learning architecture that combines Federated Learning (FL) with Knowledge Distillation (KD). As illustrated in Fig. 1, the proposed framework separates global model training, model dissemination, and on-device inference across different system layers, thereby reducing communication cost and supporting low-latency parking recommendation.

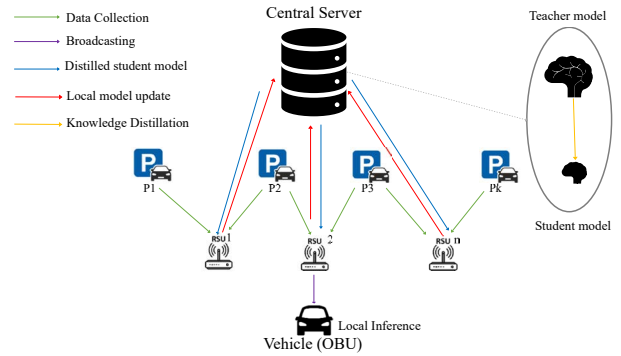


Fig. 1. Proposed communication-efficient V2X smart parking system architecture. RSUs collaboratively train the teacher model through federated learning, and the distilled student model is distributed to OBUs for local inference.

The proposed architecture consists of the following three functional layers.

A. Infrastructure Layer

The infrastructure layer serves as the computational backbone of the proposed system and consists of distributed Roadside Units (RSUs) and a central aggregation server. Each RSU performs local model training using its locally available parking data, while the central server aggregates the uploaded model updates to construct a global teacher model. After global aggregation, the server further performs knowledge distillation to transfer the learned knowledge of the teacher model to a lightweight student model. In this way, the computationally intensive training and model compression tasks are handled at the infrastructure side rather than at resource-constrained end devices.

B. Transmission Layer

The transmission layer is responsible for disseminating model updates from the infrastructure to the vehicular network. In conventional FL-based systems, repeatedly delivering large model parameters may impose a considerable communication

burden in bandwidth-constrained and highly dynamic V2X environments. In the proposed framework, only the distilled student model is delivered to vehicles for inference, which significantly reduces the downlink payload. This design improves the practicality of model distribution under unstable vehicular communication conditions.

C. User Application Layer

The user application layer resides on the On-Board Units (OBUs) of vehicles. OBUs maintain the latest student model received from nearby RSUs and execute parking occupancy prediction locally when a parking request is triggered. Because the student model is substantially smaller than the teacher model, the proposed system reduces on-device inference cost and avoids continuous uplink transmission of user-side queries. This design is suitable for resource-constrained OBUs operating in delay-sensitive Internet of Vehicles (IoV) scenarios.

IV. PROPOSED COMMUNICATION-EFFICIENT FL-KD FRAMEWORK

To reduce communication overhead while maintaining predictive performance in V2X environments, this study adopts a two-stage collaborative learning framework. In the first stage, a high-capacity teacher model is trained through federated learning across distributed RSUs. In the second stage, the global teacher model is compressed into a lightweight student model through knowledge distillation for efficient deployment on OBUs.

A. Federated Training of the Global Teacher Model

To learn global temporal patterns without centralizing raw parking data, a teacher model is trained through federated learning at the infrastructure side. Historical parking vacancy rates are used as input features, and Long Short-Term Memory (LSTM) networks are adopted as the prediction backbone. In order to preserve sufficient temporal modeling capability, the hidden state dimension of the teacher model is set to $H_{\text{teacher}} = 64$.

For global aggregation, the central server adopts a weighted federated averaging strategy. Assume that the i -th participating node contains n_i local samples, and let the total number of samples be $N = \sum_{i=1}^K n_i$. The global model weights W_{global} are updated as

$$W_{\text{global}} = \sum_{i=1}^K \frac{n_i}{N} W_i^{\text{local}}, \quad (9)$$

where W_i^{local} denotes the locally trained model parameters of the i -th node. In the implementation, each participating node performs $E_{\text{local}} = 5$ local training epochs per round, and the server conducts global aggregation over $R = 50$ communication rounds to obtain the final global teacher model.

B. Knowledge Distillation for Student Model Compression

Although federated learning improves privacy preservation by avoiding raw data sharing, the resulting teacher model still remains relatively large for direct deployment on OBUs. To reduce the model size and inference cost, a lightweight student model is trained by distilling knowledge from the global teacher model. In the proposed implementation, the student model uses a smaller hidden dimension of $H_{\text{student}} = 16$.

Since the parking prediction task is formulated as a regression problem, the distillation objective is defined using regression-based losses. Let $\hat{y}_j^T$ and $\hat{y}_j^S$ denote the teacher and student predictions for the j -th sample, respectively, and let y_j denote the corresponding ground truth. The soft distillation loss is defined as the Mean Squared Error (MSE) between the student and teacher predictions:

$$L_{\text{soft}} = \frac{1}{B} \sum_{j=1}^B (\hat{y}_j^S - \hat{y}_j^T)^2, \quad (10)$$

where B is the batch size. In addition, the student model is also supervised by the ground-truth labels through the task loss:

$$L_{\text{hard}} = \frac{1}{B} \sum_{j=1}^B (\hat{y}_j^S - y_j)^2. \quad (11)$$

The overall training objective of the student model is a weighted combination of the above two terms:

$$L_{\text{total}} = \alpha L_{\text{soft}} + (1 - \alpha) L_{\text{hard}}, \quad (12)$$

where α is a balancing coefficient. In this study, $\alpha = 0.7$ is adopted so that the student model learns both the predictive behavior of the teacher model and the supervision from the ground-truth labels.

C. Analysis of Communication Overhead and Compression Rate

A key advantage of the proposed FL-KD framework is that only the lightweight student model needs to be distributed to OBUs for real-time inference. In LSTM-based models, the number of parameters is strongly influenced by the hidden-state dimension. Therefore, reducing the hidden dimension from $H_{\text{teacher}} = 64$ to $H_{\text{student}} = 16$ yields a substantial reduction in model size.

Let $|W_{\text{teacher}}|$ and $|W_{\text{student}}|$ denote the payload sizes of the teacher and student models, respectively. The theoretical compression ratio can be approximated as

$$C_r \approx 1 - \left(\frac{H_{\text{student}}}{H_{\text{teacher}}} \right)^2 = 1 - \left(\frac{16}{64} \right)^2 = 93.75\%. \quad (13)$$

In the actual implementation, the payload is reduced from $|W_{\text{teacher}}| = 70.00$ KB to $|W_{\text{student}}| = 5.88$ KB. The gap between the theoretical approximation (93.75%) and the measured value is mainly due to input projection weights and bias terms, which scale linearly with H rather than quadratically. Accordingly, the downlink payload reduction ratio is given by

$$\eta_{\text{gain}} = \frac{|W_{\text{teacher}}| - |W_{\text{student}}|}{|W_{\text{teacher}}|} \approx 91.60\%. \quad (14)$$

These results indicate that the proposed FL-KD framework can substantially reduce the communication burden of model dissemination in V2X environments, making lightweight edge inference more practical for resource-constrained OBUs.

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

To evaluate the effectiveness of the proposed communication-efficient FL-KD framework, extensive simulations were conducted using real-world dynamic parking data from the Transport Data eXchange (TDX) [12] platform. We compare the performance of the baseline full-parameter model (Teacher) and the lightweight student model (Student) from three perspectives: communication payload, transmission latency, and prediction accuracy.

A. Communication Payload Reduction

A major challenge in V2X-based federated learning is the large size of neural network parameters that need to be delivered to OBUs. Table I summarizes the communication payload required for a single downlink transmission.

The teacher model with a hidden dimension of 64 requires a payload of 70.00 KB. In contrast, the distilled student model with a hidden dimension of 16 reduces the payload to 5.88 KB. This corresponds to a 91.60% reduction in downlink communication overhead, indicating that the proposed FL-KD framework can substantially reduce the burden of model dissemination in bandwidth-constrained vehicular networks.

B. Transmission Latency in V2X Networks

To quantify the latency reduction under constrained communication conditions, we consider a weak V2X signal environment with a limited bandwidth of 256 Kbps. The corresponding transmission latency is also reported in Table I.

In the baseline setting, downloading the complete teacher model requires 2187 ms. By contrast, the lightweight student model only requires 183 ms. This large reduction in transmission time improves the practicality of distributing model updates to fast-moving vehicles and supports more responsive parking recommendation services in dynamic V2X scenarios.

C. Prediction Accuracy and Model Generalization

A common concern in model compression is whether reducing the model size leads to degraded prediction performance. To provide an intuitive comparison, Fig. 2 and Fig. 3 present the prediction results for two representative parking lots, where the outputs of the teacher and student models are compared against the ground truth vacancy rate.

The results show that the student model maintains competitive prediction fidelity despite using substantially fewer parameters. Across the representative cases shown in Fig. 2 and Fig. 3, the student predictions closely follow both the teacher outputs and the ground truth data. This suggests that the proposed student model preserves the main predictive characteristics of the teacher model while achieving a much smaller model size.

TABLE I
PERFORMANCE COMPARISON UNDER CONSTRAINED V2X NETWORKS
(256 Kbps)

Model	Hidden Dim.	Data Payload (KB)	Transmission Latency (ms)
Teacher (Baseline)	64	70.00	2187
Student (Proposed)	16	5.88	183
Reduction	–	91.60%	~91.6%

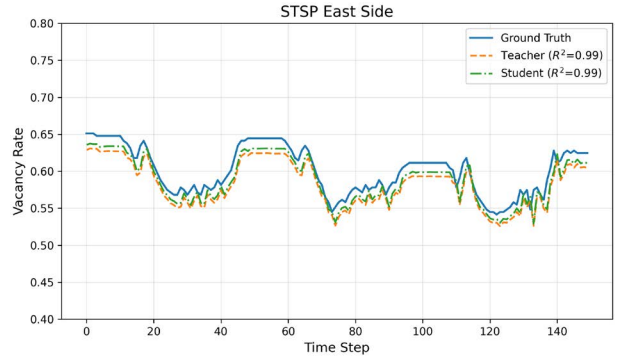


Fig. 2. Prediction performance of the Teacher model ($H = 64$) and Student model ($H = 16$) compared to the ground truth parking vacancy rate at STSP East Side.

From a quantitative perspective, the student model achieves an R-squared (R^2) comparable to that of the teacher model in the representative cases shown in Fig. 2 and Fig. 3, indicating that the reduction in model size does not lead to a substantial degradation in prediction quality. This behavior may be attributed to the regularization effect of teacher-guided supervision, which helps the student model learn smoother predictive behavior and improves generalization on edge devices.

Overall, the experimental results indicate that the proposed FL-KD framework can effectively improve communication efficiency while maintaining satisfactory predictive performance for resource-constrained IoV applications. In terms of the constraints formulated in Section II, the achieved compression ratio $|W_S|/|W_T| = 5.88/70.00 \approx 0.084$ is well below any practical $\rho_{\max}$, and the student model attains $R^2 \approx 0.99$, confirming that the predictive fidelity requirement R_{req}^2 is satisfied.

D. Dynamic Recommendation and Preference Trade-off

To evaluate the robustness and adaptability of the proposed edge inference architecture in realistic urban scenarios, we conducted a Monte Carlo simulation using static parking infrastructure data and dynamic time-series occupancy records from the TDX platform. Instead of relying on a single fixed vehicle location, we randomly generated the coordinates of 100 vehicles within the bounding box of the target city area to simulate diverse parking requests.

At the user application layer, the OBU leverages the downloaded lightweight student model to predict the vacancy rates

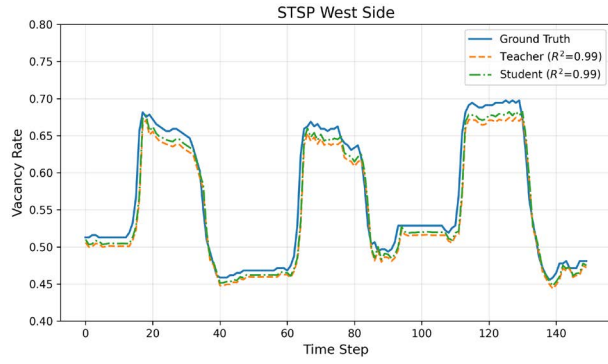


Fig. 3. Prediction performance of the Teacher model ($H = 64$) and Student model ($H = 16$) compared to the ground truth parking vacancy rate at STSP West Side.

($\hat{v}_j$) of nearby parking lots. For each candidate lot j , the geographic distance d_j is computed via the Haversine formula, and the estimated waiting time is derived as $w_j = (1 - \hat{v}_j) \times \tau$, where τ is a scaling constant. Each factor is min-max normalized across all candidates to obtain $\bar{v}_j, \bar{d}_j, \bar{w}_j \in [0, 1]$. A multi-criteria scoring function is then used to compute the recommendation score:

$$S_j = w_1 \cdot \bar{v}_j + w_2 \cdot (1 - \bar{d}_j) + w_3 \cdot (1 - \bar{w}_j), \quad (15)$$

where $w_1, w_2, w_3 \geq 0$ are user-defined preference weights satisfying $w_1 + w_2 + w_3 = 1$: w_1 controls the preference for parking availability, w_2 for proximity, and w_3 for shorter waiting time.

To demonstrate the effectiveness of this localized recommendation mechanism, we simulated three representative user preference profiles as summarized in Table II. The search radius was set to 5.0 km.

The results reveal a clear trade-off governed by the recommendation strategy. Under the distance-oriented preference ($w_1 = 0.1, w_2 = 0.8, w_3 = 0.1$), the system reduces the average travel distance to 2.42 km while maintaining a vacancy rate of 52.1%. Under the vacancy-oriented preference ($w_1 = 0.8, w_2 = 0.1, w_3 = 0.1$), the average vacancy level increases to 75.1%, at the expense of a longer average travel distance of 3.59 km. The balanced setting ($w_1 = 0.4, w_2 = 0.3, w_3 = 0.3$) achieves a practical trade-off with 72.1% vacancy and 3.05 km distance.

These findings show that combining the lightweight student model with a localized scoring function enables resource-constrained OBUs to perform personalized parking recommendation with low latency, without continuously offloading user-side query information to the central server.

VI. CONCLUSION

This paper proposed a communication-efficient collaborative learning framework that combines Federated Learning (FL) with Knowledge Distillation (KD) for smart parking in IoV environments. RSUs collaboratively train a teacher model via FL, and the server distills it into a lightweight

TABLE II
MONTE CARLO SIMULATION RESULTS

Preference	w_1	w_2	w_3	Avg. Dist. (km)	Avg. Vac. (%)
Distance-oriented	0.1	0.8	0.1	2.42	52.1%
Balanced	0.4	0.3	0.3	3.05	72.1%
Vacancy-oriented	0.8	0.1	0.1	3.59	75.1%

student model for deployment on OBUs. Experiments on real-world TDX data confirm that the student model achieves over 91% downlink payload reduction with negligible loss in prediction accuracy ($R^2 \approx 0.99$), while a localized multi-criteria recommendation module accommodates diverse user preferences without cloud offloading.

ACKNOWLEDGMENT

This study was supported by the National Science and Technology Council (NSTC), Taiwan, under Grant Nos. NSTC 114-2221-E-194-062- and NSTC 115-2221-E-194-042-MY2.

REFERENCES

- [1] A. Mishra and S. Deshpande, "Deep learning based parking prediction using lstm approach," in *Information and Communication Technology for Intelligent Systems (ICTIS)*, Ahmedabad, India, 2021.
- [2] W. Shao, Y. Zhang, B. Guo, K. Qin, J. Chan, and F. D. Salim, "Parking availability prediction with long short term memory model," in *Green, Pervasive, and Cloud Computing*, Uberlândia, Brazil, 2019.
- [3] Z. Xu, X. Tang, C. Ma, and R. Zhang, "Research on parking space detection and prediction model based on cnn-lstm," *IEEE Access*, vol. 12, pp. 30 085–30 100, 2024.
- [4] F. Zhang, K. Shang, L. Yan, H. Nan, and Z. Miao, "Prediction of parking space availability using improved mat-lstm network," *ISPRS International Journal of Geo-Information*, vol. 13, no. 5, 2024. [Online]. Available: <https://www.mdpi.com/2220-9964/13/5/151>
- [5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2023. [Online]. Available: <https://arxiv.org/abs/1602.05629>
- [6] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.
- [7] D. Maroua, "A state-of-the-art on federated learning for vehicular communications," *Vehicular Communications*, vol. 45, p. 100709, 2024.
- [8] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Zak, and Z. Wang, "Federated learning for connected and automated vehicles: A survey of existing approaches and challenges," 2023. [Online]. Available: <https://arxiv.org/abs/2308.10407>
- [9] S. B. Prathiba, G. Raja, S. Anbalagan, K. Dev, S. Gurumoorthy, and A. P. Sankaran, "Federated learning empowered computation offloading and resource management in 6g-v2x," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 5, pp. 3234–3243, 2022.
- [10] R. Song, L. Lyu, W. Jiang, A. Festag, and A. Knoll, "V2x-boosted federated learning for cooperative intelligent transportation systems with contextual client selection," 2023. [Online]. Available: <https://arxiv.org/abs/2305.11654>
- [11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: <https://arxiv.org/abs/1503.02531>
- [12] Transport data exchange (TDX). [Online]. Available: <https://tdx.transportdata.tw/>
- [13] V. K. Sarker, T. N. Gia, I. Ben Dhaou, and T. Westerlund, "Smart parking system with dynamic pricing, edge-cloud computing and lora," *Sensors*, vol. 20, no. 17, 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/17/4669>