ML-Based Delay Violation Probability-Aware Proactive Scheduling for Industrial 5G Networks

Ruofan Xu
Chair of Communication Networks
Technical University of Munich
Munich, Germany
ge83cub@tum.de

Shubhangi Bhadauria
Siemens AG
Munich, Germany
shubhangi.bhadauria@siemens.com

Wolfgang Kellerer
Chair of Communication Networks
Technical University of Munich
Munich, Germany
wolfgang.kellerer@tum.de

Fidan Mehmeti
Chair of Communication Networks
Technical University of Munich
Munich, Germany
fidan.mehmeti@tum.de

***Abstract*—Ultra-Reliable Low-Latency Communication (URLLC) is essential to realize the vision of Industry 4.0, where autonomous systems coordinate complex cyber-physical processes in real time. However, for such industrial applications, average latency minimization is insufficient, since rare delay spikes can violate strict delay bounds and compromise system stability or safety. This work proposes a machine learning (ML)-based delay violation probability (DVP)-aware proactive scheduling framework to address these rare delay spikes. The proposed framework adapts a transformer-based probabilistic delay prediction model to estimate packet-level delay-risk information and integrates the predicted risk into the 5G downlink scheduler. The considered system includes multiple user equipment (UE) devices assigned to heterogeneous quality of service (QoS) classes with different packet delay budget (PDB) and target DVP requirements. To support detailed evaluation, a cross-layer analysis framework is implemented for ns-3/5G-LENA that tracks packet transmission, reception, scheduling decisions, and DVP-related statistics across protocol layers. Simulation results in an 8-UE scenario show that the proposed scheduler substantially reduces delay violations for the delay-critical UEs with a $5\,$ms PDB and a target DVP of 10^{-4}, while preserving stable service for the remaining UEs, thereby increasing the total number of UEs satisfying their target DVP requirements.**

***Index Terms*—5G Industrial Networks, Proactive Scheduling, Delay Violation Probability, Quality of Service, Machine Learning.**

I. Introduction

Industry 4.0 has transformed cyber-physical systems (CPS) into Cyber Physical Production Systems (CPPS), with smart factories emerging as a key initiative [1]. Realizing this vision requires ultra-reliable, low-latency, and deterministic communication services, which Third Generation Partnership Project (3GPP) fifth generation (5G) networks support through ultra-reliable low-latency communications (URLLC) [2]. The corresponding quality of service (QoS) requirements, including end-to-end latency (1–50 ms) and reliability (up to 99.99999%), remain use-case specific [3]–[5]. As 3GPP sixth generation (6G) standardization progresses, Industry 5.0 introduces even more stringent requirements, with URLLC serving as a foundational service for critical industrial applications and real-time human-machine coordination; increasingly collaborative human-machine systems are expected to demand communication capabilities beyond current 5G standards [6]. Hence, URLLC remains a key enabler for this industrial revolution. In this work, we consider URLLC-oriented industrial communication under stringent latency-reliability requirements, with the most stringent evaluated configuration using a 5 ms PDB and a target DVP of 10^{-4}.

Critically, for URLLC-demanding industrial applications, only minimizing the average latency is insufficient [7]. Even if the mean end-to-end delay remains low, a small number of extreme delay spikes may violate a key QoS profile parameter—namely, the packet delay budget (PDB)—and trigger unstable or unsafe system behavior. Thus, the tail of the delay distribution becomes a critical performance aspect for such time-critical industrial applications. The challenge is modeling tail behavior accurately because tail events stem from unpredictable interactions: retransmissions, queueing delays, and processing delays. These factors interact nonlinearly, depend on prior system states, and are coupled across protocol layers.

Traditional reactive scheduling allocates resources according to the currently observed network state and can only respond after channel conditions change. Once congestion occurs, delay violations may become unavoidable—a limitation that motivates proactive scheduling, which incorporates predictions of future delay behavior into resource allocation to prevent violations before they occur [8]. Analytical approaches provide useful information, but often rely on idealized assumptions regarding traffic models, channel behavior, or queueing dynamics. This limitation reflects a broader challenge in 5G radio resource management (RRM): increasingly diverse deployment scenarios and sophisticated RRM functionalities make purely model-based approaches difficult to generalize across use cases [9].

Machine learning (ML) provides a promising alternative for proactive scheduling because it can learn temporal and cross-layer patterns directly from simulation or measurement data. Temporal sequence models, such as Transformer-based architectures [10], can learn how previous packet-level events and scheduling decisions affect future packet delay behavior [11]. However, predicting only the expected delay trajectory is insufficient, because delay violations depend on the tail behavior of the delay distribution. By combining temporal sequence models with mixture density networks (MDNs),

This work was supported by the European Commission's Horizon Europe Smart Networks and Services Joint Undertaking under the research and innovation project MULTI-X, Grant Agreement No. 101192521.

the full packet-level delay distribution can be estimated and the probability of exceeding a given PDB can be derived. Compared with mean delay, delay violation probability (DVP), defined as the probability that the end-to-end packet delay exceeds the corresponding PDB, directly captures latency-bound violations and is therefore more suitable for evaluating delay-critical services.

A key challenge emerges when attempting to deploy DVP-aware scheduling in real 5G New Radio (NR) industrial systems: while DVP-aware scheduling has been studied in wireless feedback-loop systems and URLLC cross-layer resource allocation [12], [13], existing approaches typically assume homogeneous QoS requirements, single-path channel models, or stationary user equipment (UE) deployments. However, in an indoor factory environment, UEs may support heterogeneous QoS classes, experience mobility, and encounter obstacles in the planned trajectory. Existing DVP-aware frameworks do not adequately address these challenges. In this paper, we propose an ML-based DVP-aware proactive scheduling mechanism for 5G NR industrial communication with stringent latency-reliability requirements. The proposed scheduler combines two key components: first, a probabilistic temporal sequence-based delay prediction framework is adapted to capture the entire delay distribution including tail behavior and second, a scheduler that prioritizes delay-critical UEs to help them satisfy their target DVP requirements while maintaining stable communication for non-critical UEs. By integrating prediction and scheduling jointly, the mechanism adapts to the dynamic channel and traffic conditions of indoor factory scenarios.

The main contributions of this paper can be summarized as:

- We propose an ML-based DVP-aware proactive scheduling mechanism for delay-critical 5G NR industrial communication. The scheduler aims to enable delay-critical UEs to satisfy their target DVP requirements while maintaining stable communication for the other UEs.
- We adapt a probabilistic sequence-based delay prediction framework to the considered downlink scheduling use case and integrate the adapted predictor into the scheduler.
- Lastly, a cross-layer evaluation framework in ns-3/5G-LENA is implemented to analyze how the scheduling decisions affect the DVP and end-to-end delay performance with mobility and obstacles in the trajectory of the automated guided vehicles (AGVs).

The remainder of this paper is organized as follows: Section II discusses related work. Section III presents the problem formulation. This is followed by Section IV, which introduces the ML-based delay distribution prediction model and its scheduler-oriented adaptations. Section V describes the proposed DVP-aware scheduling framework. In Section VI the performance of the approach is evaluated. Finally, Section VII concludes the paper.

II. Related Work

Over the years, many resource allocation approaches considering delay and reliability have been proposed for low-latency communication systems [14]. Yang *et al.* [15] investigated deep reinforcement learning (DRL)-based resource allocation in low-latency edge computing networks, where both DVP and decoding error probability are considered for end-to-end reliability. However, their focus is mainly on computational resource allocation in edge computing networks, rather than on packet-level radio scheduling in 5G NR downlink systems. Data-driven DVP prediction has also been explored for delay-tail control. Mostafavi *et al.* [16] proposed an active queue management framework using data-driven predictors to estimate packet-level delay violation probabilities. Different from their queue-management setting, this paper uses predicted packet-level DVP risk inside a 5G NR scheduler to adjust UE ordering and radio resource allocation. In the past few years, ML compared to traditional resource allocation procedures has gained significant traction. Compared with fourth-generation (4G) networks, 5G supports diverse applications including vehicular communications, device-to-device (D2D), machine-to-machine (M2M), mobile edge computing (MEC), cloud computing, and internet of things (IoT), while supporting enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and URLLC. This increased complexity motivates advanced radio resource management (RRM) strategies [17]. Pourmahboubi and Tabrizchi [18] proposed a long short-term memory (LSTM)-based framework for predicting 5G resource allocation percentages. Villegas *et al.* [19] proposed a DRL-based Medium Access Control (MAC) scheduler reconfiguration mechanism in Open Radio Access Network (O-RAN), using ns-3 [20]/5G-LENA [21] to improve QoS and resource efficiency.

Existing learning-based approaches mainly focus on resource allocation prediction, scheduler reconfiguration, queue management, or edge computing resource allocation. This work adopts the probabilistic delay prediction framework of Mostafavi [11], which predicts packet-delay distributions rather than point estimates [22], [23], enabling future delay distributions to be estimated. However, [11] considers uplink communication between a single stationary UE and a next-generation Node B (gNB) with multi-step prediction, whereas our downlink scheduler requires the delay-risk estimate of the immediate next packet. Their dataset is generated using OpenAirInterface (OAI) [24] and recorded by the end-to-end delay analytics framework (EDAF) [25]. The adaptations required for our scheduling scenario are described in Section IV.

Lastly, obstacle information for the AGVs is assumed to be available through integrated sensing and communication (ISAC). Prior work has investigated sensing-assisted scheduling in 5G/6G networks. For example, [26] uses a DRL agent to dynamically reconfigure scheduling based on ISAC and communication metrics. In this work, obstacle detection is assumed to be obtained through sensing measurements, while obstacle locations remain fixed during each simulation.

III. Problem Formulation

In this section, the system model is detailed first. This is followed by the problem statement.

A. System Model

We consider a 5G-enabled industrial scenario based on the 3GPP Indoor Factory (InF) channel model [3]. A single gNB provides downlink connectivity to multiple user equipment (UEs): automated guided vehicles (AGVs) and stationary robots. An industrial control system periodically generates downlink packets carrying control commands, navigation updates, obstacle or UE position, and task-related information. AGVs move towards corresponding stationary robot locations to execute transportation or handover tasks, while the robots are modeled as stationary service points. Consequently, the system comprises both mobile UEs with time-varying channel conditions and stationary UEs with relatively stable wireless links. UEs are assumed to support three QoS classes: *High*, *Medium*, and *Low*. Each class is associated with a priority level, a PDB, and a target DVP. A packet is considered delay-violating if its end-to-end delay exceeds the corresponding PDB, and the scheduling objective is to keep the DVP of each UE below its target DVP requirement.

In the considered industrial environment, we introduce environmental obstacles that affect wireless propagation. As AGVs move, varying distances to the gNB and non-line-of-sight conditions induce blockage, shadowing, multipath effects, signal-to-interference-plus-noise ratio (SINR) degradation, modulation and coding scheme (MCS) variations, retransmissions, and increased delay uncertainty.

In this work, we assume that obstacle locations and AGV trajectories are known *a priori*. This prior knowledge aligns with the sensing-assisted network architecture considered in this work, where environmental awareness – including the spatial layout of obstacles and AGV trajectories – is made available to network functions prior to communication. Accordingly, the AGVs follow predefined obstacle-aware trajectories derived from this environmental knowledge. Within such an architecture, obstacle and trajectory information is processed by dedicated sensing functions that produce situational awareness outputs. The focus of this work is not obstacle detection or path planning, but the impact of mobility, obstacle-induced channel variations, and heterogeneous QoS requirements on packet delay performance under different scheduling strategies.

For the analysis, the end-to-end delay of packet n of UE u is defined as

$$D_{u,n} = t^{\mathrm{rx}}_{u,n} - t^{\mathrm{tx}}_{u,n}, \tag{1}$$

where $t^{\mathrm{tx}}_{u,n}$ and $t^{\mathrm{rx}}_{u,n}$ denote the packet transmission time and reception time at the application layer, respectively. The DVP of UE u is defined as the probability that the packet delay exceeds its PDB:

$$\mathrm{DVP}_u = \Pr\left(D_{u,n} > \mathrm{PDB}_u\right). \tag{2}$$

B. Problem Statement

Rare delay spikes can violate PDB requirements, degrade system stability, and affect critical control operations in industrial applications. The objective of this work is to support stringent reliability for delay-sensitive QoS classes while ensuring packets meet both their PDB and target DVP requirements.

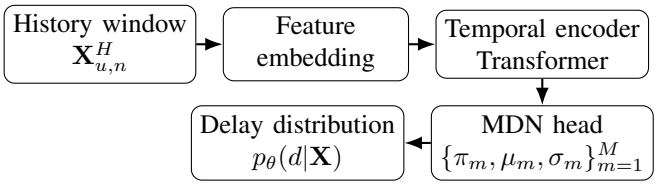


Fig. 1. Time-series probabilistic delay prediction pipeline.

In this work, the time-division multiple access QoS (TDMA-QoS) scheduler is assumed as the QoS-aware baseline scheduler. The TDMA-QoS scheduling decision depends on QoS priority, potential throughput, historical average throughput, and delay-budget-related urgency, rather than on the QoS class alone [27]. As a result, a delay-critical UE close to violating its PDB may not always receive the highest scheduling priority. In addition, TDMA-QoS allocates Orthogonal Frequency-Division Multiplexing (OFDM) symbols incrementally according to the scheduling order, assigning one OFDM symbol to one selected UE at each allocation step. Therefore, even if a delay-critical UE obtains a high scheduling weight, it may still receive insufficient resources to complete transmission within the remaining delay budget. Hence, these limitations can cause a UE with stringent PDB constraints to experience high DVP, especially under mobility-induced channel variations.

The problem is further exacerbated for very stringent PDB requirements (e.g., 5 ms). For such tight PDBs, a single retransmission triggered by low SINR or unstable channel conditions during the AGV movement can consume a large fraction of the remaining PDB, inherently causing delay violation.

These observations motivate a DVP-aware scheduling design that combines predicted packet violation risk with QoS-aware scheduling decisions. At the system level, the optimization goal is to maximize the number of supported UEs that meet their target DVP requirements. Consequently, the optimization problem in this work is formulated as

$$\begin{aligned} \max_{\pi,\mathcal{U}} \quad & |\mathcal{U}| \\ \text{s.t.} \quad & \mathrm{DVP}_u(\pi) \le \mathrm{DVP}_u^{\mathrm{target}}, \quad \forall u \in \mathcal{U}, \end{aligned} \tag{3}$$

where $|\mathcal{U}|$ is the number of supported UEs, $\mathrm{DVP}_u(\pi)$ is the DVP of UE u under scheduling policy π, and $\mathrm{DVP}_u^{\mathrm{target}}$ is its target DVP requirement.

IV. ML-Based Delay Distribution Prediction

This section discusses the ML-based delay distribution prediction framework used for the DVP-aware scheduler. The first step toward enabling DVP-aware scheduling in this work is to obtain packet-level delay distribution predictions. The overall prediction pipeline, adapted from [11], is shown in Fig. 1. Given a history window of packet-level observations $\mathbf{X}^H_{u,n}$, the input features are embedded into fixed-length tokens and

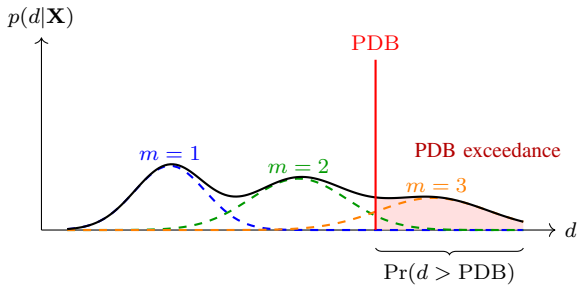


Fig. 2. Gaussian MDN output and the PDB exceedance probability.

processed by a Transformer-based temporal encoder. Through self-attention, the Transformer captures dependencies among historical packet events, such as MCS selection, retransmissions, resource allocation, and channel-related conditions. Instead of directly regressing a scalar delay, the resulting hidden representation is passed to an MDN head, from which the packet-level delay distribution can be obtained. MDNs were introduced in [28] as a way to combine neural networks with mixture models in order to represent conditional probability distributions. In this work, the MDN outputs the parameters of a Gaussian mixture model:

$$p_\theta\left(d \mid \mathbf{X}_{u,n}^H\right) = \sum_{m=1}^{M} \pi_{u,n,m} \mathcal{N}\left(d; \mu_{u,n,m}, \sigma_{u,n,m}^2\right), \quad (4)$$

where M is the number of mixture components. Each component is characterized by a mixture weight $\pi_{u,n,m}$, mean $\mu_{u,n,m}$, and standard deviation $\sigma_{u,n,m}$. A schematic example of such an MDN output is shown in Fig. 2. The dashed curves represent the individual Gaussian components, while the solid curve represents the resulting mixture distribution.

From this predicted distribution, the conditional mean delay can be derived as

$$\hat{d}_{u,n}^{\text{mean}} = \mathbb{E}_{p_\theta}\left[d \mid \mathbf{X}_{u,n}^H\right] = \sum_{m=1}^{M} \pi_{u,n,m} \mu_{u,n,m}. \quad (5)$$

Accordingly, the packet-level violation probability can be obtained as

$$p_{u,n}^{\text{vio}} = \Pr\left(D_{u,n} > \text{PDB}_u \mid \mathbf{X}_{u,n}^H\right). \quad (6)$$

For a Gaussian mixture distribution, this probability can be computed as

$$p_{u,n}^{\text{vio}} = \sum_{m=1}^{M} \pi_{u,n,m} \left[1 - \Phi\left(\frac{\text{PDB}_u - \mu_{u,n,m}}{\sigma_{u,n,m}}\right)\right], \quad (7)$$

where $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function. As illustrated in Fig. 2, this corresponds to the tail probability of the predicted delay distribution on the right-hand side of the PDB threshold. This probability is used as the packet-level ML risk signal for scheduling.

As mentioned in Section II, the authors in [11] focus on probabilistic delay forecasting for uplink communication with a single stationary UE. In contrast, this work considers a downlink industrial factory scenario with multiple UEs, heterogeneous QoS classes, mobility, and obstacle-related channel variations. Therefore, packet delay is affected by multi-UE resource competition, scheduling decisions, mobility-induced channel changes, and QoS-specific PDB and target DVP requirements, which are not considered in [11].

The proposed DVP-aware scheduler primarily requires delay-risk estimates for the immediate next packet. Therefore, the original model's [11] initialization strategy which assumes the decoder input at future step 0 with an all-zero start-of-sequence token, may introduce an information gap that weakens the immediate next-packet prediction, which is critical for online scheduling. To address this, we implement targeted adaptations to optimize the framework for one-step scheduler-oriented prediction. The prediction dataset is regenerated using ns-3/5G-LENA simulations for the considered downlink industrial scenario. Unlike the original single-UE measurement traces, the new dataset includes multiple UEs, heterogeneous QoS classes, QoS-specific PDBs, mobility, obstacle conditions, and scheduling-related information. This allows the model to learn the delay behavior caused by resource allocation, channel variation, and QoS heterogeneity.

The input feature set is extended from the original packet-level traffic and transmission features, such as packet size, inter-arrival time, packet arrival slot, MCS index, and retransmission counts. Additional scheduler- and channel-related features are introduced, including SINR, allocated physical resource blocks (PRBs), allocated OFDM symbols, and the distance between the gNB and the UE, so that the predictor can better capture the effects of channel quality, resource allocation, and mobility.

To enhance the first future step prediction, the decoder input is modified. Instead of using the zero start-of-sequence token, it is initialized with the embedding of the last historical packet. This strengthens the connection between the latest observed packet state and the immediate future prediction.

With these adaptations, the framework in [11] is adapted from a general probabilistic delay forecaster into a scheduler-oriented delay-risk estimator. The predicted delay distribution is used to estimate packet-level violation probability, which is then incorporated into the proposed DVP-aware scheduling decision.

V. Proposed DVP-Aware Scheduling Framework

As shown in Fig. 3, when packets from multiple UEs arrive at the MAC layer, the scheduler observes the active downlink queues and identifies the QoS class and PDB requirement of each UE. For each active UE, recent packet-level and cross-layer observations are used by the ML-based predictor to estimate future delay behavior and violation risk. The predicted risk is then tracked in a per-UE DVP state, which determines whether the UE should be regarded as delay-critical. This state is incorporated into the scheduling decision through QoS-DVP-aware UE ordering and resource protection. UEs without elevated risk follow the baseline TDMA-QoS

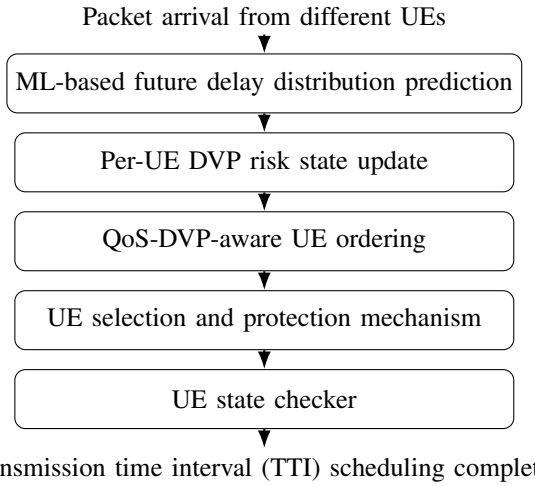


Fig. 3. Workflow of the proposed DVP-aware scheduling framework.

behavior, whereas risky UEs may receive higher scheduling priority and additional protection to reduce delay violations.

The following subsections describe the details of the proposed framework. The per-UE DVP risk state is introduced to explain how delay violation risk is estimated and tracked over time. The QoS-DVP-aware ordering and protection mechanism is then presented to show how the risk information is incorporated into the scheduling decision.

A. Per-UE DVP Risk State

To incorporate the predicted delay violation risk into scheduling, the proposed scheduler maintains a DVP risk state for each UE. For UE u, the scheduler maintains two historical statistics: V_u, the number of delay violations observed so far, and N_u, the number of previously observed packets.

The current empirical DVP is computed as

$$\mathrm{DVP}_u^{\mathrm{cur}} = \frac{V_u}{N_u}. \tag{8}$$

The current DVP violation condition is then defined as

$$C_u^{\mathrm{cur}} = \mathbf{1}\left(\mathrm{DVP}_u^{\mathrm{cur}} > \mathrm{DVP}_u^{\mathrm{target}}\right), \tag{9}$$

where $\mathrm{DVP}_u^{\mathrm{target}}$ is the target DVP requirement of UE u.

Given the packet-level violation probability $p_{u,n}^{\mathrm{vio}}$ predicted by the ML model, the scheduler further estimates the next-step DVP after packet n as

$$\mathrm{DVP}_u^{\mathrm{next}} = \frac{V_u + p_{u,n}^{\mathrm{vio}}}{N_u + 1}. \tag{10}$$

This value indicates how the DVP of UE u is expected to change if no additional protection is applied to the current packet.

The corresponding next-step risk ratio is defined as

$$R_u^{\mathrm{next}} = \frac{\mathrm{DVP}_u^{\mathrm{next}}}{\mathrm{DVP}_u^{\mathrm{target}}}. \tag{11}$$

A larger R_u^{next} indicates the DVP risk is higher relative to the target level.

B. QoS-DVP-Aware Ordering

The proposed scheduler introduces a QoS-DVP-aware ordering rule to address the limitation that the baseline TDMA-QoS weight does not always prioritize delay-critical UEs. At each scheduling opportunity, active UEs are ordered according to the following criteria: (1) QoS class priority, (2) current empirical DVP condition C_u^{cur}, (3) predicted next-step risk ratio R_u^{next}, and (4) baseline TDMA-QoS scheduling weight.

With this ordering rule, UEs with stricter QoS requirements are prioritized. Among UEs belonging to the same QoS class, those with higher current DVP pressure or higher predicted next-step delay risk are scheduled earlier. The baseline TDMA-QoS weight is retained as the final ordering criterion, so the proposed scheduler preserves the original QoS-aware scheduling behavior when the DVP risk levels are similar.

C. Protection Mechanism

After the QoS-DVP-aware ordering, the scheduler identifies whether a UE should enter protection mode. Here, we consider two activation conditions based on the predicted mean delay and the predicted violation probability. The first condition is triggered when the predicted mean delay of the next packet exceeds the PDB of UE u, indicating that the packet is likely to violate its latency requirement. The second condition is triggered when the predicted probability of exceeding the PDB is sufficiently high. This probability-based condition captures tail-risk cases in which the mean delay may still remain below the PDB, while the probability mass beyond the PDB is non-negligible. As the empirical DVP may be unreliable when only a small number of packets have been observed, a two-stage risk identification mechanism is used.

We define $N_{\min}$ based on the target DVP requirement; a more stringent target DVP requires a larger $N_{\min}$ to obtain a sufficiently stable UE-level DVP estimate. When $N_u < N_{\min}$, the scheduler relies on the ML-predicted violation probability. The current packet is identified as risky if

$$p_{u,n}^{\mathrm{vio}} > \tau_{\mathrm{early}}, \tag{12}$$

where τ_{early} is an empirically selected early-stage risk threshold, defined as

$$\tau_{\mathrm{early}} = \max\left(c \cdot \mathrm{DVP}_u^{\mathrm{target}}, 10^{-3}\right), \tag{13}$$

where c controls the sensitivity of early-stage risk detection. The value of c is selected empirically through preliminary simulations covering different UE configurations and operating scenarios, with the aim of balancing timely protection activation against unnecessary protection.

When $N_u \geq N_{\min}$, the scheduler uses a history-aware DVP budget. To satisfy the target DVP, the projected next-step DVP in (10) should not exceed the target requirement:

$$\frac{V_u + p_{u,n}^{\mathrm{vio}}}{N_u + 1} \leq \mathrm{DVP}_u^{\mathrm{target}}. \tag{14}$$

From this constraint, the remaining violation budget can be expressed as

$$B_u = \mathrm{DVP}_u^{\mathrm{target}}(N_u + 1) - V_u. \tag{15}$$

The current packet is then identified as risky if its predicted violation probability exceeds the remaining budget:

$$p_{u,n}^{\mathrm{vio}} > B_u. \tag{16}$$

Accordingly, protection is activated if the predicted mean delay exceeds the PDB or if the DVP risk condition described above is satisfied. However, the proposed scheduler still follows the TDMA-based allocation model, where a scheduled UE occupies the full configured bandwidth [29]. Therefore, protection is mainly applied through additional OFDM-symbol allocation and MCS backoff.

For resource allocation, a UE in protection mode can receive additional OFDM symbols beyond the baseline TDMA-QoS allocation, subject to the maximum available symbols in the current slot. This increases the chance that delay-critical packets are transmitted immediately rather than waiting for later scheduling opportunities.

For link robustness, an MCS backoff can be applied to the protected UEs. By selecting a more conservative MCS, the scheduler reduces the probability of Hybrid Automatic Repeat Request (HARQ) retransmissions. This is especially useful under the stringent 5 ms PDB QoS requirement, where even one retransmission may lead to delay violation.

Since MCS backoff lowers spectral efficiency, it is combined with additional OFDM-symbol allocation. The additional symbols compensate for the reduced data rate, while the lower MCS improves transmission reliability. Therefore, the protection mechanism jointly addresses scheduling-related delay and retransmission-induced delay.

To avoid keeping a UE in protection mode indefinitely, a UE state checker is implemented. The checker monitors the latest DVP state and the predicted risk of each UE. Protection is maintained when the UE remains DVP-critical, and it is released when the recent risk state falls below the configured release condition. This prevents unnecessary resource consumption and reduces the risk of performance degradation of other UEs.

D. Cross-Layer Analysis Framework

Although ns-3/5G-LENA provides trace functions for different protocol layers, the default traces are not sufficient to directly track all packet-level information required for DVP analysis, such as the associated UE identifier, QoS class, PDB, scheduling decision, reception status, delay violation flag, and packet loss flag. Therefore, this work implements a cross-layer analysis framework to track each packet while traveling through the system.

The framework records information from multiple layers, including application-layer transmission and reception timestamps, MAC scheduling decisions, assigned OFDM symbols, allocated PRBs, selected MCS, HARQ feedback, packet reception status, and packet delay. These records make it possible to link the end-to-end delay of each packet with the corresponding scheduling behavior and wireless link conditions. Based on the collected traces, packet-level delay violations and empirical DVP are computed for each UE.

The main evaluation metric is the empirical DVP. A packet is counted as a violation if it is either not successfully received or received after its corresponding PDB. Therefore, the empirical DVP is computed as

$$\mathrm{DVP} = \frac{N_{\mathrm{lost}} + N_{\mathrm{late}}}{N_{\mathrm{tx}}}, \tag{17}$$

where N_{tx} is the number of transmitted packets, N_{lost} is the number of packets that are not successfully received, and N_{late} is the number of received packets whose delay exceeds the corresponding PDB. This framework provides the basis for the system-level performance analysis in the next section.

VI. Performance Evaluation

In this section, the proposed ML-assisted DVP-aware scheduling framework is evaluated for both delay distribution prediction and scheduling. The prediction model is trained offline using packet-level traces collected from ns-3/5G-LENA simulations, exported to the Open Neural Network Exchange (ONNX) format [30], and integrated into the scheduler in ns-3 through ONNX Runtime [31] for online delay-risk inference. The resulting DVP-aware scheduler is evaluated under different traffic loads, UE densities, and mobility conditions, and is compared with the TDMA-QoS baseline in terms of DVP and the number of UEs satisfying the target DVP. Model training and simulations are performed on an NVIDIA DGX Spark workstation.

A. ML Prediction Performance

The ML model is first evaluated independently from the scheduler before being integrated into the scheduling decision. This evaluation is necessary because the proposed scheduler relies on the model output to estimate whether a packet is likely to violate its PDB. Therefore, the predictor should capture both the expected packet delay and the probabilistic delay behavior required for delay-risk-aware scheduling.

The evaluation is conducted on a dataset, which is generated from ns-3/5G-LENA simulations under the stringent QoS constraints. The dataset contains eight UEs, including one *High*-priority UE, one *Medium*-priority UE, and six *Low*-priority UEs. The dataset covers four scenario configurations that differ in mobility and obstacle settings. For each configuration, simulations are performed with packet generation intervals of 10, 20, 50, and 100 ms, resulting in 16 simulation settings in total. In each simulation, 1000 packets are transmitted to each UE, resulting in 8000 packet-level records per simulation. In total, the raw dataset contains $128,000$ packet-level records collected from 128 UE streams. Each stream corresponds to one radio network temporary identifier (RNTI) and contains packet-level traffic, scheduling, and delay-related information extracted from the simulator. The derived prediction samples are split into training, validation, and test sets with a ratio of approximately $0.8 : 0.1 : 0.1$.

Inspired by [11], the model observes a history window H of 20 packets. For the multi-step setting, a future window F of 20 is used, because the Transformer-based model achieved

TABLE I
FIRST-FUTURE-PACKET PREDICTION PERFORMANCE

Model	Accuracy	MAE [ms]	Std. NLL
Reference model (H20/F20)	0.736	1.3316	0.4939
Proposed model (H20/F20)	**0.944**	**0.2419**	**-0.4717**
Proposed model (H20/F1)	**0.982**	**0.1751**	**-0.5346**

its best performance with this window size [11]. Additionally, this work evaluates a scheduler-oriented one-step prediction, where the model predicts only the first future packet. This approach aligns well with the online scheduling requirement, because the scheduler must make scheduling decisions in real-time and only needs the delay-risk estimate for the next packet.

The prediction performance is evaluated using three metrics. The mean-hit accuracy measures the fraction of samples whose predicted mean delay is within a tolerance of $\Delta = 1$ ms from the true delay:

$$\mathrm{Acc}_{\Delta} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left(\left|\hat{d}_n^{\mathrm{mean}} - d_n\right| \le \Delta\right), \tag{18}$$

where N is the number of test samples, d_n is the true packet delay, and $\hat{d}_n^{\mathrm{mean}}$ is the predicted mean delay. The mean absolute error (MAE) is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\left|\hat{d}_n^{\mathrm{mean}} - d_n\right|. \tag{19}$$

While these two metrics evaluate the point-level delay prediction, the negative log-likelihood (NLL) evaluates the quality of the predicted delay distribution:

$$\mathrm{NLL} = -\frac{1}{N}\sum_{n=1}^{N}\log p_\theta\left(d_n \mid \mathbf{X}_n^H\right). \tag{20}$$

A lower NLL indicates that the predicted distribution assigns higher likelihood to the observed delay values, and therefore represents the probabilistic delay behavior more accurately.

Table I reports the prediction performance at the first future prediction step, which is the most relevant one for scheduling, because it corresponds to the immediate future packet used by the scheduler.

Under the same H20/F20 setting, the proposed model outperforms the reference model across all metrics, increasing the mean-hit accuracy from 0.736 to 0.944 and reducing the MAE from 1.3316 ms to 0.2419 ms. The standardized NLL also decreases from 0.4939 to -0.4717, indicating a better predicted delay distribution. The H20/F1 model further improves the first-future-packet prediction performance, achieving a mean-hit accuracy of 0.982, an MAE of 0.1751 ms, and a standardized NLL of -0.5346. Therefore, the H20/F1 model is used as the scheduler-oriented delay-risk predictor in the system-level evaluation.

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Scenario	3GPP Indoor factory
Factory size	$100 \times 100 \times 10$ m^3
Carrier frequency	3.5 GHz
System bandwidth	20 MHz
Numerology	0
Subcarrier spacing	15 kHz
Slot duration	1 ms
OFDM symbols per slot	14
PRBs per OFDM symbol	106
gNB transmission power	30 dBm
AGV moving speed	3 km/h
Obstacle attenuation	3.0 dB
Packet generation interval	20 ms
Packet size	20 B
Transmitted Packets	10^6

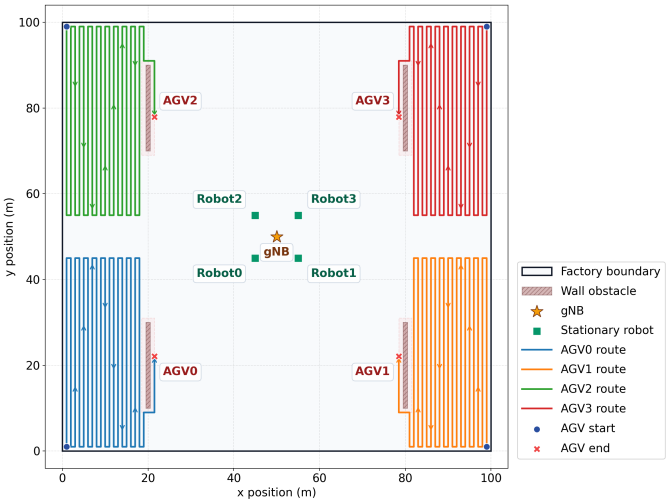


Fig. 4. Indoor factory scenario.

B. DVP-Aware Scheduler Performance

The system-level evaluation is performed using ns-3/5G-LENA. We consider mobile and stationary UEs in an indoor factory scenario, where AGVs move towards their corresponding stationary robots while avoiding obstacles. An example 8-UE deployment is shown in Fig. 4. The main simulation parameters are summarized in Table II. We evaluate the scheduler performance in three cases, as summarized in Table III.

The target DVP requirements are 10^{-4}, 10^{-3}, and 10^{-2} for *High*-, *Medium*-, and *Low*-priority traffic, respectively. Cases A and B are evaluated with 4, 6, and 8 UEs, including one *High*-priority AGV and one *Medium*-priority AGV, whereas Case C considers an 8-UE setting with two *High*-priority AGVs, two *Medium*-priority AGVs, and four *Low*-priority UEs.

Figs. 5–7 report the DVP results for the evaluated cases. The detailed plots focus on the delay-critical traffic classes, where delay violations are most likely to occur.

The results show that the proposed scheduler consistently reduces the DVP compared with the baseline scheduler across the evaluated PDB configurations. In Case A, the *Medium*- and

TABLE III
EVALUATION CASES FOR THE MIXED MOBILE AND STATIONARY UE SCENARIOS

Case	*High*	*Medium*	*Low*	*High* UE	*Medium* UE
A	5 ms	20 ms	100 ms	1	1
B	5 ms	8 ms	10 ms	1	1
C	5 ms	8 ms	10 ms	2	2

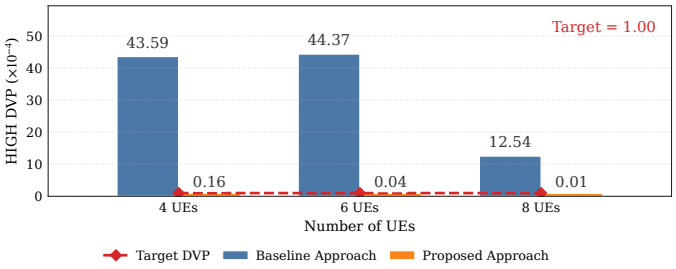
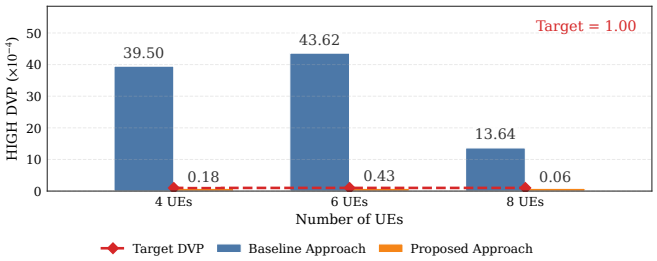


Fig. 5. Case A: *High*-priority DVP comparison under the stringent PDB setting, i.e., *High* PDB = 5 ms.

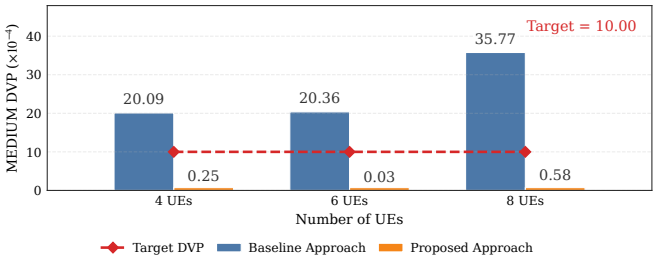
Low-priority traffic have relatively relaxed PDB requirements, so the baseline scheduler can already satisfy their target DVP requirements. The main improvement in this case is observed for the *High*-priority UE with the stringent 5 ms PDB requirement. In Case B, the improvement is observed under a stringent multi-class PDB setting, where both *High*- and *Medium*-priority traffic have strict delay budgets. In Case C, the proposed scheduler remains effective when the number of delay-critical mobile UEs is increased. Overall, the proposed scheduler reduces the DVP and enables more UEs to satisfy their target DVP requirements by combining DVP-aware ordering, additional OFDM-symbol allocation, and MCS back-off.

Table IV reports the number of UEs whose measured DVP does not exceed the corresponding target DVP, where BA denotes the baseline TDMA-QoS scheduler and PA denotes the proposed approach. This metric provides a system-level view of the improvement, as it reflects how many UEs can meet their individual target DVP requirements. As shown in Table IV, the proposed scheduler increases the number of UEs satisfying their target DVP in all evaluated configurations. In Cases A and B, all evaluated UEs meet their target DVP under the proposed scheduler for 4-, 6-, and 8-UE scenarios. In Case C, where more mobile UEs are assigned to delay-critical classes, the proposed scheduler increases the number of UEs meeting the target DVP from 4 to 8. This confirms that the proposed scheduler not only improves the DVP performance of delay-critical UEs, but also preserves stable communication for the remaining UEs, thereby improving overall system-level delay-reliability satisfaction.

The experimental results also show that the baseline TDMA-QoS scheduler can already satisfy the target DVP requirements of UEs with less stringent PDB and DVP constraints in the considered mobility scenarios. Simulations considering only stationary UEs show negligible delay violations even when

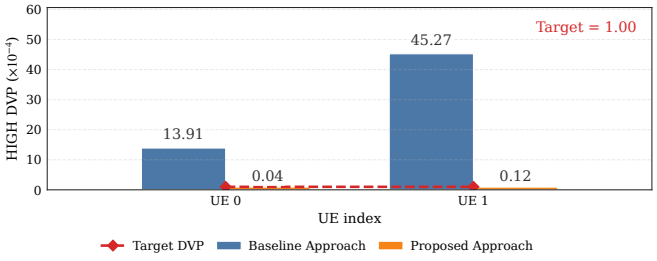


(a) *High*-priority traffic, PDB = 5 ms.

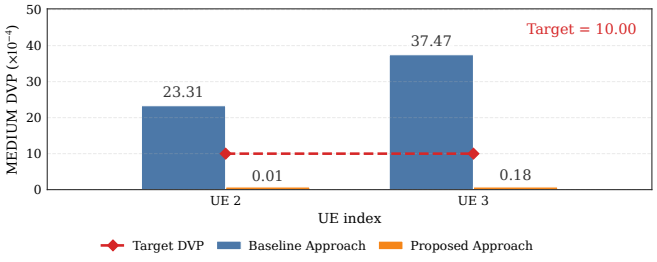


(b) *Medium*-priority traffic, PDB = 8 ms.

Fig. 6. Case B: DVP comparison under the highly stringent PDB setting.



(a) *High*-priority traffic, PDB = 5 ms.



(b) *Medium*-priority traffic, PDB = 8 ms.

Fig. 7. Case C: DVP comparison under the highly stringent multi-critical-UE configuration.

the number of UEs reached 16 within the factory layout. Hence, the reliability bottleneck is not solely determined by UE density, but mainly arises from the interaction between mobility-induced channel variations and scheduling decisions. Moreover, the number of UEs does not have a strictly monotonic impact on delay violations, because changes in the active UE set can affect the TDMA-QoS scheduling order through QoS weights, proportional-fair throughput, and delay-budget-related factors. The proposed scheduler improves its robustness by dynamically adjusting both the transmission priority and resource allocation based on each UE's predicted

TABLE IV
NUMBER OF UES MEETING THEIR TARGET DVP UNDER DIFFERENT PDB CONFIGURATIONS

Case	4 UEs		6 UEs		8 UEs	
	BA	**PA**	**BA**	**PA**	**BA**	**PA**
A	3	**4**	5	**6**	7	**8**
B	2	**4**	4	**6**	6	**8**
C	–	–	–	–	4	**8**

delay violation risk.

VII. CONCLUSION

In this paper, we designed and implemented an ML-based DVP-aware scheduler for delay-critical 5G NR industrial communication. The proposed scheduler integrates delay behavior prediction into TDMA-QoS scheduling and applies targeted protection through DVP-aware UE ordering, additional OFDM-symbol allocation, and MCS backoff when needed. Simulation results demonstrate that the proposed scheduler improves delay-reliability satisfaction in mobile industrial scenarios by reducing delay violations for delay-critical UEs and increasing the number of UEs achieving their target DVP requirements. Validation under more stringent URLLC regimes, including a 1 ms PDB and DVP targets such as 10^{-6}, remains for future work, since reliable empirical estimation at this level would require on the order of 10^8 transmitted packets and is constrained by the computational and storage costs of detailed ns-3 simulations and the available hardware resources. Future work will also validate the proposed approach on a real 5G testbed, including a comparison of the delay-trace characteristics generated by ns-3/5G-LENA with measured delay behavior, and investigate different access schemes and real-time trajectory adaptation through ISAC.

REFERENCES

[1] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0—Inception, conception and perception," *J. Manuf. Syst.*, vol. 61, 2021.

[2] International Telecommunication Union (ITU), *IMT Vision—Framework and overall objectives of the future development of IMT for 2020 and beyond*, ITU-R Recommendation M.2083-0, 2015.

[3] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP TR 38.901, V18.0.0, Release 18, 2024.

[4] 3GPP, "System architecture for the 5G System (5GS)," 3GPP TS 23.501, V16.6.0, Release 16, 2020.

[5] 3GPP, "Service requirements for cyber-physical control applications in vertical domains," 3GPP TS 22.104, V17.7.0, Release 17, 2022.

[6] A. M. Ibrahim, R. Nordin, Y. S. M. Khamayseh, A. Amphawan, and M. B. Jasser, "URLLC for 6G-enabled Industry 5.0: A taxonomy of architectures, cross-layer techniques, and time-critical applications," *ICT Express*, 2026.

[7] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk, and scale," *Proc. of IEEE*, no. 10, vol. 106, 2018.

[8] S. Hendaoui, F. Hendaoui, and N. Zangar, "Dynamic proactive–reactive scheduling for URLLC in 5G: Leveraging XGBoost and network virtualization," *Physical Communication*, vol. 68, 2025.

[9] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo, and P. Soldati, "Learning radio resource management in 5G networks: Framework, opportunities, and challenges," *IEEE Commun. Mag.*, no. 9, vol. 56, 2018.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NeurIPS*, 2017.

[11] S. Mostafavi, G. P. Sharma, A. Traboulsi, and J. Gross, "Probabilistic delay forecasting in 5G using recurrent and attention-based architectures," *arXiv preprint arXiv:2503.15297*, 2025.

[12] S. Zoppi, J. P. Champati, J. Gross, and W. Kellerer, "Scheduling of wireless edge networks for feedback-based interactive applications," *IEEE Trans. Commun.*, no. 5, vol. 70, 2022.

[13] Y. Wang and W. Chen, "Minimizing delay violation probability in URLLC over fading channels: A cross-layer approach," in *Proc. of IEEE GLOBECOM*, 2020.

[14] M. E. Haque, F. Tariq, M. R. A. Khandaker, K.-K. Wong, and Y. Zhang, "A survey of scheduling in 5G URLLC and outlook for emerging 6G systems," *IEEE Access*, vol. 11, 2023.

[15] T. Yang, Y. Hu, M. C. Gursoy, and A. Schmeink, "Deep reinforcement learning based resource allocation in low latency edge computing networks," in *Proc. of ISWCS*, 2018.

[16] S. Mostafavi, N. Roy, G. Dán, and J. Gross, "Active queue management with data-driven delay violation probability predictors," in *Proc. of IEEE GLOBECOM*, 2023.

[17] I. A. Bartsiokas, P. K. Gkonis, D. I. Kaklamani, and I. S. Venieris, "ML-based radio resource management in 5G and beyond networks: A survey," *IEEE Access*, vol. 10, 2022.

[18] A. Pourmahboubi and H. Tabrizchi, "LSTM-based framework for 5G resource allocation prediction," in *Proc. of IST*, 2024.

[19] N. Villegas, J. L. Herrera, L. Diez, D. Scotece, L. Foschini, and R. Agüero, "DRL-based dynamic MAC scheduler reconfiguration in O-RAN," in *Proc. of IEEE ICC*, 2025.

[20] G. F. Riley and T. R. Henderson, "The ns-3 network simulator," in *Modeling and Tools for Network Simulation*, Springer, 2010.

[21] N. Patriciello, S. Lagén, B. Bojović, and L. Giupponi, "An E2E simulator for 5G NR networks," *Simulation Modelling Practice and Theory*, vol. 96, 2019.

[22] D. C. Moreira, I. M. Guerreiro, W. Sun, C. C. Cavalcante, and D. A. Sousa, "QoS predictability in V2X communication with machine learning," in *Proc. of IEEE VTC-Spring*, 2020.

[23] S. Barmpounakis, L. Magoula, N. Koursioumpas, R. Khalili, J. M. Perdomo, and R. P. Manjunath, "LSTM-based QoS prediction for 5G-enabled connected and automated mobility applications," in *Proc. of IEEE 5GWF*, 2021.

[24] OpenAirInterface Software Alliance, "OpenAirInterface: Open-source software-based implementation of 5G/4G radio access network and core network," [Online]. Available: https://openairinterface.org/.

[25] S. Mostafavi, M. Tillner, G. P. Sharma, and J. Gross, "EDAF: An end-to-end delay analytics framework for 5G-and-beyond networks," in *Proc. of IEEE INFOCOM WKSHPS*, 2024.

[26] N. Villegas, J. L. Herrera, L. Diez, D. Scotece, L. Foschini, and R. Agüero, "ISAC-assisted DRL for dynamic MAC scheduler reconfiguration in O-RAN," *IEEE Open J. Commun. Soc.*, vol. 7, 2026.

[27] K. Koutlia, S. Lagén, and B. Bojović, "Enabling QoS provisioning support for delay-critical traffic and multi-flow handling in ns-3 5G-LENA," in *Proc. of WNS3*, 2023.

[28] C. M. Bishop, "Mixture density networks," Neural Computing Research Group, Aston University, Tech. Rep. NCRG/94/004, 1994.

[29] A. Larrañaga, M. C. Lucas-Estañ, S. Lagén, Z. Ali, I. Martinez, and J. Gozálvez, "An open-source implementation and validation of 5G NR configured grant for URLLC in ns-3 5G-LENA: A scheduling case study in Industry 4.0 scenarios," *J. Netw. Comput. Appl.*, vol. 215, 2023.

[30] Open Neural Network Exchange, "ONNX: Open Neural Network Exchange," [Online]. Available: https://onnx.ai/.

[31] Microsoft, "ONNX Runtime," [Online]. Available: https://onnxruntime.ai/.