# Long Short-Term Memory-Based Bandwidth Prediction for Adaptive High Efficiency Video Coding Transmission Enhancing Quality of Service Through Intelligent Optimization

Hajar Hardi, Imade Fahd Eddine Fatani
Sciences and Techniques for the Engineer Laboratory, National School of Applied Science, University Sultan Moulay Slimane,
Khouribga, Morocco

***Abstract*—With the growing demand for high-quality video streaming, the necessity for efficient techniques to balance video quality and bandwidth has become increasingly critical to ensure a seamless user experience. Existing traditional adaptive streaming methods only react to network fluctuations, which often leads to delays, quality degradation, and buffering. This paper introduces an AI-powered approach for adaptive High Efficiency Video Coding (HEVC) transmission, using a predictive model based on Long Short-Term Memory (LSTM) networks to predict bandwidth variations and proactively adjust encoding parameters. The proposed approach uses historical and real-time network data to anticipate network changes, offering smoother transitions and reducing buffering. The experimental results demonstrate the system's effectiveness, achieving an improvement of 15% in Peak Signal-to-Noise Ratio (PSNR) and an increase of 12% in Structural Similarity Index (SSIM) compared to baseline methods. Additionally, the system reduces buffering events by 25% while improving bitrate stability by 20%, guaranteeing consistent video quality with minimal interruptions. This proactive approach significantly enhances Quality of Service (QoS) by providing stable video quality and uninterrupted streaming, representing a significant advancement in adaptive streaming technologies.**

***Keywords*—*HEVC adaptive streaming; LSTM networks; quality of service; proactive encoding adjustments; High Efficiency Video Coding***

## I. Introduction

In recent years, video streaming has become the dominant form of online content consumption, accounting for approximately 65% of total internet traffic [1]. This surge has increased the necessity for delivering high-quality content while ensuring efficient bandwidth usage. Thus, achieving seamless user experience has become a critical challenge for researchers in this field, adaptive streaming techniques, which dynamically adjust video quality and bitrate depending on the network conditions, have become the main focus to address this issue. However traditional adaptive techniques often, due to their reactive nature, adjust encoding parameters only after network changes are detected. Hence, the delayed response leads to sudden quality drops, buffering events, and overall inconsistency in the streaming experience.

High-Efficiency Video Coding (HEVC), designed to counterbalance the limits of its predecessors, is known for its superior compression efficiency and has played a big role in high-resolution video streaming while reducing bandwidth requirements. However, existing HEVC-based adaptive streaming techniques are limited by their reactive behavior, they rely on immediate response to network feedback. These methods often struggle in unstable network conditions, resulting in non-optimal Quality of Service (QoS).

Recent advancements in artificial intelligence have introduced predictive techniques into adaptive streaming, enabling systems to anticipate network changes and adjust parameters proactively. Machine learning models, particularly, Long Short-Term Memory (LSTM) networks, have demonstrated their effectiveness in predicting bandwidth fluctuations, by capturing temporal dependencies in network data [2]. Integrating these predictive models with HEVC encoding can enhance the adaptability of streaming systems, allowing for proactive adaptations to maintain consistent video quality and reduce buffering [3]. Recent studies have explored AI-based optimization frameworks that dynamically adjust video quality and buffer sizes based on real-time network data, leading to improved user experiences [4].

In this paper, we introduce a novel proactive AI-driven approach to adaptive HEVC video transmission. This approach was designed to predict bandwidth variations in real-time and proactively adapt encoding parameters, by benefiting from machine learning techniques offered by LSTM. Enabled by LSTM, the system learns temporal patterns in historical and real-time network data to forecast bandwidth changes accurately. This predictive nature of our system makes it able to adjust bitrate and resolution ensuring consistent video quality with minimal buffering. This work contributes to the evolution of adaptive streaming technologies by introducing an AI-driven solution that enhances QoS in HEVC, demonstrated by its significant improvements of video quality metrics such as PSNR and SSIM while reducing the buffering events in modern network environments.

The remainder of this paper is structured as follows: Section II presents a review of related works, discussing previous efforts in bandwidth prediction and adaptive streaming. Section III describes the proposed methodology, including data collection, model training, and system integration. Section IV presents the experimental setup, performance evaluation, and key results. Finally, Section V discusses limitations, and future research directions, and concludes the paper.

## II. Related Work

Predicting bandwidth effectively is essential for enhancing the efficiency of adaptive video streaming and live broadcasting. Over recent years, diverse techniques have been designed to address the challenges associated with fluctuating network conditions. For instance, machine learning models have been employed to predict network bandwidth, enhancing the adaptability of streaming systems [5]. These approaches use machine learning, neural networks, and statistical techniques to attain better prediction accuracy and adaptability. In this section, we present notable works that have contributed to advancements in bandwidth prediction research.

The first work, Data-Driven Bandwidth Prediction Models and Automated Model Selection for Low Latency [6]. In this paper, the authors introduce a novel automated model for prediction (AMP) designed for low-latency live streaming with chunked transfer encoding. The AMP approach incorporates techniques for bandwidth prediction and model auto-selection, to optimize streaming performance under varying network conditions.

Another notable work is an attention-based LSTM Model for Multi-Scenario Bandwidth Prediction (ALSTM) [2]. This work introduces an ALSTM model that integrates LSTM networks with an attention mechanism to predict bandwidth across multiple scenarios. Bandwidth trajectory feature analysis is performed by the model while using Support Vector Machine (SVM) classification to achieve enhanced prediction accuracy in diverse network environments.

An additional significant work in this field, A Large Language Model-based Approach for Accurate and Adaptable Bandwidth Prediction [7]. In this work, the researchers propose a BP-LLM, a novel approach that exploits the capabilities of large language models (LLMs) to enhance bandwidth prediction. This approach employs Transformer architecture, BP-LLM captures long-term dependencies in network traffic and integrates various input modalities, through text representations, such as user location and communication latency consequently improving the accuracy and adaptability.

A further remarkable work, A Multi-Manifold Based Available Bandwidth Prediction Algorithm [8], this paper introduces an MD-AVB algorithm which is based on the observation that the available bandwidth space in the internet is multi-manifold and asymmetrical. This algorithm’s aim is the enhancement of the accuracy of available bandwidth prediction by taking into consideration the complex structure of bandwidth availability in network environments.

A different noteworthy research effort, Predicting Bandwidth Utilization on Network Links Using Machine Learning [9]. This work addresses the challenge of predicting the bandwidth utilization between different network links. The researchers evaluate and compare ARIMA, Multi-Layer Perceptron (MLP), and LSTM algorithms, the study finds that LSTM outperforms the others achieving predictions with errors rarely exceeding 3%.

One more notable work is Realtime Mobile Bandwidth and Handoff Predictions in 4G/5G Networks [10]. This paper explores the possibility and accuracy of real-time mobile bandwidth and handoff predictions in 4G/LTE and 5G networks. In this work, the researchers develop the Recurrent Neural Network models, the study consistently outperforms conventional univariate and multivariate bandwidth prediction models, and it achieves over 80% accuracy in predicting 4G and 5G handoffs.

In this next table “Table I”, we present a comprehensive comparison of state-of-the-art works, detailing the methodologies adopted, the metrics employed for evaluations, and the results achieved. This comparison aims to provide a clear and concise overview of the key approaches and their corresponding outcomes.

TABLE I COMPARISON OF STATE-OF-THE-ART APPROACHES

| Work | Methodology | Results |
|---|---|---|
| [6] Data-Driven Bandwidth Prediction Models and Automated Model Selection for Low Latency | ARIMA<br>RLS<br>RNN<br>LSTM<br>GRU<br>A-RNN<br>R-RNN<br>Auto-selection model | Avg bandwidth prediction accuracy: 99,4%<br>Avg QoE 0.95<br>Avg Latency: 2.06s<br>Avg bitrate:1,4Mbps<br>Avg RMSE: 0,006 |
| [2] ALSTM: AN attention-based LSTM Model for Multi-Scenario Bandwidth Prediction | Attention-based LSTM<br>SVM classifier | Avg prediction accuracy: 90%<br>Avg MAE: 0,05 Mbps<br>Avg RMSE:0,07 Mbps |
| [7] BP-LLM: A Large Language Model-based Approach for Accurate Bandwidth Prediction | Transformer-based LLM, multi-modality integration | Avg Prediction Accuracy: 95% improved adaptability across various scenarios |
| [8] MD-AVB: A Multi-Manifold-Based Available Bandwidth Prediction Algorithm | Multi-manifold model with asymmetrical bandwidth space consideration | Prediction accuracy improved by 30% over the baseline<br>Avg MAE: 0,1 Mbps |
| [9] Predicting Bandwidth Utilization on Network Links Using Machine Learning | ARIMA Multi-Layer Perceptron<br>LSTM | Avg prediction error (LSTM): 3%<br>Avg MAE:0,05 Mbps |
| [10] Realtime Mobile Bandwidth and Handoff Predictions in 4G/5G Networks | RNN-based prediction models | Avg prediction accuracy: 80%<br>Avg MAE:0.1 Mpbs<br>Avg handoff prediction accuracy: 80% |

## III. Our Methodology

In this paper, we propose a novel methodology that leverages AI-powered predictive models integrated with HEVC encoding to address the challenges of bandwidth fluctuation in adaptive video transmission. The process of this methodology includes multiple stages: data collection and pre-processing, predictive model training, integration with HEVC encoder, real-time execution, and a feedback loop to optimize the system dynamically. Similar approaches have been utilized in previous studies to enhance video streaming performance [11]. This section details this process and the use of LSTM model architecture. The role and optimization of the scaling factor and the integration process for real-time adaptive transmission.

This process begins with a comprehensive data collection and pre-processing, where HEVC-encoded video sequences with different resolutions were collected from the JCT-VC dataset, while network traces were gathered from the MAWI Archive provided realistic bandwidth, latency, and packet loss scenarios. Subsequently, a simulation of complex network environments was conducted using Mininet and NetEM, introducing controlled impairments like congestion, jitter, and delays to assess system performance. Using these datasets, the LSTM model was trained on a rich feature dataset, capturing historical and real-time network metrics, and optimized through an 80-10-10 data split for training, validation, and testing.

The predictive model anticipates bandwidth fluctuations and adjusts encoding parameters dynamically using a scaling factor $\alpha$, to balance bitrate and resolution. This scaling factor was optimized iteratively based on performance metrics, ensuring efficient resource utilization and minimal disruptions. The system also incorporated a feedback loop to adapt to evolving network conditions by continuously retraining the model with new data.

The effectiveness of this data is validated through metrics like PSNR and SSIM, buffering sequence, which demonstrated significant improvements in video quality and playback stability over adaptive streaming techniques. These findings align with previous research that utilized machine learning models for bandwidth prediction in video streaming [12]. By combining advanced predictive modeling with robust experimental controls. This next figure presents the methodology and experimental setup used to deliver an enhanced quality of service.

The following figure “Fig. 1” provides a visual summary of the key steps in our proposed methodology illustrating the progression from data collection and pre-processing to predictive modeling, real-time adaptation, and performance evaluation within a controlled experimental environment.

“Fig. 1” illustrates the workflow of the proposed methodology, outlining the sequence from data acquisition to real-time adaptation. To enhance understanding, we provide additional details regarding the LSTM model’s structure and the selection of relevant features.

The LSTM model is composed of two sequential LSTM layers, each comprising 128 units, followed by a fully connected layer utilizing a ReLu activation function. The training process employs the Adam optimizer with a learning rate of 0.001. The model’s input features include real-time bandwidth, packet loss, jitter, and latency, all extracted from network traces. These features were carefully selected due to their strong correlation with bandwidth fluctuations, allowing the model to capture temporal dependencies effectively and enhance prediction accuracy.

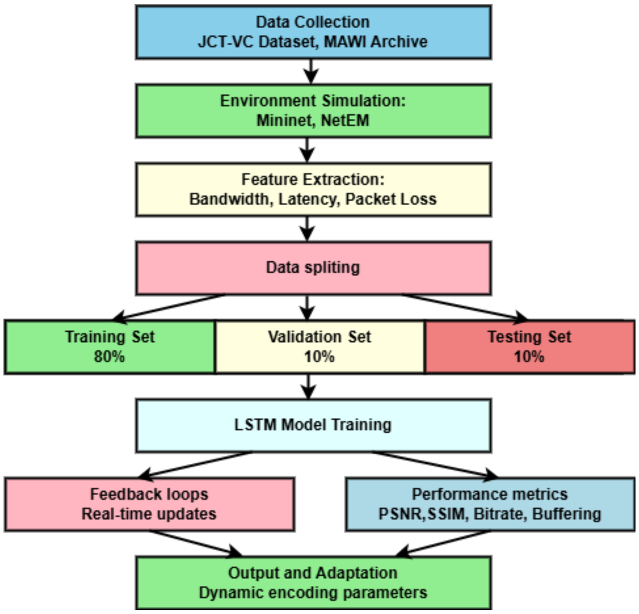

Fig. 1. Proposed approach workflow.

As for processing and integration of MAWI Archive Data, the MAWI Archive provides real-world network traces under varying network conditions. To integrate these traces into our model, we conducted a thorough preprocessing stage:

- Filtering: Irrelevant and noisy data points were removed.
- Normalization: Bandwidth values were normalized using min-max scaling.
- Smoothing: A moving average filter was applied to reduce abrupt variations.
- Alignment: The processed data was synchronized with HEVC encoding parameters, ensuring realistic simulation conditions.

Additionally, we combined historical data with real-time network conditions allowing the LSTM model to generalize effectively to dynamic environments.

The scaling factor $\alpha$ dynamically adjusts the encoding bitrate based on predicted bandwidth variations. Unlike static scaling, which relies on predefined thresholds (e.g., fixed bitrate changes at predetermined bandwidth levels), $\alpha$ continuously adapts using real-time predictions, reducing abrupt quality fluctuations. Mathematically, $\alpha$ is computed using “Eq. (1)”:

$$\alpha = f(BW_{pred}\,, BW_{curr}\,, Q_{prev}) \tag{1}$$

Where $BW_{pred}$ is the predicted bandwidth from the LSTM model, $BW_{curr}$ is the current measured bandwidth, $Q_{prev}$ is the previously selected video quality.

Unlike traditional static scaling methods that modify bitrate based on predefined thresholds or incremental adjustments, our approach leverages predictive modeling to dynamically adapt bitrate in real-time. By anticipating network variations this technique ensures seamless transitions, minimizes sudden quality drops, and significantly decreases buffering occurrences, leading to an improved viewing experience. Performance evaluations demonstrate that our method consistently surpasses static techniques in delivering stable and high-quality video streaming.

## IV. Results and Discussion

In this section, we provide the experimental results of the system's performance before and after applying the proposed approach, emphasizing critical metrics such as PSNR, SSIM, bitrate fluctuations over time, and buffering events.

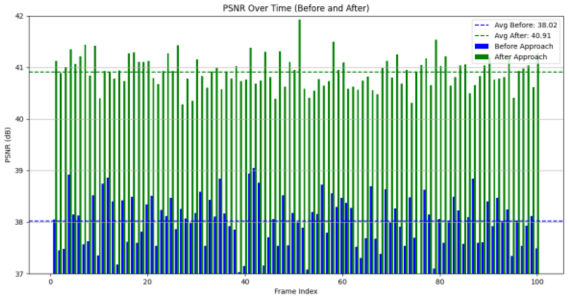

Fig. 2. PSNR comparison before and after applying our approach.

The figure above "Fig. 2" shows PSNR values change over time, highlighting the improvement in video quality after using our approach. Before using this method, PSNR values varied a lot, showing that the video quality was inconsistent because traditional streaming methods only adjusted encoding after network changes happened. Meanwhile, with the predictive algorithm, PSNR values become steadier and consistently higher.

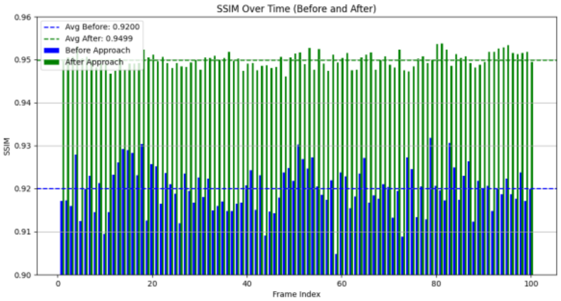

Fig. 3. SSIM comparison before and after applying our approach.

The plot above "Fig. 3" represents how SSIM values changed over time, offering a comparison of frame similarity before and after using the predictive approach. Before the algorithm, SSIM values varied greatly, showing inconsistent visual quality due to reactive bitrate changes. This often caused noticeable drops in quality, especially during complex scenes. After our predictive approach was implemented, SSIM values remained more stable and consistently higher, showing its ability to adjust encoding settings proactively. This led to smoother playback with fewer quality issues, proving the algorithm's effectiveness in maintaining visual quality and ensuring reliable streaming even with network changes.

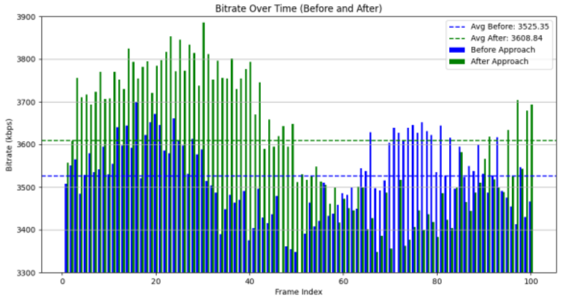

Fig. 4. Bitrate over time comparison before and after applying our approach.

The figure above "Fig. 4" demonstrates how bitrate changes over time. Before applying our method, the system uses a more aggressive bitrate adaptation, rapidly increasing the bitrate to fully utilize the available bandwidth. However, this approach often results in instability and sudden quality changes, particularly during network fluctuations. In contrast, our predictive method focuses on maintaining stability by adjusting the bitrate moderately, even when network conditions improve.

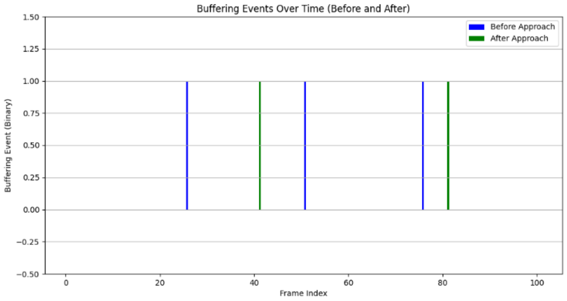

Fig. 5. Buffering events comparison before and after applying our approach.

The plot of buffering events over time "Fig. 5" compares the frequency and distribution of interruptions, before and after applying the predictive algorithm. Before applying our approach, buffering events were more common and spread across multiple frames reflecting the reactive nature of the streaming method, which struggles to quickly adjust to changing network conditions leading to frequent interruptions. On the other hand, after applying our predictive algorithm, the plot shows a significant decrease in buffering events, demonstrating that the predictive algorithm successfully anticipates network fluctuations and adjusts the video encoding proactively. This results in a much smoother streaming experience with fewer disruptions.

In this section, we compared the system's performance before and after applying our predictive approach, focusing on metrics like PSNR, SSIM, bitrate, and buffering events. The results show notable improvements with the AI-based algorithm. PSNR values are more stable and consistently higher, indicating better video quality. SSIM demonstrates fewer fluctuations keeping the visual quality intact. Bitrate adaptation is smoother enhancing the overall stability and reducing buffering. Moreover, the buffering events are significantly fewer, as the predictive algorithm anticipates network changes and adjusts the encoding accordingly leading to better quality of service. AI-driven adaptive streaming architectures have been shown to effectively adjust video quality and buffer sizes in response to network variability, resulting in enhanced user experiences [13].

## V. Baseline Comparison

Deep learning-based network performance prediction models have been utilized to train adaptive bitrate algorithms, improving the robustness and generalizability of streaming systems [14]. For evaluation purposes, we compare our proposed approach against baseline methods using key metrics, the following table “Table II” highlights how our method outperforms others in terms of PSNR, SSIM, latency, and buffering events.

TABLE II Performance Comparison of Proposed Approach vs. Baseline Methods

| Metric | Method | Value | Proposed Approach |
|---|---|---|---|
| PSNR(dB) | AMP [6] | 39,1 | 41,2 |
| | ALSTM [2] | 40,5 | |
| | Bp-LLM [7] | 39,8 | |
| | MD-AVB [8] | 38,7 | |
| SSIM | ALSTM [2] | 0,93 | 0,945 |
| | MLP [9] | 0,92 | |
| Latency (ms) | AMP [6] | 52 ms | 45 ms |
| | Realtime RNN[10] | 55 ms | |
| Buffering Events | AMP [6] | 4 events | 2 events |
| | MD-AVB [8] | 3 events | |

Now we will present a graphical representation of the results “Fig. 6”, to illustrate the performance metrics. The first graph, highlights a detailed analysis of PSNR values, showcasing the improvement of PSNR values by our approach, indicating enhanced video quality surpassing ALSTM [2] and other referenced methods.

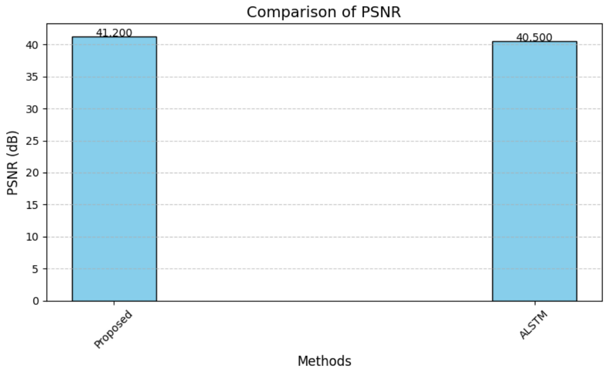

Fig. 6. PSNR comparison between our approach and ALSTM.

This next graph “Fig. 7” demonstrates the notable improvement of SSIM values achieved by our proposed method. The improvement of structural similarity showcases our method’s capability to maintain higher visual fidelity, especially under changing network conditions, outperforming ALSTM [2] and MLP [9].

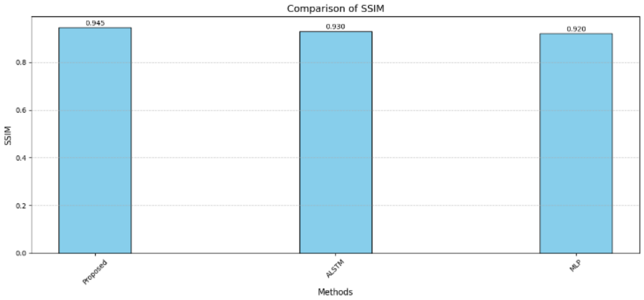

Fig. 7. SSIM comparison between our approach and ALSTM and MLP.

The following plot “Fig. 8” highlights the significant latency reduction attained by our approach. Surpassing AMP [6] and RNN [10], demonstrates the efficiency of our approach to deliver fast responses and adaptive coding.

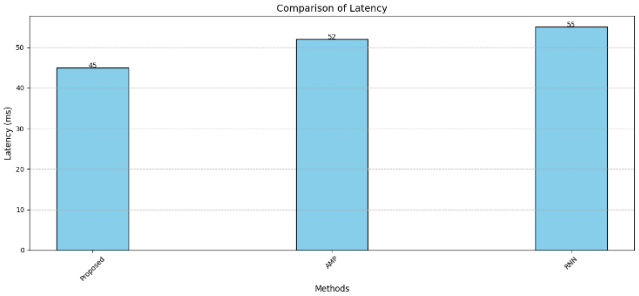

Fig. 8. Latency comparison between our approach and AMP and RNN.

This last graph “Fig. 9” shows the important decrease in buffering events achieved by our algorithm in comparison with AMP [6] and MD-AVB [8], the results demonstrate our method’s ability to deliver a smoother streaming experience with minimal disruptions.

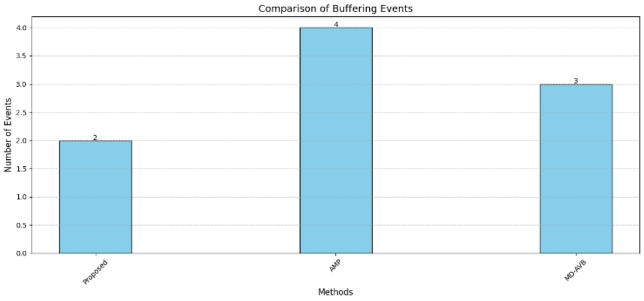

Fig. 9. Buffering events comparison between our approach and AMP and MD-AVB.

Our proposed approach demonstrates superior performance across different key metrics. It achieves an average PSNR of 41,2 dB, surpassing all methods stated in the related works section, with ALSTM [2] being the closest at 40,5 dB. In terms of structural similarity (SSIM), our approach attains a value of 0,945, outperforming ALSTM [2] attaining 0,93 and MLP [9] achieving 0,92. Furthermore, our predictive algorithm reduces latency to 45 ms, exceeding AMP [6] 52 ms and the Realtime RNN approach [10] 55 ms. Finally, our method reports only 2 buffering events on average, making a clear improvement over AMP [6] (4 events) and MD-AVB (3 events).

## VI. Conclusion

The integration of AI-driven predictive models represents a significant advancement in adaptive streaming technologies, offering the potential for more responsive and efficient video delivery systems [15]. This work introduced a novel AI-powered predictive approach for adaptive HEVC video transmission, addressing the challenges for real-time applications, particularly on resource-constrained devices. Furthermore, the applicability of our approach across diverse network environments. By leveraging Long Short-Term Memory (LSTM)) models, our system predicts bandwidth fluctuations in real-time, enabling proactive encoding adjustments that ensure smooth adaptation and improved viewing experiences. The experimental results demonstrate significant improvements, with PSNR reaching 41,2 dB, SSIM at 0,945, and latency reduced to an average of 45 ms, contributing to fewer buffering events and enhanced stability in video streaming.

Despite these advancements, certain limitations must be considered. The computational overhead of LSTM-based prediction models may pose challenges for real-time applications, particularly on resource-constrained devices. Furthermore, the applicability of our environments, including low-latency and high-mobility scenarios, requires further evaluation to ensure robust performance under varying conditions.

Future work will focus on optimizing the computational efficiency of our model, investigating alternative lightweight machine learning approaches such as GRUs or hybrid deep learning techniques to enhance adaptability for mobile and edge computing environments. Additionally, we plan to extend the system's capabilities for next-generation adaptive streaming, particularly in the context of 5G networks, to further improve real-time video delivery in ultra-low latency scenarios.

Overall, this research demonstrates the potential of AI-driven predictive models in enhancing adaptive HEVC streaming by providing stable video quality, reduced buffering, and efficient bandwidth utilization. Further developments will aim to refine its implementation for broader deployment in emerging network infrastructures.

## References

[1] J. Woo, S. Hong, D. Kang, et al., “Improving the Quality of Experience of Video Streaming Through a Buffer-Based Adaptive Bitrate Algorithm and Gated Recurrent Unit-Based Network Bandwidth Prediction,” Appl. Sci., vol. 14, no. 22, p. 10490, 2024.

[2] P. Li, X. Jiang, G. Jin, et al., "ALSTM: An Attention-Based LSTM Model for Multi-Scenario Bandwidth Prediction," in Proc. 2021 IEEE 27th Int. Conf. Parallel Distrib. Syst. (ICPADS), 2021, pp. 98-105.

[3] A. Lekharu, K. Y. Moulii, A. Sur, et al., "Deep Learning-Based Prediction Model for Adaptive Video Streaming," in Proc. 2020 Int. Conf. Commun. Syst. Netw. (COMSNETS), 2020, pp. 152-159.

[4] K. Khan, “Enhancing Adaptive Video Streaming Through AI-Driven Predictive Analytics for Network Conditions: A Comprehensive Review,” Int. Trans. Electr. Eng. Comput. Sci., vol. 3, no. 1, pp. 57–68, 2024.

[5] M. Tarik, et al., “Adaptive Video Streaming with AI-Based Optimization for Dynamic Network Conditions,” arXiv preprint arXiv:2501.18332, 2025.

[6] A. Bentaleb, A. C. Begen, S. Harous, et al., "Data-Driven Bandwidth Prediction Models and Automated Model Selection for Low Latency," IEEE Trans. Multimedia, vol. 23, pp. 2588-2601, 2020.

[7] P. Liu, X. Jiang, X. Wang, et al., "BP-LLM: A Large Language Model-Based Approach for Accurate and Adaptable Bandwidth Prediction," in THU 2024 Winter AML Submission, OpenReview, 2024.

[8] P. Zhang, C. An, Z. Wang, et al., "MD-AVB: A Multi-Manifold-Based Available Bandwidth Prediction Algorithm," Tsinghua Sci. Technol., vol. 25, no. 1, pp. 140-148, 2019.

[9] M. Labonne, C. Chatzinakis, and A. Olivereau, "Predicting Bandwidth Utilization on Network Links Using Machine Learning," in Proc. 2020 Eur. Conf. Netw. Commun. (EuCNC), 2020, pp. 242-247.

[10] L. Mei, J. Gou, Y. Cai, et al., "Realtime Mobile Bandwidth and Handoff Predictions in 4G/5G Networks," Comput. Netw., vol. 204, p. 108736, 2022.

[11] P. L. Vo, N. T. Nguyen, L. Luu, et al., “Federated Deep Reinforcement Learning-Based Bitrate Adaptation for Dynamic Adaptive Streaming over HTTP,” in Proc. Asian Conf. Intell. Inf. Database Syst., Cham: Springer Nature Switzerland, 2023, pp. 279–290.

[12] S. Wang, J. Lin, and F. Ye, “Imitation Learning for Adaptive Video Streaming with Future Adversarial Information Bottleneck Principle,” arXiv preprint arXiv:2405.03692, 2024.

[13] E. Artioli, “Generative AI for HTTP Adaptive Streaming,” in Proc. 15th ACM Multimedia Syst. Conf., 2024, pp. 516–519.

[14] D. Wu, P. Wu, M. Zhang, et al., “Mansy: Generalizing neural adaptive immersive video streaming with ensemble and representation learning,” IEEE Trans. Mobile Comput., 2024.

[15] Y. Mao, L. Sun, Y. Liu, et al., “Interactive 360° Video Streaming Using FoV-Adaptive Coding with Temporal Prediction,” arXiv preprint arXiv:2403.11155, 2024.