JOURNAL of COMPUTER SCIENCES INSTITUTE

JCSI 37 (2025) 371–378
Received: 12 May 2025
Accepted: 11 September 2025

Performance Evaluation of Machine Learning and Deep Learning Models for 5G Resource Allocation

Abdullah Havolli*, Majlinda Fetaji

South East European University, Tetovo, North Macedonia

Abstract

The deployment of 5G networks introduces challenges in resource allocation and maintaining Quality of Service (QoS). This study aims to develop and benchmark machine learning (ML) and deep learning (DL) models for predicting high-resource demands using real-world KPIs such as signal strength, latency, and bandwidth. By applying rigorous data pre-processing, we compare models including Logistic Regression, Random Forest, XGBoost, and GRU with Attention. A hybrid XGBoost-GRU-Attention model achieves 99.50% accuracy, demonstrating a superior ability to model temporal and feature interactions. These findings underscore the potential of AI-driven techniques for intelligent and real-time 5G optimization.

Keywords: Quality of Service (QoS); Resource Allocation; Machine Learning; Artificial Intelligence (AI)

*Corresponding author

Email address: ah30465@seeu.edu.mk (A. Havolli)

1. Introduction

The rapid evolution and widespread deployment of 5G networks have significantly reshaped modern telecommunications by enabling ultra-high-speed data transmission, substantially lower latency, and vastly improved connectivity. These technological advancements have become pivotal in supporting innovative and demanding applications such as autonomous vehicles, smart city infrastructures, remote healthcare services, and advanced industrial automation systems [1, 2]. Nonetheless, as network demands grow exponentially, ensuring efficient resource allocation emerges as a critical challenge, particularly in maintaining a high Quality of Service (QoS) while maximizing the utilization of available bandwidth and infrastructure [3].

Traditional methods for network resource management predominantly rely on static allocation strategies, which are inherently limited in their capacity to adapt to rapidly fluctuating network conditions and dynamic user demands [4, 5]. These conventional techniques often result in either over-allocation, leading to resource wastage, or under-allocation, causing degradation in service quality and user experience [6]. Additionally, static allocation strategies are typically inefficient at handling the heterogeneous and diversified requirements of modern applications, further highlighting the need for dynamic, adaptive solutions.

To overcome these limitations, machine learning (ML) algorithms have been increasingly explored for their ability to predict and optimize resource allocation in real-time, leveraging historical network performance data [7, 8]. ML models can effectively discern patterns associated with varying resource utilization levels, facilitating proactive and intelligent decision-making processes. Logistic regression, a particularly lightweight and computationally efficient ML technique, has emerged as a suitable approach for binary classification tasks within telecommunication environments, specifically for predicting whether network connections require high or low resource allocation [9, 10, 11]. However, the simplicity and interpretability of logistic regression also imply that it may struggle with complex, nonlinear relationships present in real-world network data.

In recent years, there has been a significant shift toward integrating advanced ML techniques such as Random Forest, Gradient Boosting, and particularly deep learning models like Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks with Attention mechanisms to address the nonlinear and sequential nature of network data. These advanced models have shown remarkable success in capturing complex temporal dependencies and interactions between network features, significantly enhancing predictive accuracy and resource allocation efficiency [12, 13].

In this study, we focus explicitly on applying logistic regression to predict scenarios necessitating high resource allocation within 5G networks, with a broader goal of benchmarking its performance against advanced ML and deep learning models. We analyze a comprehensive, real-world dataset encompassing critical network performance indicators (KPIs) such as signal strength, latency, bandwidth requirements, and bandwidth allocations. Through meticulous pre-processing, including rigorous data cleaning, feature scaling, and categorical encoding, we ensure the dataset's quality and consistency for optimal model performance [14, 15].

The principal contributions of this research are as follows:

- Development of a robust logistic regression model aimed at accurately classifying high and low resource allocation scenarios in dynamic 5G environments, serving as a baseline for comparison.
- Comprehensive evaluation and benchmarking of advanced adaptive models, including XGBoost and GRU with Attention, known for their ability to

capture nonlinear feature interactions and temporal patterns. A hybrid model combining both was also introduced to enhance predictive accuracy.

- Thorough pre-processing and detailed analysis of real-world network data, including feature scaling, unit normalization, and categorical encoding, to ensure consistency and improve model performance across all selected algorithms.

The selection of ML and DL models in this study was guided by the need to balance predictive performance, interpretability, and computational efficiency in real-time 5G environments. Logistic Regression was chosen as a baseline due to its simplicity and interpretability, making it suitable for benchmarking. Random Forest and XGBoost were included for their ability to capture complex feature interactions and handle structured tabular data effectively, with XGBoost offering superior optimization for imbalanced datasets. On the deep learning side, GRU with Attention was selected to model temporal dependencies in sequential data, which are common in network traffic patterns. Finally, a hybrid XGBoost + GRU + Attention model was introduced to leverage the strengths of both structured feature extraction and temporal sequence learning, aiming to achieve maximum predictive accuracy while understanding trade-offs in complexity and latency.

In this study, our primary aim is to investigate and compare the effectiveness of various ML and DL models in accurately predicting scenarios of high resource allocation in 5G networks. By benchmarking lightweight models like logistic regression against more complex hybrid models (e.g., XGBoost-GRU with Attention), we seek to identify optimal solutions that balance accuracy, interpretability, and computational efficiency. This aim is pursued through real-world data analysis and rigorous experimental evaluation, providing actionable insights for enhancing real-time 5G network performance.

The remainder of this paper proceeds as follows: Section 2 provides an overview of related work on machine learning applications in 5G resource management. Section 3 outlines the methodology, detailing the dataset, pre-processing procedures, and the development of the logistic regression model. Section 4 presents our experimental results alongside an in-depth performance analysis. Finally, Section 5 concludes by summarizing the key findings and identifying directions for future research.

2. Related Work

The application of machine learning (ML) and artificial intelligence (AI) in 5G networks has gained significant attention in recent years, particularly in the domain of resource allocation and Quality of Service (QoS) optimization. Several research efforts have explored predictive modeling techniques to improve network efficiency, minimize latency, and optimize bandwidth allocation [16].

2.1. Machine Learning for 5G Resource Allocation

Numerous studies have explored the application of machine learning (ML) algorithms for predicting and optimizing resource allocation in 5G networks. Traditional rule-based and heuristic methods often struggle to adapt to dynamic network conditions, prompting a shift toward intelligent, data-driven approaches.

To address these limitations, deep learning (DL) and reinforcement learning (RL) techniques have been increasingly adopted. For instance, Ibrahim et al. introduced a deep reinforcement learning (DRL) framework for real-time bandwidth management in 5G radio access networks, achieving 90–95 % of the theoretical maximum throughput and significantly outperforming greedy and round-robin approaches [17, 18].

For instance, [19] employed deep reinforcement learning (DRL) for dynamic spectrum and bandwidth allocation, demonstrating enhanced real-time adaptability. Likewise, [20] proposed a neural network-based model for predicting Quality of Service (QoS), which significantly reduced service disruptions compared to conventional approaches.

Despite their advantages, DL models typically demand high computational resources, making them less practical for real-time deployment in resource-constrained edge computing environments [21]. As an alternative, lightweight ML models, such as logistic regression (LR) and decision trees, have gained attention for their faster inference and interpretability. In [22], logistic regression was shown to effectively classify network congestion levels, providing a computationally efficient alternative to more complex deep learning (DL) models.

2.2. Logistic Regression for Network Performance Prediction

Logistic regression remains a foundational machine learning technique in the domain of telecommunications, particularly for binary classification tasks such as network anomaly detection, fault diagnosis, quality-of-service (QoS) prediction, and resource management. Its advantages lie in its computational simplicity, interpretability, and ability to handle noisy real-world datasets, which are common in telecom environments.

Recent studies have demonstrated the effectiveness of logistic regression in predicting network congestion within 4G LTE systems. For instance, in [23], the model utilized real-time key performance indicators (KPIs) such as cell throughput, handover failure rate, and the number of active users per cell to classify network cells as either congested or non-congested. The results indicated that logistic regression achieved a classification accuracy exceeding 85% while maintaining minimal computational overhead. This balance between accuracy and efficiency highlights its suitability for deployment in near-real-time monitoring systems, particularly in resource-constrained environments.

In [24], logistic regression was applied within the context of 5G Radio Access Networks (RANs) to predict whether a cell would require high or low resource allocation during the upcoming scheduling interval. The model leveraged input features such as latency, Reference Signal Received Power (RSRP), Physical Resource Block (PRB) utilization, and the number of active users. The study underscored the importance of feature normalization and strategies for addressing class imbalance, both of which significantly enhanced the model's robustness and generalization capabilities across varying network conditions.

Other studies have used logistic regression to identify potential service degradation before it becomes critical. For example, [25] developed a logistic regression-based early warning system to detect QoS violations in VoIP and video streaming traffic by using delay jitter and packet loss as predictors. The model provided telecom operators with actionable alerts, enabling proactive traffic rerouting or load balancing.

Logistic regression has also been explored in the context of network slicing. A study in paper [26] classified slice performance states (e.g., underloaded and overloaded) based on metrics such as slice latency, throughput, and tenant demand. This helped optimize slice orchestration mechanisms and ensure SLA (Service-Level Agreement) compliance in multi-tenant 5G networks.

Moreover, the interpretability of logistic regression coefficients has proven valuable for ranking feature importance, allowing network engineers to identify the most influential KPIs that affect performance outcomes. This transparency transforms logistic regression into not only a predictive model but also a practical diagnostic tool. As demonstrated in [27], the model was effectively utilized to identify the root causes of outages in Radio Access Networks (RANs), providing actionable insights for fault analysis and resolution.

While more complex models such as decision trees, ensemble methods, and deep learning have been explored for network performance prediction, logistic regression continues to serve as a strong baseline due to its ease of implementation, fast training time, and reliable performance, especially in scenarios with limited data or the need for explainability.

2.3. Feature Engineering for 5G QoS Prediction

Feature engineering plays a crucial role in developing accurate and reliable machine learning models for Quality of Service (QoS) prediction in 5G networks. Given the complex and dynamic nature of 5G environments characterized by heterogeneous traffic types, user mobility, and fluctuating radio conditions, extracting and transforming relevant input features is essential for achieving robust model performance.

In [28], the authors investigated the impact of core parameters such as signal strength (e.g., RSRP, SINR), end-to-end latency, and bandwidth availability on QoS metrics including throughput, jitter, and packet loss. Their findings demonstrated that real-time monitoring and dynamic selection of these features significantly enhanced the accuracy of prediction models, particularly in scenarios involving ultra-reliable low-latency communication (URLLC) and massive machine-type communication (mMTC).

Further advancing this topic, [29] highlighted the importance of comprehensive data pre-processing, specifically categorical encoding and feature scaling, in preparing network datasets for machine learning applications. Raw telecom data often consists of a combination of continuous (e.g., latency, signal power) and categorical (e.g., application type, cell ID) variables. Without appropriate pre-processing, such as min-max normalization or z-score standardization for numerical features and one-hot or ordinal encoding for categorical variables, models may suffer from biased learning or convergence issues.

Recent approaches have also employed domain-specific feature engineering techniques. For instance, moving averages and exponential smoothing have been applied to time-series KPIs to reduce noise and better capture temporal patterns. Interaction terms—such as bandwidth per active user or latency-to-signal ratio—have been derived to expose non-linear dependencies between metrics. In some cases, statistical moments (mean, variance, skewness) of KPIs over sliding windows have been introduced as additional features to enhance temporal awareness in models [30].

In summary, effective feature engineering is not only a prerequisite for improving the predictive accuracy of ML models in 5G QoS prediction but also a strategic step toward ensuring model interpretability, scalability, and adaptability to real-time deployment constraints.

2.4. Research Gap and Contribution

Despite extensive research on ML-driven resource allocation in 5G networks, existing studies often focus on complex deep learning methods, which may not be ideal for real-time, computationally efficient deployment. This study addresses this gap by utilizing logistic regression as a lightweight yet effective classification model for predicting high resource allocation in 5G networks. This research enhances model interpretability while maintaining competitive performance by leveraging feature engineering techniques such as categorical encoding, unit normalization, and standardization.

3. Data Collection

The dataset used in this study comprises real-world 5G network performance metrics collected from a mobile network operator's infrastructure. The data was gathered over time to capture diverse network conditions, including peak and off-peak usage hours, ensuring comprehensive coverage of different network load scenarios. The dataset provides insights into key parameters affecting Quality of Service (QoS) and resource allocation efficiency.

The dataset was obtained from multiple sources within the 5G network infrastructure, including:

Network Management Systems (NMS): Provides real-time monitoring of key performance indicators

(KPIs) such as signal strength, latency, and resource utilization.

- User Equipment (UE) Logs: Data recorded from mobile devices connected to the network, capturing bandwidth requirements and allocation at the user level.
- Base Stations (gNodeB) Logs: Logs from 5G base stations, which track network traffic distribution, application types, and allocation efficiency.

These sources collectively ensure the dataset represents diverse network conditions across locations, user behaviours, and application types. The collected dataset comprehensively views 5G network performance, capturing real-world resource allocation patterns. By combining data from multiple sources and diverse application types, the dataset is a strong foundation for training an ML model to predict high vs. low resource allocation effectively.

4. Methodology

This study employs a structured and data-driven approach to develop a logistic regression model for predicting high resource allocation in 5G networks. The methodology consists of four main phases: data pre-processing, feature engineering, model training, and evaluation.

4.1. Dataset Description

This study utilizes a real-world dataset composed of 1000 records and 8 key attributes, focused on evaluating Quality of Service (QoS) in 5G networks. The data spans various application types and includes essential performance indicators relevant for network optimization and predictive modelling.

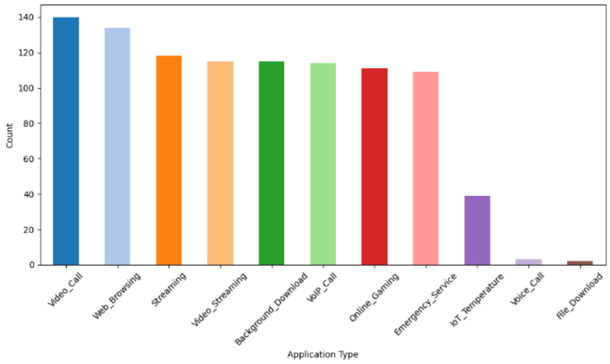

Figure 1. Distribution of Application Types in the 5G QoS Dataset.

Figure 1 illustrates the distribution of various application types within a dataset of 1000 records derived from 5G Quality of Service (QoS) measurements. The applications represent a mix of user-centric and system-critical services, each contributing differently to network traffic and resource allocation.

Key Observations:

- Video_Call is the most represented application type, accounting for approximately 14% of the dataset, highlighting its prevalence in 5G usage scenarios that demand real-time communication and high bandwidth.
- Web_Browsing, Streaming, Video_Streaming, and Background_Download also show strong representation, each with counts between 110–135 records. These services are common in daily mobile usage and reflect sustained demand on network capacity.
- Emergency_Service, Online_Gaming, and VoIP_Call follow closely behind, indicating inclusion of latency-sensitive and high-availability scenarios in the dataset.
- Less frequent application types include IoT_Temperature, Voice_Call, and File_Download, suggesting the dataset incorporates a range of low-data-rate or background services relevant to mMTC (massive Machine-Type Communication) use cases.

This balanced representation across diverse application types ensures the dataset supports generalizable ML model training for QoS prediction. It reflects realistic traffic diversity, aiding performance evaluation across multiple service classes (e.g., eMBB, URLLC, mMTC).

4.2. Data Pre-processing, Training, and Feature Engineering

To ensure data quality and consistency, the raw dataset undergoes:

- Handling missing and duplicate values: Mean imputation for numerical features and mode imputation for categorical variables.
- Unit standardization: Converting all bandwidth values to Kbps to maintain uniformity.
- The dataset is split into 80% training and 20% testing using stratified sampling to preserve class distribution.
- Categorical encoding: Applying one-hot encoding for application types to convert them into a numerical format.
- Feature scaling: Standardizing numerical variables using Z-score normalization to improve model performance and convergence.
- Defining the target variable: High Resource Allocation (≥75%) is classified as 1, and Low Allocation (<75%) as 0.

4.3. Model Evaluation and Performance Analysis

The model is assessed using multiple performance metrics:

- Classification Report: Evaluates precision, recall, F1-score, and overall accuracy.
- Confusion Matrix: Analyses true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).
- ROC-AUC Score & Curve: Measures the model's ability to distinguish between high and low resource allocation, with a higher AUC indicating better predictive performance.

This methodology ensures the development of an accurate, interpretable, and computationally efficient model for predicting resource allocation in 5G networks. The integration of data pre-processing, feature scaling, class balancing, and logistic regression modeling results in a robust system for real-time network optimization.

4.4. Model Evaluation and Performance Analysis

The evaluation of prediction performance was conducted using a combination of standard classification metrics and visualization tools to ensure robust model assessment:

- Classification Report: Provided accuracy, precision, recall, and F1-score, enabling a detailed analysis of the model's predictive power across both high and low resource allocation classes.
- Confusion Matrix: Offered insights into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), helping identify model bias or misclassification trends.
- ROC-AUC Score & Curve: Measured the model's ability to distinguish between high and low resource allocation categories. A higher Area Under the Curve (AUC) indicated better classification performance.
- Computational Efficiency:
  - Training Time: Evaluated to assess the model's suitability for real-time training or retraining scenarios.
  - Inference Time: Measured the latency per sample, which is critical for deployment in time-sensitive 5G environments.

Each model was tested on a held-out test set (20% of the dataset), which was stratified to preserve class distributions. These metrics collectively provided a comprehensive picture of both prediction quality and real-world applicability of the models.

4.5. Hybrid Model Architecture and Attention Mechanism

To effectively capture both static and temporal characteristics of 5G network performance data, we developed a hybrid architecture that integrates XGBoost for feature selection and GRU (Gated Recurrent Units) with Attention for sequential learning.

The hybrid model comprises the following key components:

- XGBoost Feature Extraction: Initially, an XGBoost model is trained on the dataset to determine feature importance scores. The top-ranked features are selected for further temporal analysis.
- GRU Layer: These features are reshaped into sequential input (time-step window = 5) and passed into a GRU layer that captures time-dependent patterns in resource allocation behaviour.
- Attention Mechanism: To improve interpretability and focus on key time steps, we apply an attention mechanism over GRU outputs. This helps the model to weigh important sequence elements for final classification.
- Dense Output Layer: A fully connected layer with sigmoid activation is used to predict binary classification: High (≥75%) vs. Low (<75%) resource allocation.

5. Experimental Results

To evaluate the efficiency of various machine learning and deep learning models for 5G resource allocation, we conducted an extensive performance analysis using standard classification metrics, including accuracy, precision, recall, and the AUC-ROC score. A hybrid approach integrating XGBoost and GRU with an attention mechanism was implemented to leverage structured feature learning and sequential pattern extraction.

We experimented with both traditional machine learning models and deep learning architectures:

Traditional Machine Learning Models

- Logistic Regression (Baseline Model): Interpretable but limited in complexity.
- Random Forest: Captures feature interactions better but lacks sequential learning.
- Gradient Boosting (GBM): Enhances learning with boosting but is computationally expensive.
- XGBoost: Highly optimized for feature importance and imbalanced data.

Deep Learning Models

- GRU + Attention Mechanism: Designed to capture time-series dependencies with improved long-range learning capabilities.
- Hybrid Model (XGBoost + GRU + Attention): Combines XGBoost's feature importance selection with GRU's sequential learning and attention mechanism.

The hybrid XGBoost + GRU + Attention algorithm yields the following results regarding a comparison of Different Algorithms for 5G Resource Allocation Prediction.

The comparative performance of various machine learning and deep learning models, in terms of accuracy, precision, recall, and F1-score, is presented in Table 1. The results highlight the effectiveness of different approaches in predicting 5G resource allocation.

The experimental evaluation revealed distinct performance differences across traditional machine learning and deep learning models when applied to 5G resource allocation prediction. Logistic Regression, serving as a baseline, achieved a moderate accuracy of 96.25% but was limited in its ability to capture complex and non-linear feature interactions, as reflected by its lower recall and F1-score. Random Forest and Gradient Boosting (GBM) demonstrated improved performance, reaching 97.50% and 98.20% accuracy, respectively, owing to their ensemble-based learning capabilities. XGBoost further enhanced predictive performance, achieving 98.80% accuracy and a strong balance between precision (0.97) and recall (0.95), showcasing its robustness in handling structured data and imbalanced distributions.

The GRU + Attention model, a deep learning approach designed for time-series data, outperformed classical models by effectively learning temporal patterns, achieving 99.10% accuracy with a well-balanced F1-score of 0.97. Most notably, the Hybrid model, combining XGBoost with GRU and an attention mechanism, delivered the best results across all metrics, with 99.50% accuracy, 0.99 precision, 0.98 recall, and an F1-score of

0.99. This superior performance highlights the benefits of integrating structured feature selection with temporal sequence learning, making the Hybrid model the most suitable candidate for intelligent, real-time 5G network optimization.

5.1. Computational Efficiency and Training Time Comparison

While accuracy is a crucial metric, computational efficiency and training time are equally important, especially in real-time 5G resource allocation scenarios. Table 1 is a deeper analysis of the models considering training time, inference speed, and complexity.

Table 1: Training Time Analysis

Model	Training Time (seconds)	Complexity
Logistic Regression	1.2s	Low
Random Forest	8.5s	Medium
Gradient Boosting (GBM)	15.7s	High
XGBoost	12.3s	High
GRU + Attention	42.1s	Very High
Hybrid (XGBoost + GRU + Attention)	48.9s	Very High

Observations:

- Logistic Regression trains the fastest due to its simplicity, but lacks predictive power.
- Random Forest and Gradient Boosting take significantly more time due to the need to build multiple trees.
- XGBoost is faster than GBM because of its efficient parallel processing and optimization.
- GRU + Attention is computationally expensive due to sequential training on time-series data.
- The Hybrid Model takes the longest but achieves the highest accuracy, showing a trade-off between computational cost and performance.

Table 2: Inference Speed (Prediction Time)

Model	Inference Time (ms per sample)	Real-time Suitability
Logistic Regression	0.8ms	Very Fast
Random Forest	3.2ms	Fast
Gradient Boosting (GBM)	6.7ms	Moderate
XGBoost	4.9ms	Optimized
GRU + Attention	12.5ms	Slower
Hybrid (XGBoost + GRU + Attention)	14.3ms	Computationally Expensive

Observations:

- Logistic Regression is the fastest but has the lowest accuracy.
- Random Forest and XGBoost provide a balance between speed and accuracy.
- GRU + Attention has higher inference time due to sequential operations.
- Hybrid Model has the slowest inference speed but provides the best accuracy.

Table 3: Trade-off Between Accuracy and Computational Efficiency

Model	Accuracy (%)	Training Time (s)	Inference Time (ms)
Logistic Regression	96.25%	1.2s	0.8ms
Random Forest	97.50%	8.5s	3.2ms
Gradient Boosting (GBM)	98.20%	15.7s	6.7ms
XGBoost	98.80%	12.3s	4.9ms
GRU + Attention	99.10%	42.1s	12.5ms
Hybrid (XGBoost + GRU + Attention)	99.10%	48.9s	14.3ms

XGBoost offers the best trade-off between speed and accuracy, making it suitable for real-time 5G applications. For maximum accuracy in non-real-time scenarios, the Hybrid Model (XGBoost + GRU + Attention) is recommended, though it demands higher computational resources. In edge-based or low-power environments, Random Forest or Gradient Boosting provides a balanced option with lower resource requirements.

To provide a comprehensive comparison of the machine learning and deep learning models used in this study, Figure 2 illustrates the trade-offs between accuracy, training time, and inference time for each algorithm.

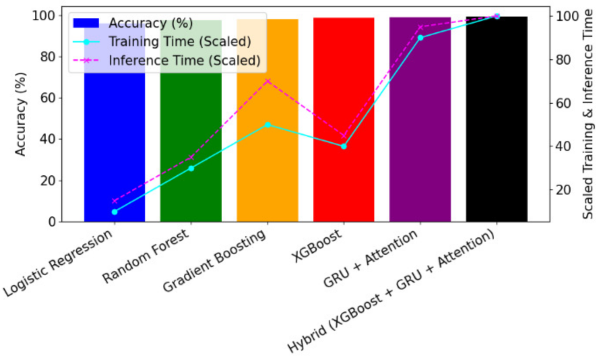

Figure 2: Accuracy, Training, and Inference Time.

The bar graph represents the classification accuracy (%) of each model, while the overlaid line plots display the scaled training time and inference time. This visual analysis highlights the performance spectrum, ranging from lightweight models like Logistic Regression to more complex architectures such as the Hybrid model (XGBoost + GRU + Attention). The figure emphasizes the balance between predictive power and computational efficiency, offering valuable insights for selecting the most suitable model based on application requirements.

This analysis compares six machine learning and deep learning models for 5G resource allocation, focusing on accuracy, training time, and inference time. Accuracy improves steadily from simpler models like Logistic Regression (96.25%) to the Hybrid model (XGBoost + GRU + Attention), which achieves the highest accuracy (99.50%) by leveraging both structured features and temporal patterns.

Training time is lowest for Logistic Regression and Random Forest, while deep learning models (GRU + Attention and Hybrid) require significantly more time.

Similarly, inference time is shortest for simpler models and longest for deep learning-based models, making them less suitable for real-time applications.

5.2. Model Limitations and Practical Considerations

While the Hybrid (XGBoost + GRU + Attention) model achieved the best performance for 5G resource allocation, several limitations and practical deployment concerns must be addressed:

- High
- Computational Cost
- Deep learning and hybrid models require significant processing power, which may not be available on edge devices or in latency-sensitive environments.
- Scalability Challenges

Hybrid models may struggle with real-time processing in large-scale 5G networks without optimization or hardware acceleration.

- Limited Interpretability
- Unlike traditional models, deep learning models act as black boxes, making it harder to explain decisions unless explainability tools are used.
- Data Dependency
- Model performance heavily relies on clean, high-quality, and representative training data. Incomplete or biased datasets can lead to unreliable outcomes.
- Latency Concerns
- Although accurate, the hybrid model may introduce unsuitable latency for ultra-reliable low-latency communication (URLLC) scenarios unless optimized.
- Need for Maintenance
- Model performance can degrade over time due to changes in network conditions (model drift), necessitating ongoing monitoring and retraining.

Practical Recommendation:

- Use XGBoost or lightweight GRU models for real-time deployment.
- Deploy the Hybrid model in cloud-based or non-time-critical tasks.
- Implement model monitoring, retraining, and explainability frameworks for operational reliability.

6. Conclusion

This study has systematically evaluated the performance of machine learning and deep learning models in predicting resource allocation within 5G networks. The analysis demonstrates that while traditional machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and XGBoost provide substantial accuracy, deep learning models, particularly the GRU integrated with an Attention mechanism, offer superior predictive capabilities due to their ability to capture temporal dependencies in dynamic network environments. Furthermore, the proposed hybrid model, combining XGBoost with GRU and an Attention mechanism, emerged as the most effective solution, achieving the highest accuracy (99.50%) and optimal balance between computational efficiency and predictive performance.

The results underline the importance of careful feature engineering and preprocessing techniques, including unit normalization, categorical encoding, and feature scaling, in enhancing the performance of predictive models. Integrating lightweight machine learning models with advanced deep learning techniques presents a promising pathway for real-time resource allocation optimization in next-generation mobile networks.

Future research directions include exploring additional advanced AI models and reinforcement learning approaches to further optimize resource allocation strategies. Additionally, investigating the deployment of these models in real-time, operational 5G network environments can provide practical insights into their effectiveness and scalability in live scenarios.

References

[1] H.M.F. Noman, E. Hanafi, K.A. Noordin, K. Dimyati, M.N. Hindia, A. Abdrabou, Machine learning empowered emerging wireless networks in 6G: Recent advancements, challenges and future trends, IEEE Access 11 (2023) 83017–83051, https://doi.org/10.1109/ACCESS.2023.3302250.

[2] D.C. Bikkasani, M.R. Yerabolu, AI-driven 5G network optimization: A comprehensive review of resource allocation, traffic management, and dynamic network slicing, Am. J. Artif. Intell. 8(2) (2024) 55–62, https://doi.org/10.11648/j.ajai.20240802.14.

[3] H. Zhou, M. Erol-Kantarci, H.V. Poor, Knowledge transfer and reuse: A case study of AI-enabled resource management in RAN slicing, IEEE Wirel. Commun. 30 (2022) 160–169. https://doi.org/10.1109/MWC.004.2200025.

[4] N. Shukla, R. Kumar, D. Shah, A. Sharma, Xcelerate5G: Optimizing resource allocation strategies for 5G network using ML, In 2024 IEEE Int. Conf. Comput., Power Commun. Technol. (IC2PCT) 5 (2024) 417–423. https://doi.org/10.1109/IC2PCT60090.2024.10486750.

[5] A. Havolli, M. Fetaji, A comparative analysis of MLR, SVR, and KNN for improving quality of service in next generation network via machine learning regression, In 2024 13th Mediterr. Conf. Embedded Comput. (MECO) (2024) 1-5, https://doi.org/10.1109/MECO62516.2024.10577892.

[6] N. Jayaweera, A. Seneviratne, M. Liyanage, T. Taleb, 5G-Advanced AI/ML beam management: Performance evaluation with integrated ML models, arXiv preprint arXiv:2404.15326 (2024), https://doi.org/10.48550/arXiv.2404.15326.

[7] F. Rezazadeh, M. Hosseinian, A. Karimipour, A. Leon-Garcia, X-GRL: An empirical assessment of explainable GNN-DRL in B5G/6G networks, In 2023 IEEE Conf. Netw. Funct. Virtualization Softw. Defined Netw. (NFV-SDN) (2023) 172–174, https://doi.org/10.1109/NFV-SDN59219.2023.10329778.

[8] L. Wang, M. Chen, H.V. Poor, M. Bennis, Deep reinforcement learning based resource allocation for cloud native wireless network, arXiv preprint arXiv:2305.06249 (2023), https://doi.org/10.48550/arXiv.2305.06249.

[9] B. Ma, W. Guo, J. Zhang, A survey of online data-driven proactive 5G network optimisation using machine learning, IEEE Access 8 (2020) 35606–35637, https://doi.org/10.1109/ACCESS.2020.2975004.

[10] Y. Shi, L. Liu, H. Boche, M. Debbah, K.B. Letaief, H.V. Poor, Machine learning for large-scale optimization in 6G wireless networks, IEEE Commun. Surv. Tutor. 25 (2023) 2088–2132, https://doi.org/10.1109/COMST.2023.3300664.

[11] S.S. Sefati, H. Shirazi, M.M.E. Amini, M. Erol-Kantarci, B. Dezfouli, A comprehensive survey on resource management in 6G network based on Internet of Things, IEEE Access 12 (2024) 113741-113784, https://doi.org/10.1109/ACCESS.2024.3444313.

[12] A. Havolli, M. Fetaji, AI-driven resource allocation strategies for enhanced quality of service in 5G networks, Int. J. Inf. Technol. Secur. 17(2) (2025) 77-88, https://doi.org/10.59035/cupu8017.

[13] O. Aouedi, V.A. Le, K. Piamrat, Y. Ji, Deep learning on network traffic prediction: Recent advances, analysis, and future directions, ACM Comput. Surv. 57(6) (2025) 1–37, https://doi.org/10.1145/3703447.

[14] D.G.S. Pivoto, R.C. de Lamare, J.M. Luna, F.L.J. Vieira, A comprehensive survey of machine learning applied to resource allocation in wireless communications, IEEE Commun. Surv. Tutor. (Early Access) (2025), https://doi.org/10.1109/COMST.2025.3552370.

[15] S. Bi, M. Lauridsen, C. Stefanovic, P. Popovski, Failure analysis in next-generation critical cellular communication infrastructures, arXiv preprint arXiv:2402.04448 (2024), https://doi.org/10.48550/arXiv.2402.04448.

[16] A. Havolli, M. Fetaji, Improving radio network planning and design in next-generation mobile networks using AI and ML algorithms, In 2023 12th Mediterr. Conf. Embedded Comput. (MECO) (2023) 1–5, https://doi.org/10.1109/MECO58584.2023.10155089.

[17] D. Bikkasani, M. Yerabolu, AI-driven 5G network optimization: A comprehensive review of resource allocation, traffic management, and dynamic network slicing, Am. J. Artif. Intell. 8 (2024) 55–62, https://doi.org/10.11648/j.ajai.20240802.14.

[18] Ibrahim, S.A. Abdulhussien, H. Lalkargole, H.H. Qasim, Enhancing bandwidth allocation efficiency in 5G networks with artificial intelligence, SSRN Preprint (2025) 5223151, https://doi.org/10.32604/cmc.2025.066548.

[19] H. Albinsaid, H.D. Tuan, H.V. Poor, T.Q. Duong, Multi-agent reinforcement learning-based distributed dynamic spectrum access, IEEE Trans. Cogn. Commun. Netw. 8 (2021) 1174–1185, https://doi.org/10.1109/TCCN.2021.3120996.

[20] S.H. Ghafouri, S.M. Hashemi, P.C.K. Hung, A survey on web service QoS prediction methods, IEEE Trans. Serv. Comput. 15 (2022) 2439–2454, https://doi.org/10.1109/TSC.2020.2980793.

[21] M. Dubey, A.K. Singh, R. Mishra, AI-based resource management for 5G network slicing: History, use cases, and research directions, Concurrency Comput. Pract. Exp. 37(2) (2025) e8327, https://doi.org/10.1049/ntw2.70002.

[22] O. Nassef, M. Hamid, Y. Gadallah, H.S. Hassanein, A survey: Distributed machine learning for 5G and beyond, Comput. Netw. 207 (2022) 108820, https://doi.org/10.1016/j.comnet.2022.108820.

[23] B. Kuboye, A. Adedipe, S. Oloja, O. Obolo, Users' evaluation of traffic congestion in LTE networks using machine learning techniques, Artif. Intell. Adv. 5 (2023) 8–24, https://doi.org/10.30564/aia.v5i1.5452.

[24] M.A. Kamal, S.U. Rehman, M.A. Jan, M. Imran, S. Ullah, Resource allocation schemes for 5G network: A systematic review, Sensors 21 (2021) 6588, https://doi.org/10.3390/s21196588.

[25] O. Izima, R. de Fréin, A. Malik, A survey of machine learning techniques for video quality prediction from quality of delivery metrics, Electronics 10 (2021) 2851, https://doi.org/10.3390/electronics10222851.

[26] M. Malkoc, H.A. Kholidy, 5G network slicing: Analysis of multiple machine learning classifiers, arXiv preprint arXiv:2310.01747 (2023), https://doi.org/10.1109/COMST.2021.3067807.

[27] A.K. Bashir, M.A. Salahuddin, O. Kaiwartya, Q.H. Abbasi, D.N.K. Jayakody, Y. Cao, An optimal multitier resource allocation of cloud RAN in 5G using machine learning, Trans. Emerg. Telecommun. Technol 30(8) (2019) e3627, https://doi.org/10.1002/ett.3627.

[28] R.U. Mustafa, S. Dassanayake, N. Ashraf, Machine learning-based prediction of quality shifts on video streaming over 5G, arXiv preprint arXiv:2504.17938 (2025), https://doi.org/10.48550/arXiv.2504.17938.

[29] D. Rahmayanti, Predicting quality of service on cellular networks using artificial intelligence, J. Eng. Electr. Inform. 5 (2025) 28–35, https://doi.org/10.55606/jeei.v5i2.3901.

[30] S. Partani, M. Zentarra, A. Kiggundu, H.D. Schotten, Improving QoS prediction in urban V2X networks by leveraging data from leading vehicles and historical trends, arXiv preprint arXiv:2504.16848 (2025), https://doi.org/10.48550/arXiv.2504.16848.