Journal of Network and Computer Applications 242 (2025) 104260

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Predictive modeling of RRC inactive transitions and latency impacts for energy optimization in live NR SA networks

Roopesh Kumar Polaganga [a,b,*], Qilian Liang [b]

[a] T-Mobile US Inc, 98012, Bellevue, WA, USA
[b] The University of Texas at Arlington, 76010, Arlington, TX, USA

ARTICLE INFO

Handling Editor: Dr. M. Atiquzzaman

Index Terms:
Machine Learning (ML)
Weighted ensemble learning
5G New Radio (NR)
RRC Inactive
6G
Energy savings
Live telecommunication networks

ABSTRACT

With the rapid evolution of 5G and the anticipated advancements in future 6G networks, machine learning is unlocking unprecedented opportunities for network optimization. Among the most significant advancements in 5G Standalone (SA) networks is the Radio Resource Control (RRC) Inactive state, a feature that is critical for achieving low-latency performance. Building on this foundation, our study is categorized into two key contributions. First, we present a novel application of ensemble machine learning to predict transitions from the RRC Inactive state, specifically distinguishing between RRC Resume and RRC Fallback requests. This predictive capability, developed using real-world New Radio (NR) SA network data, offers insights into previously unexplored transition behavior. Second, we demonstrate how this predictive capability can be applied to optimize gNodeB (gNB) operations, proactively managing User Equipment (UE) contexts to minimize unnecessary paging and processing overhead. Our findings show that the proposed framework achieves considerable energy savings while maintaining latency requirements critical to RRC Inactive mechanisms. These results underscore the practicality and scalability of machine learning-driven approaches to enhance network resource allocation and operational efficiency in 5G SA networks, providing a pathway to sustainable and high-performing next-generation networks.

1. Introduction

Amidst the rapid advancements in wireless communication, Machine Learning (ML) is transforming the landscape of innovation. As the industry shifted from 4G LTE—renowned for its reliable high-speed mobile internet—to 5G NR, which promises significantly faster data rates and reduced latency, we are setting the stage for the emergence of 6G networks (Chen et al., 2023). In this transition, ML takes a leading role, offering intelligent solutions that unlock new capabilities. Its proficiency in analyzing large datasets, uncovering complex patterns, and making data-driven decisions has positioned ML as a key driver for gaining deeper network insights. Moreover, ML's ability to predict user behavior enables it to optimize network performance and enhance the overall efficiency of telecommunications systems (Li et al., 2021). As network operators started to reap the benefits of 5G use cases by deploying Standalone (SA) core network, features like Radio Resource Control (RRC) Inactive state are being introduced that plays a critical role in improving both power efficiency and latency by preserving the user equipment (UE) context during periods of inactivity (Dagiuklas, 2023). This state allows for rapid transitions back to active communication, significantly reducing signaling overhead compared to traditional RRC Idle to RRC Connected transitions (Da Silva et al., 2016). As 5G networks expand and evolve, understanding and managing these transitions effectively becomes increasingly important. Machine learning offers a powerful tool for this task, enabling operators to predict state changes, such as transitions from RRC Inactive to either RRC Resume or RRC Resume Fallback (or simply referred to as RRC Fallback). By leveraging predictive models, network operators can optimize resource allocation, minimize latency, and ensure more efficient network performance, all while adapting to the dynamic behavior of users and devices. The ability to anticipate these state transitions in real-time allows for smarter network management and enhanced user experience, making machine learning a key enabler in the future of 5G and beyond.

As 5G networks introduce increasingly complex use cases and dynamic user behaviors, the demand for advanced modeling techniques to address these challenges becomes increasingly critical. Ensemble learning-based traffic classification has proven to be highly efficient in wireless networks, streamlining the selection, optimization, and

* Corresponding author. T-Mobile US Inc, 98012, Bellevue, WA, USA.
E-mail addresses: RoopeshKumar.Polaganga@Mavs.Uta.edu (R.K. Polaganga), Liang@Uta.edu (Q. Liang).

https://doi.org/10.1016/j.jnca.2025.104260
Received 1 March 2025; Received in revised form 18 May 2025; Accepted 18 June 2025
Available online 18 June 2025

deployment of machine learning models (Wang et al., 2024). This approach allows network operators to proactively tackle challenges such as resource allocation and network optimization, ensuring that 5G networks can seamlessly adapt to the evolving demands of modern applications. Building on this foundation, our work extends the efficacy of ensemble learning by applying it in a real-world 5G SA network to classify RRC Inactive state transitions—specifically distinguishing between RRC Resume and RRC Fallback requests—in its novelty. This pioneering application highlights the potential of ensemble learning in addressing previously unexplored aspects of 5G SA network behavior, providing actionable insights for network resource management. Furthermore, our work introduces a novel framework to leverage these predictions for energy optimization, enabling proactive gNB operations that minimize unnecessary paging and processing overhead. This contribution is particularly significant as energy efficiency emerges as a cornerstone for the sustainability of future generations of networks (Giordani et al., 2020). By addressing both operational efficiency and sustainability, our framework aligns with the pressing need for scalable, adaptive, and environmentally conscious network solutions.

This study is structured as follows: The remainder of Section I provides an introduces RRC Inactive state as well as the ensemble learning methods used followed by a review of related work on the 5G RRC Inactive state and energy savings. Section II introduces the proposed framework to optimize energy in 5G SA networks based on RRC State predictability. Section III delves into the characteristics of the real-world 5G NR SA network data employed in this study. In Section IV, we present the results of classification predictions of RRC States accompanied by a regression prediction of their latency impacts. Proposed energy optimized framework is simulated based on real-world network parameters to quantify savings. Finally, Section V concludes the study with key takeaways and discusses potential directions for future research.

1.1. RRC inactive state

The RRC Inactive state, introduced in 5G NR Standalone (SA) networks, is designed to reduce signaling overhead and improve network latency by keeping the user equipment (UE) context intact during periods of inactivity. This allows for rapid transitions back to active communication without the need for full reconnection procedures, making it particularly useful in scenarios involving intermittent data transfer. Fig. 1 represents the transitions between RRC states, with the blue sections denoting the current 5G NR SA standard, which includes three states: RRC Connected, RRC Inactive, and RRC Idle. The orange sections highlight the pre-Release 17 standard, where only two states: RRC Connected and RRC Idle were allowed. The introduction of the RRC Inactive state in Release 17 marks a significant improvement in signaling reduction, allowing faster transitions and more efficient handling of intermittent data transfers, which were not possible in earlier standards.

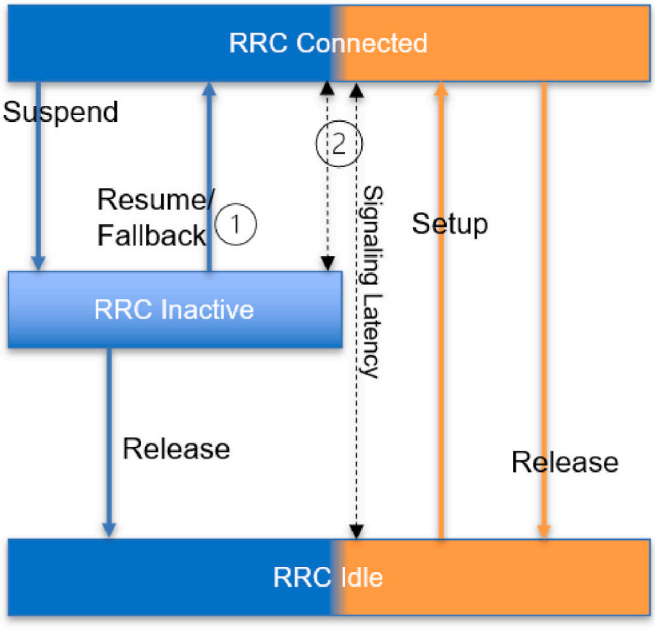


Fig. 1. RRC State Transitions and Latency illustration in 5G SA Network with RRC Inactive State.

Two key procedures emerge when the UE needs to transition from the RRC Inactive state: RRC Resume and RRC Fallback. In the RRC Resume procedure (labeled as ① in the diagram), the UE reconnects with the original gNB using the saved context, allowing for a swift transition back to the RRC Connected state with minimal latency. However, if the UE is unable to reconnect with the original gNB, the RRC Fallback procedure (also labeled ①) is triggered. This requires establishing a new connection with a different gNB, leading to more signaling overhead and higher latency.

This study employs ensemble learning that will be introduced in the subsequent sections to predict whether a UE will undergo RRC Resume or RRC Fallback, as well as the associated signaling latency (labeled as ② in the figures with dotted lines). Shorter dotted line represents the latency to transition from RRC Inactive to RRC Connected state while the longer dotted line represents the latency to transition from idle to RRC connected state. When the incoming request from UE to get from Inactive to connected state is 'Fallback' procedure, it takes a lot longer than a typical RRC Resume approach. As shown in Fig. 2, it in fact takes the same time as going from Idle to connected mode as gNB has lost its context and it must be set up again just like any other UE coming in from Idle to connected state. By accurately predicting these outcomes and quantifying the corresponding latency impacts, this work aims to optimize network energy consumption while maintaining overall user experience in 5G SA networks.

1.2. Ensemble learning

In the ever-evolving telecommunications landscape, where precision, adaptability, and performance are critical, ensemble methods stand out as highly effective tools. These methods leverage the combined strength of multiple machine learning models to boost predictive accuracy and reliability (Breiman, 1996). By aggregating diverse models such as decision trees, random forests, gradient boosting machines, and neural networks, ensemble techniques help minimize biases and errors inherent in individual models. For example, a random forest pools predictions from several decision trees, while gradient boosting

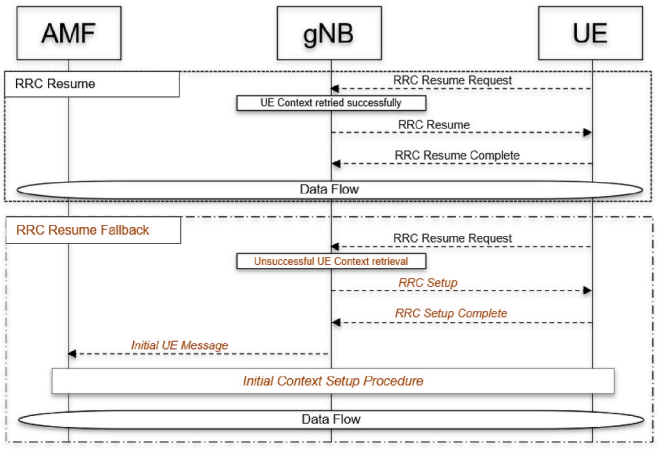


Fig. 2. Call Flow for RRC Resume and RRC Resume Fallback procedures.

sequentially refines weak models to create a strong overall predictor (Freund et al., 1996). Fig. 3 illustrates a weighted ensemble architecture, where N base models generate predictions that are then combined by an ensemble meta-model using various weight optimization strategies. These strategies may depend on each model's accuracy regarding a specific target metric or even on user-defined preferences. The collaborative power of ensemble models not only enhances performance but also offers a flexible framework applicable across numerous fields, including telecommunications, finance, and healthcare (Dietterich, 2000).

Unlike traditional ensemble models, which treat all base models equally, weighted ensembles assign varying influence on individual models, capitalizing on their strengths while minimizing their weaknesses (Mohr et al., 2018). This tailored approach not only boosts predictive accuracy but also provides deeper insights into complex network behaviors and user interactions. In scenarios with distinct data patterns or anomalies, weighted ensembles offer greater flexibility, allowing for more precise control over each model's contribution, ultimately enhancing performance and adaptability. In telco applications, where optimizing machine learning models is crucial, the weighted ensemble method stands out as a powerful and strategic solution for combining diverse models effectively (Stojanović et al., 2021).

A multi-layer stacking strategy provided by AutoGluon framework is employed in this study that combines predictions from multiple base models to maximize predictive accuracy and robustness. The process begins with data preprocessing, where the input dataset (X, Y) is structured for supervised learning. To implement k-fold bagging – a method that combines the ideas of k-fold cross-validation and bagging – the dataset is partitioned into k disjoint subsets $\{X^j, Y^j\}_{j=1}^{k}$. For each model m in the model family $\mathscr{M}$, training is conducted on the $k-1$ folds while predictions are made on the held-out fold, resulting in out-of-fold (OOF) predictions $\hat{Y}_{m,i,j}$. These OOF predictions are concatenated across all base models to form a new feature set X', which serves as input for a meta-model in the next stacking layer. This hierarchical process is repeated across L stacking layers, allowing the meta-model to optimally combine the strengths of the base models. The prediction from the weighted ensemble model is represented in equation (1), where

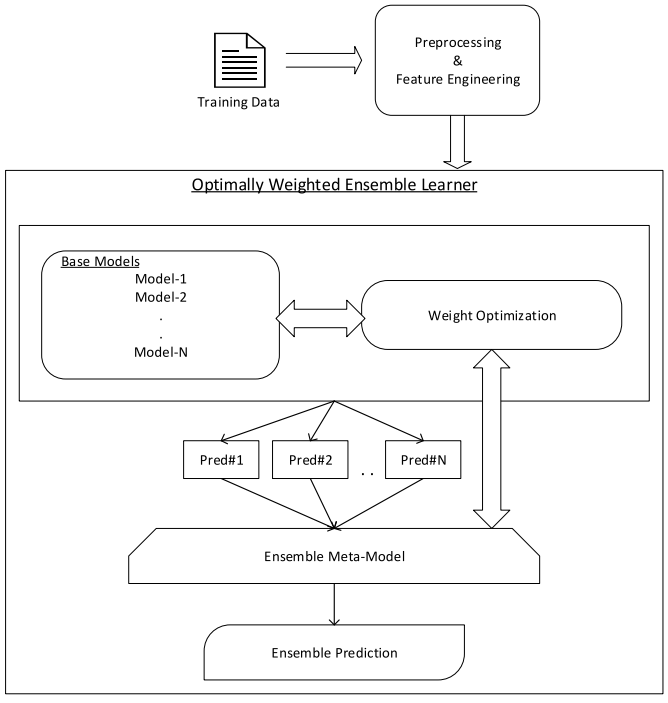


Fig. 3. Ensemble learning architecture.

$\hat{y}ensemble$ is the final prediction, N is the number of base models, w_i is the weight assigned to the ith base model, and $\hat{y}_i$ is the corresponding prediction (Song et al., 2022):

$$\hat{y}_{ensemble} = \sum_{i=1}^{N} w_i . \hat{y}_i \tag{1}$$

Using the Mean Squared Error (MSE) as the metric, weights are computed as shown below:

$$w_i = \frac{f_i(metric)}{\sum_{j=1}^{N} f_i(metric)} \tag{2}$$

where $f_i(metric)$ represents the performance of the i^{th} model on a selected metric. This approach ensures that models with lower MSE values contribute more to the ensemble prediction, enhancing overall accuracy. For instance, consider two models with MSE values of 0.05 and 0.15. The weights for these models would be calculated as follows: For the model with MSE of 0.05, the weight would be $w_1 = (1/0.05)/((1/0.05)+(1/0.15)) = 0.75$ and for the model with MSE of 0.15, the weight would be $w_2 = (1/0.15)/((1/0.05)+(1/0.15)) = 0.25$. This process ensures that the better-performing model (with lower MSE) receives a greater influence in the final prediction. While Algorithm 1 outlines this training strategy, post preprocessing the real-world data, each stacking layer is allocated a time budget, denoted as T_{total}/L, where T_{total} is the total time available for training and prediction, and L is the number of stacking layers. For this study, T_{total} is set to 48 h, although this limit does not come into play. The training time required is initially estimated and if this exceeds the time available for the current layer, the process moves to the next stacking layer. k-fold bagging approach greatly reduces prediction variance by partitioning the data into k-disjoint subsets and training copies of each model on the remaining subsets, leaving one chunk out in each iteration, referred to as cross-validated committees, useful for real-world network scenarios (Parmanto et al., 1996).

Algorithm 1. AutoGluon-Tabular Training Strategy (multi-layer stack ensembling + n- repeated k-fold bagging)

Require: data (X, Y), family of models $\mathscr{M}$, # of layers L
Preprocess data to extract features
for $l = 1$ **to** L **do** {stacking}
  for $i = 1$ **to** n **do** { n-repeated}
    Randomly split data into k chunks $\{X^j, Y^j\}_{j=1}^{k}$
    for $j = 1$ **to** k **do** { k-fold bagging}
      for each model type m in $\mathscr{M}$ **do**
        Train a type-m model on X^{-j}, Y^{-j}
        Make predictions $\hat{Y}_{m,i}^{j}$ on OOF data X^j
    end for
  end for
  end for
  Average OOF predictions $\hat{Y}_m = \left\{\frac{1}{n}\sum_i \hat{Y}_{m,i}^{j}\right\}_{j=1}^{k}$
  X ← concatenate $(X, \{\hat{Y}_m\}_{m\in\mathscr{M}})$
end for

In the context of RRC Inactive state transitions, (X, Y) represents features and labels derived from real-world 5G NR SA network data, with the ensemble framework applied to classify RRC Resume and RRC Fallback states. The diverse base models capture variability in network behaviors, while the meta-model integrates these predictions for superior classification performance. To ensure reliability, k-fold bagging is employed to reduce variance, and intermediate models are checkpointed to guarantee robustness under time constraints. This approach not only improves predictive accuracy but also dynamically optimizes gNB operations, minimizing unnecessary paging and processing, thereby achieving measurable energy savings without compromising latency requirements.

1.3. Related work

Anticipating changes in user behavior patterns, informed by early predictions of session duration, is invaluable for operators in crafting effective management strategies and mitigating operational risks. To achieve this, operators can leverage insights from historical mobile broadband (MBB) data within the telecommunications domain (Luo et al., 2016). explored regular 5G user activity predictions based on real-world MBB data, but their work focused on LTE with no extensions for the latest 5G NR SA networks or its features. More recent work by (Brezov and Burov, 2023) utilized ensemble learning, but their focus was on environmental data rather than network data (Wilhelmi et al., 2021). employed network simulators to develop machine learning-assisted 5G/6G networks, while (Upadhyay et al., 2022) leveraged real-world network data to determine whether a user was connected to a 5G network, though they did not explore session duration or state transitions. In the Radio Access Network (RAN) domain (Sun et al., 2022), applied deep reinforcement learning to deploy machine learning models, and (Wee et al., 2023) focused on predictive churn modeling in the telecom industry.

Several key contributions have emerged in studies specific to the RRC Inactive state. For instance (Da Silva et al., 2016), introduced optimizations for resource allocation and efficient state transitions (Gao et al., 2019). demonstrated efficient uplink multi-beam access for inactive UEs, enhancing beam recovery and access efficiency in mmWave networks, further improving network adaptability (Maheshwari and Gupta, 2021). focused on optimal resource allocation during RRC connection establishment, addressing challenges related to resource wastage when transitioning UEs from idle or inactive states to active, without leveraging machine learning (Khlass and Laselva, 2021). proposed novel methods for efficiently handling small data transmissions in the RRC Inactive state, while (Ryoo et al., 2018) explored the latency impacts of RRC state transition procedures without using machine learning models (Polaganga and Liang, 2024a). extended ensemble modeling to predict RRC session duration for LTE and NR users, highlighting the impact of new 5G services like Fixed Wireless Access (FWA) on the need for better data categorization to improve prediction accuracy. Recent work by (Song et al., 2022) and (Erickson et al., 2020) particularly leveraged AutoGluon's ensemble learning approach on real-world network data to improve resource allocation while demonstrating its robustness and accuracy, particularly in handling structured data. No work has been done to extend ensemble learning to the latest 3GPP Release 17-based 5G NR SA feature of RRC-Inactive state, particularly leveraging real-world network data, which is a key contribution to this study.

On the other hand, predictive modeling for energy efficiency in wireless networks primarily focus on broader frameworks and different contexts. For instance (Liu and Kung, 2023), proposed a proactive energy-saving framework for ultra-dense networks, leveraging mobility predictions to dynamically switch cells on or off based on anticipated user movement. Similarly (Fa-Long Luo, 2020), introduced machine learning-based energy optimization schemes by extracting traffic features to predict optimal network configurations. However, these works do not address the specific challenges introduced by RRC Inactive state transitions in 5G SA networks. Our study bridges this gap by targeting the optimization of RRC state transitions—specifically RRC Resume and RRC Fallback—through ensemble learning techniques. This forms the second key contribution of our work: the development of a novel energy-saving framework that leverages predictive modeling to optimize gNB operations.

1.4. Key contributions

This work significantly advances the state of the art by focusing on the 5G NR SA-specific RRC Inactive state, particularly through predictive classification of RRC Resume versus RRC Fallback transitions, an aspect not addressed in prior research. Existing studies largely overlook the granular dynamics of RRC Inactive transitions and their operational impact on gNB behavior, often treating session management as a coarse-grained duration prediction task or relying on synthetic simulations. In contrast, we utilize real-world 5G SA deployment data to train and validate our models, ensuring practical generalizability under live network conditions.

We introduce latency-aware regression modeling to quantify signaling delays specific to each transition type, addressing critical Quality of Service (QoS) considerations in real time. To operationalize these insights, we propose a novel energy optimization framework that proactively manages UE context retention based on transition predictions. Unlike prior work that optimizes either energy or latency in isolation, our approach jointly addresses both dimensions in a unified framework, ensuring a balance between energy efficiency and user experience. This framework is designed to be deployable on Release-17-compliant baseband units without additional hardware requirements. Unlike our previous work, which focused on generic session duration prediction across LTE and NR using ensemble methods (Polaganga and Liang, 2024a), this study delivers a holistic, adaptive, and field-validated solution—bridging prediction, latency modeling, and gNB-side energy optimization for live 5G SA networks.

2. Proposed framework

2.1. Proposed energy optimization framework

Effective energy management at the gNB level is essential to support the growing demands of these networks while minimizing operational costs and environmental impact. In this work, we propose an ensemble learning-driven framework that leverages predictive modeling of RRC Inactive transitions to achieve energy savings without compromising latency performance. This framework integrates seamlessly into gNB operations, providing a scalable and efficient solution for real-time energy optimization.

The proposed framework, as illustrated in Fig. 4, begins with continuous traffic monitoring within the gNB, capturing real-time UE activity and transitions between RRC states. At the core of this framework lies an ensemble-based predictive mechanism that leverages machine learning models trained on historical data (X, Y), where X represents network features and Y denotes transition labels of RRC Resume (S_r) or RRC Fallback (S_f). Real-world data used in this work is explained in greater details in subsequent sections. For each UE n, the predictive model computes a transition state $\widehat{y_n} \in \{S_r, S_f\}$ with high accuracy, enabling dynamic classification and proactive resource management. UEs predicted to fallback (S_f) trigger the deletion of their context information from the gNB's baseband and radio layers, minimizing redundant processing and paging. Conversely, the UEs predicted to resume (S_r) retain their contexts, ensuring minimal latency for incoming requests. This classification mechanism aligns with the gNB scheduler's periodic intervals, dynamically adapting to real-time traffic patterns and UE behavior.

The energy optimization process is modeled mathematically by analyzing power consumption at each time step t. The total active sessions N_t are divided into Resume $(N_r(t))$ and Fallback $(N_f(t))$ categories such that $N_t = N_r(t) + N_f(t)$. The energy consumption denoted by $E(t)$ is computed as:

$$E(t) = P_a \cdot \frac{N_r(t)}{N_t} + P_s \cdot \frac{N_f(t)}{N_t} \tag{3}$$

where P_a and P_s denote the power consumption for active and sleep states, respectively. Energy savings are realized by maximizing the fallback ratio $\frac{N_f(t)}{N_t}$ during low-traffic periods while ensuring that $P_s \ll P_a$. This formulation ensures that energy savings are maximized during low-

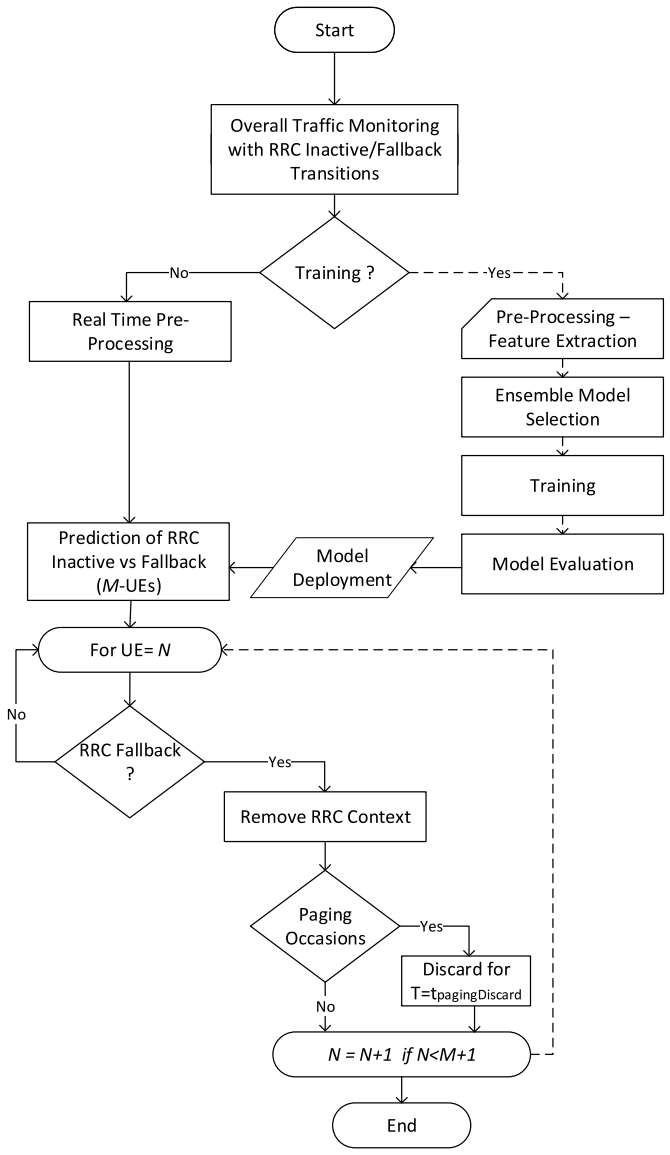


Fig. 4. Proposed energy optimization framework.

traffic periods, where fallback ratios dominate, while scaling efficiently with increased traffic. Paging occasions $(P_o(t))$ are further optimized for fallback UEs, and discarded paging attempts $(D_p(t))$ are minimized based on prediction confidence, modeled as:

$$P_o(t) = \alpha \cdot N_f(t) \text{ and } D_p(t) = \beta \cdot P_o(t) \quad (4)$$

with α and β being the scaling factors depending on traffic load and predictive accuracy. Latency performance remains a critical metric, preserved through the framework's prioritization of Resume contexts. The latency for RRC Resume (L_r) and Fallback (L_f) states is modeled using observed means from real-world data to ensure deployment feasibility. Accuracy of predictive models further ensures reliable state classification, minimizing unnecessary paging and processing while maintaining latency within acceptable bounds. By integrating predictive modeling into gNB operations, this framework delivers measurable energy savings, quantified by comparing baseline energy consumption (with no state differentiation) to predictive energy consumption (with optimized state management). This novel approach reduces gNB energy consumption during low-traffic periods while maintaining operational efficiency and user experience, demonstrating its suitability for deployment in live 5G SA networks (Fa-Long Luo, 2020).

For model training and deployment, this study employs a diverse ensemble of eight base models, encompassing ensemble methods, boosting algorithms, neural networks, and traditional regression techniques, to create a comprehensive predictive framework. These models were selected for their complementary strengths in capturing various data patterns, ensuring robustness and accuracy. Ensemble-based models such as Random Forest (RF) and Extremely Randomized Trees (XT) aggregate predictions from multiple decision trees to enhance performance and reduce overfitting (Sagi and Rokach, 2021). While RF builds decision trees with random subsets of data, XT introduces additional randomness in feature selection, further improving robustness and reducing variance (Upadhyay et al., 2022). The boosting algorithms include XGBoost (XGB), LightGBM (GBM), and CatBoost (CAT), each excelling in specific aspects. XGB offers exceptional predictive accuracy by sequentially training weak learners through gradient-based boosting (Ke et al., 2017). GBM is highly scalable and efficient, making it ideal for large datasets, while CAT stands out for its ability to handle categorical features effectively with minimal preprocessing (Dorogush et al., 2018). Neural network models, such as FastAI and NN_TORCH, leverage advanced architectures to capture complex patterns in data (Mendoza et al., 2016). FastAI simplifies deep learning workflows for rapid experimentation, while PyTorch-based NN_TORCH provides flexibility and precise control over training, making it suitable for modeling intricate relationships in tabular data (Guo and Berkhahn, 2016). This diverse model ensemble ensures a robust framework capable of addressing the complexities and variability inherent in real-world 5G NR SA network data, thereby enhancing predictive accuracy and enabling reliable network optimization. For all base models used within the ensemble, hyperparameters were tuned using grid search to optimize performance. For the final ensemble classification, AutoGluon's built-in multi-layer ensembling and hyperparameter optimization framework was employed, which automatically selected the best-performing model stack based on cross-validation performance.

To ensure real-time adaptability, the proposed framework is designed to make predictions and scheduling decisions at discrete TTI intervals, enabling dynamic updates aligned with evolving traffic conditions. Ensemble learning models, known for their robustness to data variability, are well-suited for generalizing under sudden traffic surges or rare scenarios. Furthermore, since latency is modeled and monitored alongside RRC state classification, the system can suppress context deletion or energy-saving actions during high-traffic periods where fallback transitions or paging discards might otherwise increase. These features collectively support responsive adaptation to traffic spikes without compromising performance or user experience.

3. Experimental setup

3.1. Data collection & evaluation

The datasets utilized in this study were collected from a live 5G NR Standalone (SA) network operated by a major U.S. mobile network operator over a continuous two-week period in September 2024. The data encompasses a strategically diverse selection of gNBs deployed across both urban metropolitan and rural low-density regions in various geographical areas within the United States. To mitigate potential biases—such as overrepresentation of urban base stations—careful sampling methods were employed, ensuring a balanced representation from different environmental and operational scenarios. Specifically, the selected gNB sites included deployments across low-band (600 MHz), mid-band Time Division Duplex (TDD, 2.5 GHz), and mid-band Frequency Division Duplex (FDD, 1900 MHz) spectrum bands, providing a comprehensive range of propagation conditions and network behaviors. The study combines two complementary data sources to enable robust and holistic analysis: (i) User Session Records and (ii) Network Performance Records, which were consolidated based on location and timestamp alignment.

The User Session Records dataset tracks detailed individual sessions for every User Equipment (UE) connected to the selected gNBs. Data entries were directly streamed from gNBs to storage servers, capturing session-level specifics such as UE characteristics, location coordinates, session timings, and critical radio frequency (RF) metrics like throughput, data volume, timing advance, and RF signal quality indicators for both uplink and downlink communications. Table 1 lists the 25 curated features included in this dataset. A total of approximately 100,000 user sessions were captured, ensuring validity and completeness across all essential features. To confirm the dataset's representativeness, session records included varied radio conditions, uniformly covering cell-center, cell-middle, and cell-edge scenarios as well as diverse user mobility patterns (stationary and mobile). Given that RRC Inactive is a specific 3GPP Release-17 feature, although network-enabled, compatibility was limited to recent device models. Out of the original complete session-level records, approximately 10 % were associated with RRC Inactive transitions (i.e., RRC Resume or RRC Fallback), yielding 10,000 relevant entries. This subset was filtered to ensure alignment with Release-17-compatible UEs and to guarantee data quality with valid latency and feature values. The selected volume offers a practical balance between modeling complexity and operational realism, as it reflects the actual adoption rate of RRC Inactive transitions in live 5G SA networks.

Latency values for RRC Resume and RRC Fallback transitions were derived from the filtered dataset, computed as the average time interval from the RRC Resume Request to DRB Setup Completion. These averaged latency benchmarks—15 ms for RRC Resume and 50 ms for RRC Fallback—reflect empirical observations spanning multiple days and multiple gNBs. Notably, observed Resume latencies consistently fell within the 10–20 ms range, while Fallback latencies typically ranged from 40 ms to over 100 ms, influenced by handover complexity. These selected latency benchmarks conservatively represent near-90th percentile real-world performance, ensuring realistic and practically meaningful simulation inputs without overstating potential optimization gains. Each session entry represented either pure data sessions or combined voice-and-data sessions, with data sessions potentially spanning varying channel conditions and non-guaranteed bit rate (non-GBR) channel quality indicators. Nevertheless, these variations did not materially impact the implementation or performance assessment of the RRC Inactive feature due to the consistency in RAN vendor configurations.

The second dataset, Network Performance Records, comprises aggregated Key Performance Indicators (KPIs) collected at each gNB every 15 min, irrespective of traffic levels, and stored systematically in the Operations Support System (OSS). Table 2 presents the 25 carefully selected KPI features, capturing essential network performance dimensions such as access performance, Data Radio Bearer (DRB) establishment performance, NG Context Setup effectiveness, throughput metrics, data volumes, RRC connection establishment performance, and mobility handover performance. These features provide critical context, capturing broader network-level behavior and resource utilization patterns potentially influencing signaling latency and overall network performance (Polaganga and Liang, 2024b). While User Session Records offer granular, UE-level insights, they inherently lack visibility in broader network load and resource conditions. Conversely, Network Performance Records provide comprehensive visibility into network performance trends and resource allocation but aggregate information across multiple users. Thus, integrating these datasets offers a robust analytical foundation that accounts for individual user experiences within broader network operational contexts.

To ensure data integrity and quality, thorough cleaning methods were employed, including removing incomplete session records and applying statistical methods, such as interquartile range (IQR) filtering, to detect and exclude outlier data points (Vinutha et al., 2018). These steps guarantee the reliability and accuracy of our dataset, underpinning robust, generalizable conclusions.

The signaling latency duration is calculated as the time difference between the start time (RRC Resume Request) and the end time (DRB setup with data flow), averaged across all supported UEs. To provide a contextual understanding, sessions to get from RRC inactive to RRC connected state data in comparison to regular RRC idle to connected state has relatively larger duration interval with average session duration of 192s while minimum and maximum are 0.76s and 7668s (~2.13 h) respectively. The computational infrastructure employed for executing predictions and obtaining performance results on these datasets comprises a virtual Central Processing Unit (vCPU) configuration of 64 cores, coupled with a memory allocation of 52 gigabytes. While this configuration supports high-throughput offline analysis, the trained ensemble models can be further pruned, requiring only minimal runtime resources. Inference operations for each prediction were

Table 1
List of features collected in user session records.

Features	Units
End Latitude	Degrees
End Longitude	Degrees
Model	Categorical
Make	Categorical
Service Type	Categorical
Start Time	Seconds
Duration	Seconds
Start Type	Categorical
End Time	Seconds
5G NR RSRQ	dB
5G SA Data Volume DL	Bytes
5G SA Data Volume UL	Bytes
5G SA Throughput DL	Kbps
5G SA Throughput UL	Kbps
Average Number of NR Carrier Components	#
NR RRC Establishment Cause	Categorical
NR RRC Resume Result	Categorical
NR Carrier Aggregation Service Rate	%
Maximum Number of NR Carrier Components	#
NR UE Power Headroom	dB
NR UL SINR	dB
NR TA Distance	meters
NR Start Timing Advance Cell	Categorical
NR Start Timing Advance	meters
5G NR DL SINR	dB

Table 2
List of features collected in network performance records.

Features	Units
RACH Attempts	#
RACH Failures	#
RACH Success Rate	%
RACH Successes	#
DRB Establishment Attempts	#
DRB Establishment Failures	#
DRB Establishment Success Rate	%
DRB Establishment Successes	#
Initial Context Setup Attempts	#
Initial Context Setup Failures	#
Initial Context Setup Success Rate	%
Initial Context Setup Successes	#
RRC Total Establishment Attempts	#
RRC Total Establishment Failures	#
RRC Total Establishment Success Rate	%
RRC Total Establishment Successes	#
Downlink MAC Cell Throughput	Mbps
DL MAC Data Volume	MB
UL MAC Cell Throughput	Mbps
UL MAC Data Volume	MB
UL MAC UE Throughput	Mbps
HO Exec Attempts	#
HO Exec Failures	#
HO Exec Success Rate	%
HO Exec Successes	#

observed to complete within milliseconds, making them well-suited for deployment on existing gNB baseband units, which typically have multi-core processing and dedicated machine learning acceleration capabilities. As such, the proposed framework does not necessitate any additional hardware investment and can be seamlessly integrated into current baseband architectures for real-time execution.

Evaluation metrics serve as crucial benchmarks for assessing the performance of predictive models across multiple dimensions. In this study, a total of eleven key metrics were utilized to comprehensively compare the models. Five of these metrics were applied to the classification problem, which predicts whether an incoming request is RRC Resume or RRC Fallback. These metrics include Accuracy, Precision, Recall, F1 Score, and Area Under Curve (AUC), providing a multifaceted evaluation of the classification model's ability to distinguish between the two RRC states. The remaining six metrics were employed to evaluate the numeric prediction of signaling latency in the corresponding RRC Resume and Fallback scenarios. These metrics, which include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), the coefficient of determination (R^2), SHapley Additive exPlanations (SHAP) values for interpretability, and inference latency, offer detailed insights into the accuracy and effectiveness of the models in predicting the duration of signaling latency. This robust evaluation approach ensures that both the classification accuracy and the numeric latency predictions are effectively measured, providing a holistic understanding of model performance across various dimensions (Willmott and Matsuura, 2005).

4. Results

4.1. Classification prediction of RRC states

The classification problem of RRC State prediction has been approached as outlined in earlier sections by leveraging weighted ensemble learning approach on real-world user data session records demonstrated optimal accuracy values. Table 3 presents a comprehensive evaluation of the classification model used to predict whether an incoming request will result in an RRC Resume or an RRC Fallback. The performance of the model is assessed across five key metrics: Accuracy, Precision, Recall, F1 Score, and AUC, for both RRC Resume and RRC Fallback scenarios.

The Accuracy metric shows that the ensemble learning model correctly classified 98.568 % of the requests in both cases, indicating that the model effectively differentiates between RRC Resume and RRC Fallback. This aligns with prior studies demonstrating that ensemble methods improve generalization and reduce overfitting in complex real-world datasets (Zhou, 2012). Precision, which measures the model's ability to correctly identify positive predictions (i.e., reducing false positives), stands at 99.803 % for RRC Resume and 94.309 % for RRC Fallback. This indicates that the model is highly precise for RRC Resume requests but slightly less so for RRC Fallback. The Recall metric, which captures the model's ability to identify all relevant instances (i.e., minimizing false negatives), is 98.374 % for RRC Resume and 99.283 % for RRC Fallback. This indicates that the model is very effective at identifying the majority of true RRC Resume and RRC Fallback requests. The F1 Score, which balances Precision and Recall, further emphasizes the model's robustness, with a score of 99.083 % for RRC Resume and 96.732 % for RRC Fallback, suggesting a slightly better performance for RRC Resume requests in balancing false positives and false negatives. Finally, the AUC value for both classes is 0.999, indicating an almost perfect ability to distinguish between RRC Resume and RRC Fallback classes across various classification thresholds. These findings confirm the overall strong performance of the classification model in predicting RRC state transitions with minimal errors, making it highly suitable for real-world deployment in 5G SA networks.

Table 3
Classification prediction evaluation.

Metric	RRC Resume	RRC Fallback
Accuracy	98.568 %	98.568 %
Precision	99.803 %	94.309 %
Recall	98.374 %	99.283 %
F1 Score	99.083 %	96.732 %
AUC	0.999	0.999

Feature importance for the RRC Resume vs. Fallback classification model is illustrated in Fig. 5, which ranks the top 5 features using SHAP values. The most dominant feature is 5G SA Data Volume UL bytes, contributing approximately 26 % to the overall model decision-making process. This highlights that the amount of uplink data exchanged prior to state transition is a strong discriminator of the RRC behavior. Higher uplink volumes are often correlated with active and successful session continuity, reinforcing the likelihood of a Resume event rather than a fallback. The second most important feature is Model, contributing ~19 %, indicating device-level variability in how UEs utilize RRC Inactive. This could be attributed to differences in UE firmware, modem capabilities, or even manufacturer-specific timer configurations that influence transition efficiency. The influence of 5G SA Throughput UL kbps (11.2 %) further supports the importance of uplink performance—consistent throughput often signals an ongoing data session that favors quick resume transitions. 5G NR RSRQ dB, contributing 9.7 %, captures the downlink signal quality. Poor RSRQ may result in degraded handover or resume performance, leading to increased fallback behavior. Finally, Duration, which measures the time from session initiation to transition trigger, contributes moderately. This suggests that the elapsed idle duration before resumption or fallback influences how gNB timers and UE behavior interact in triggering specific transitions.

To better understand how these top features interact, Fig. 6 presents the correlation heatmap among them. Notably, 5G SA Data Volume UL bytes shows a moderate positive correlation with Duration (0.51) and UL Throughput (0.55), implying that UEs generating higher uplink traffic tend to have longer active sessions and better throughput, both of which are hallmarks of a successful Resume path. On the other hand, Model shows minimal correlation with other features, reinforcing its role as a unique categorical contributor rather than one influenced by radio conditions or user behavior. Together, Figs. 5 and 6 provide a comprehensive view of how both individual and interrelated features drive the model's classification logic. These insights not only enhance interpretability but also inform how operators can prioritize real-time features for low-latency predictions. For instance, uplink volume, throughput, and RSRQ can be monitored continuously to preemptively estimate transition behavior, while device-specific profiling can help tailor RRC timer configurations or optimization strategies across UE categories.

Fig. 7 is an illustration of Sankey diagram representing the flow of predictions and actual outcomes for the model's classification of incoming requests into either Connection Resume or Connection Fallback. The leftmost column indicates the total number of predictions,

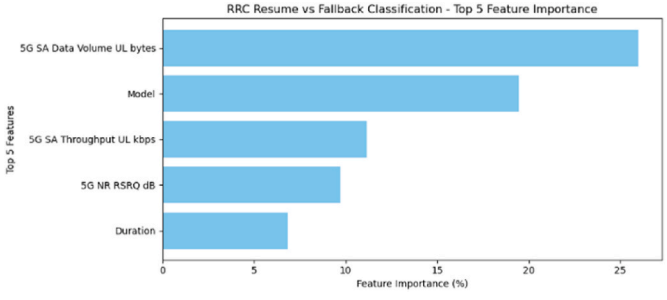


Fig. 5. Feature Importance of RRC Resume vs Fallback Classification.

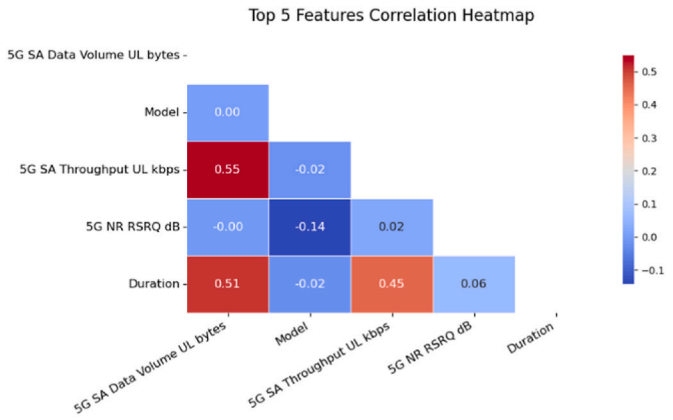


Fig. 6. Feature correlation heatmap of top 5 features.

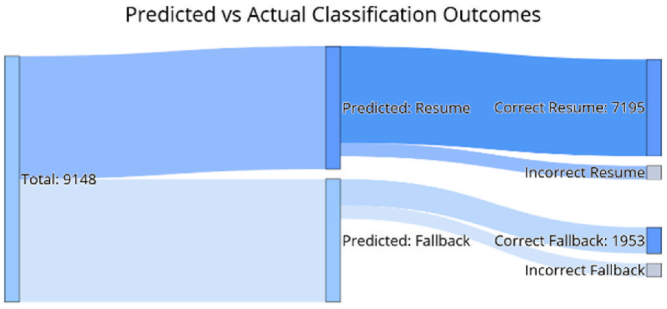


Fig. 7. Prediction Flow of RRC Resume vs. RRC Fallback Requests.

while the middle column represents the model's predicted labels (Resume or Fallback). The rightmost column reflects whether these predictions were correct or incorrect. For Connection Resume, the model is correct 99.719 % of the time, meaning that of the predicted Resume requests, almost all of them are correctly identified as Resume. Additionally, for instances where the true label in the dataset is Connection Resume, the model correctly predicted this 98.61 % of the time, highlighting the model's effectiveness in identifying Resume requests accurately. For Connection Fallback, the model shows similarly strong performance, correctly identifying Fallback requests 99.35 % of the time. The high accuracy in both categories demonstrates the reliability of the model in classifying RRC requests, with only a minimal proportion of incorrect classifications. The diagram visually emphasizes the strong correspondence between the predicted and actual labels, with most of the flow directed towards correct predictions. This reinforces the overall high accuracy and effectiveness of the classification model.

The confusion matrix shown in Fig. 8 illustrates the performance of the classification model in predicting RRC Resume versus RRC Fallback requests. The matrix is divided into four quadrants, representing the true labels versus the predicted labels. The top-left quadrant shows the true positives (TP) where the model correctly predicted RRC Resume, with 7078 instances being accurately classified. The bottom-right quadrant represents the true negatives (TN), where the model correctly predicted RRC Fallback, with 1939 instances being classified correctly. In contrast, the top-right quadrant represents the false positives (FP), where the model incorrectly predicted RRC Fallback when the true label was RRC Resume, resulting in 117 misclassifications. The bottom-left quadrant shows the false negatives (FN), where the model predicted RRC Resume instead of RRC Fallback, leading to 14 misclassifications. Overall, the model demonstrates strong performance with a high number of correct predictions (7078 TP and 1939 TN) and a relatively small number of incorrect predictions (117 FP and 14 FN). This high accuracy is visually apparent in the concentration of values in the true positive and true

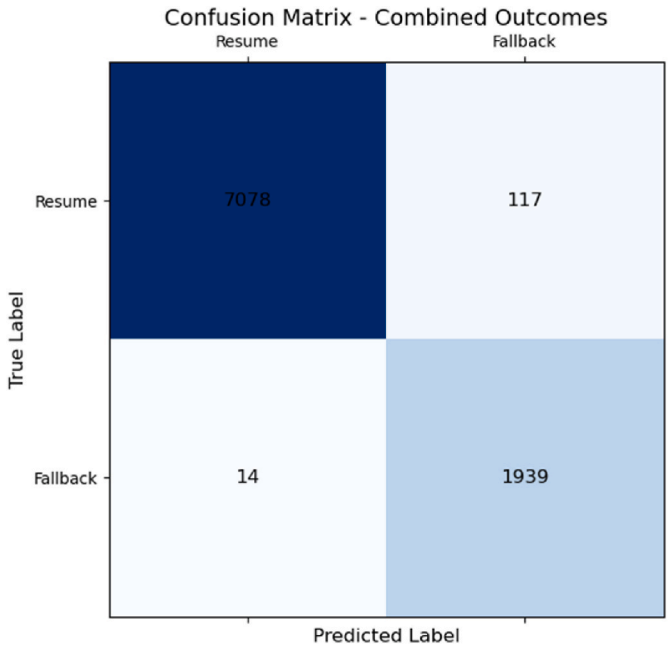


Fig. 8. Confusion matrix.

negative quadrants.

Fig. 9 presents the precision-recall curve for the classification model, illustrating the trade-off between precision and recall across different decision thresholds. The curve demonstrates that the model maintains a high level of precision even as recall approaches 1, indicating that the model can effectively identify RRC Resume and RRC Fallback requests while minimizing false positives. Area Under Precision-Recall Curve (AUPRC) is calculated to be 0.999, signifying near-perfect performance in distinguishing between the two classes. The tight curve near the top-right corner of the plot further emphasizes the model's robustness and ability to maintain high precision and recall simultaneously.

4.2. *Regression prediction of latency*

A weighted ensemble approach is implemented on Network Performance Records to predict the latency of both RRC Resume and Fallback incoming events for which the classification prediction is performed in the earlier sub-section. 6 base models were preferred for this regression namely - XGB, GBM, CAT, XT, RF, LR. Table 4 summarizes the performance metrics for the regression models predicting both RRC Resume and RRC Fallback latencies. For RRC Resume, the MAE is 1.371, indicating that, on average, the model's predictions deviate from the actual latency by 1.371 units. This metric provides a clear and interpretable

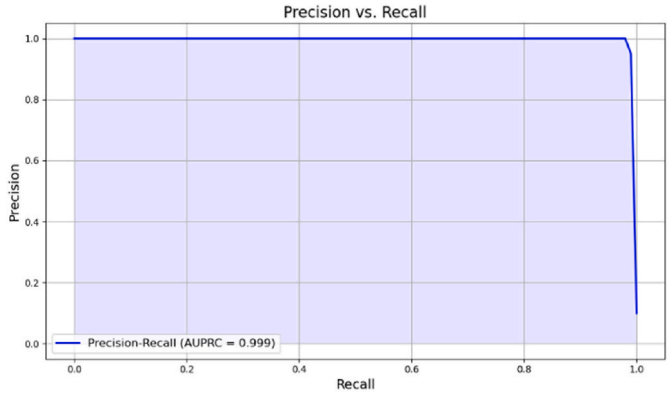


Fig. 9. Precision-Recall Curve for RRC Resume vs. RRC Fallback Classification.

Table 4
Regression prediction of RRC resume vs Fallback latency.

Metric	RRC Resume	RRC Fallback
MAE	1.371	1.420
MSE	5.844	5.923
RMSE	2.417	2.434
R^2	0.211	0.999
Inference Latency	0.240	0.192

measure of the model's accuracy by focusing on the magnitude of the prediction errors. The MSE is 5.844, which emphasizes larger errors due to its squared nature, while RMSE is 2.417, providing an interpretable error metric in the same units as the target variable. The coefficient of determination (R^2) is relatively low at 0.211, suggesting that only 21.1 % of the variance is explained by the model. This is expected given the dynamic and unpredictable nature of RRC Resume transitions, which are affected by factors such as gNB paging latency, UE inactivity duration, heterogeneous timer configurations, and session-level variability (Lundberg and Lee, 2017a). These introduce noise and non-linear interactions that limit deterministic regression performance. Nevertheless, the achieved RMSE remains well within operational tolerances for real-time gNB scheduling. The model's inference latency of 0.240 s further supports its viability for deployment. For comparison, the RRC Fallback regression model yields a significantly higher R^2, confirming the more predictable and deterministic nature of fallback behavior, and highlighting the comparative modeling complexity of RRC Resume transitions.

For RRC Fallback, the MAE is slightly higher at 1.420, reflecting a marginally larger average prediction error. The MSE for RRC Fallback is 5.923, similar to that of RRC Resume, while the RMSE is slightly higher at 2.434, again indicating the model's accuracy in terms of error magnitude. The R^2 for RRC Fallback is notably high at 0.999, signifying that the model effectively captures 99.9 % of the variability in the predicted outcomes, making it highly robust in predicting RRC Fallback latency. These findings are consistent with prior work showing ML's potential in predicting latency-critical events in RAN systems (Nikaein et al., 2015). The inference latency for the RRC Fallback model is 0.192 s, highlighting its efficient prediction speed. The overall performance metrics suggest that while the RRC Resume model leaves room for improvement in explaining variance, the RRC Fallback model is highly effective in delivering accurate predictions with minimal unexplained variance and low error rates.

The feature importance plot for the regression model predicting RRC Resume Latency is shown in Fig. 10. The horizontal bar chart highlights the top 5 features that contribute most significantly to the model's ability to predict the latency associated with an RRC Resume event. The most important feature, DRB Establishment Attempts, accounts for over 12 % of the feature importance, indicating that the number of DRB establishment attempts strongly influences the latency experienced during RRC Resume. RRC Total Establishment Attempts is the second most influential feature, contributing nearly 11 %, which suggests that the total number of establishments attempts across the network plays a critical role in latency prediction. Other significant features include HO Exec Attempts and Avg. PUCCH SINR, highlighting the impact of handover attempts and signal quality on latency. Features such as Initial Context Setup Attempts, UL MAC Cell Throughput, and HO Exec Successes also make notable contributions, suggesting that network resource allocations and successful handovers are key factors influencing latency. Additional features like SECTOR, DL MAC Cell Throughput Mbps, and Period Start Time make smaller contributions but still affect the model's predictions. Overall, this plot provides insight into which network performance metrics and conditions are most influential in predicting the latency of RRC Resume events, allowing for targeted optimizations in network management. SHAP-based feature importance enhances interpretability by quantifying the marginal contribution of each input variable (Lundberg and Lee, 2017b).

Similarly, Fig. 11 illustrates the feature importance for the regression model predicting RRC Fallback Latency. The horizontal bar plot ranks the top 5 features based on their contribution to the model's ability to predict the latency associated with an RRC Fallback event. The most dominant feature, HO Exec Successes, accounts for approximately 78 % of the feature importance, indicating that the number of successful handovers is a crucial factor influencing RRC Fallback Latency. HO Exec Attempts is the second most significant feature, contributing around 12 %, suggesting that the number of handover attempts also plays an important role in determining the latency during an RRC Fallback. The remaining features, including DRB Establishment Attempts, Initial Context Setup Successes, and RRC Total Establishment Successes, make comparatively minor contributions to the model's predictions, each contributing less than 1 %. These results highlight that successful handovers and handover attempts are by far the most critical factors affecting RRC Fallback latency, while other features such as signal quality and data throughput have a minimal influence on the regression model. This plot provides key insights into the factors driving latency in RRC Fallback scenarios, pointing to network handover dynamics as the primary determinants of latency performance in this context. The feature importance insights derived from training datasets can be strategically leveraged to streamline real-time data preprocessing, thereby reducing prediction latency and enhancing the efficiency of real-time decision-making processes.

As shown in Table 4 and Fig. 12, the RMSE for RRC Resume latency prediction is consistently higher than that of RRC Fallback. This discrepancy can be attributed to the increased variability in Resume transitions, which are influenced by a combination of UE-specific inactivity timers, network paging delays, and resume procedure handling at the gNB. In contrast, Fallback transitions are often deterministic – resulting from timer expirations or resume failures, making their latency patterns more consistent and easier to learn. The learning curve in Fig. Y

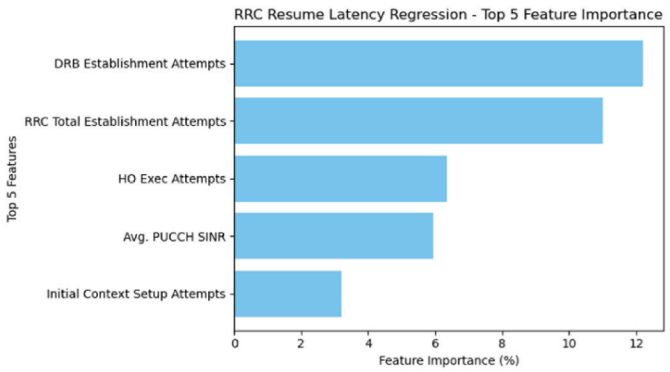


Fig. 10. Feature importance of RRC resume latency.

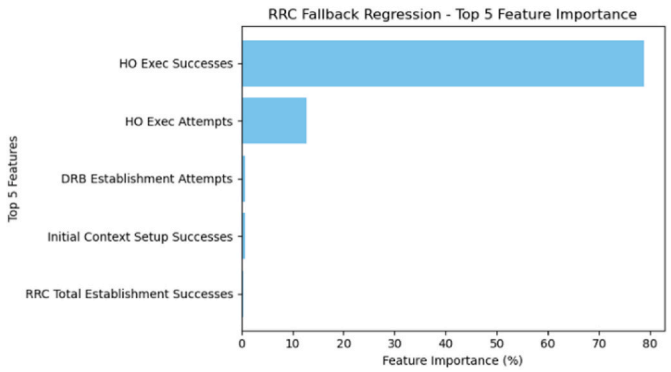


Fig. 11. Feature importance of RRC fallback latency.

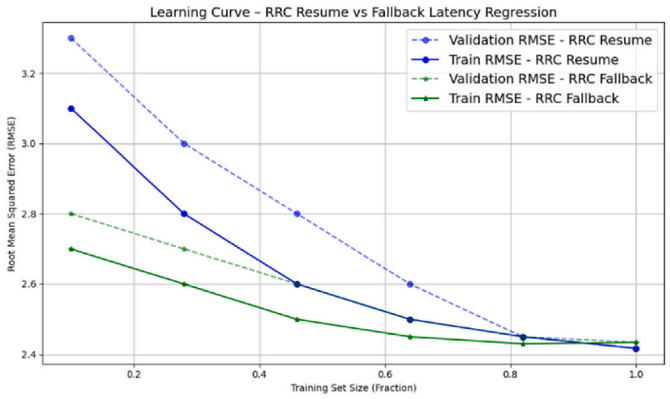


Fig. 12. Learning curve for RRC fallback latency regression.

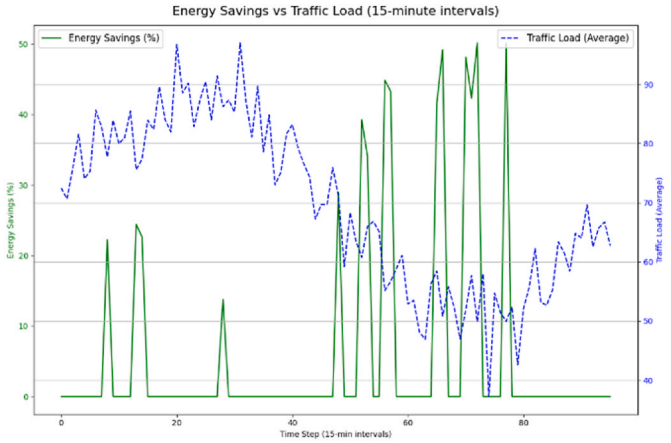


Fig. 13. Energy Optimization with traffic trend.

further reinforces this observation: the validation RMSE for RRC Resume decreases significantly with more data, indicating that the model generalizes well once sufficient training coverage is achieved. Meanwhile, RRC Fallback maintains relatively stable performance across training sizes, suggesting a simpler and more predictable latency profile.

4.3. Prediction based energy optimization

The proposed energy optimization framework, as illustrated in the flowchart of Fig. 4, is implemented under realistic constraints by simulating a single gNB scenario. The simulation is conducted over a 24-h period with 96-time steps of 15-min intervals, incorporating diurnal traffic patterns and random noise to emulate real-world network conditions. This approach captures variations in traffic load across the day, providing a comprehensive analysis of the framework's adaptability under dynamic network conditions. The framework's performance is analyzed in terms of energy savings and latency impacts, two critical metrics for ensuring its applicability in live 5G SA networks.

The power consumption assumptions used in the simulation are derived from operational benchmarks observed in live 5G networks, ensuring alignment with practical deployments. For the active state, where the gNB actively processes traffic and retains UE contexts for RRC Resume transitions, power consumption is modeled at 500 W (Liang et al., 2019). The idle state, reflecting periods of reduced activity while maintaining operational readiness, is modeled at 100 W. The sleep state, which occurs when UEs transition fully to the RRC Fallback state and their contexts are deleted, achieves maximum energy savings by minimizing baseband and radio-level activities with a power consumption of 20 W. These power states are critical to accurately evaluating the energy savings potential of the proposed framework.

The energy savings achieved through the predictive mechanism are depicted in Fig. 13, where the savings are plotted against traffic load. The results reveal that energy savings are inversely correlated with traffic loads. During periods of low traffic, the proportion of UEs predicted to transition into the fallback state increases, enabling the gNB to proactively delete unnecessary UE contexts and reduce processing overhead. This proactive management allows the gNB to transition into lower power states, such as idle or sleep, significantly reducing energy consumption. However, during high-traffic periods, the majority of UEs are in the Resume state, necessitating context retention to ensure seamless service, thereby reducing the opportunity for energy savings. At its peak, the savings reached 50 % within a 15-min reporting interval. From a baseline energy consumption of 45225.72 Wh to an optimized 33332.19 Wh on a daily basis, the predictive mechanism achieved an energy savings rate of 26.32 %, highlighting its capability to dynamically adapt to network conditions and optimize energy consumption. This ability to scale energy usage based on traffic patterns demonstrates the framework's robustness and effectiveness in real-world scenarios.

The predictive mechanism's accuracy, derived from real-world network data, averaged 98 % and was simulated with realistic variability to reflect operational conditions. This high accuracy underpins the framework's ability to reliably predict RRC transitions, ensuring the energy optimization process is both effective and robust.

Latency performance, shown in Fig. 14, is a critical metric in ensuring the feasibility of implementing RRC Inactive mechanisms. The latency comparison between baseline (without predictions) and predictive (with predictions) scenarios indicates minimal differences. The average baseline latency is 19.76 ms, while the average predictive latency is 19.64 ms, demonstrating that the predictive mechanism maintains latency performance. The latency values for RRC Resume and Fallback states were modeled with means of 15 ms and 50 ms, respectively, based on observations from live network data, ensuring realistic latency distributions in the simulation. These results validate that the predictive mechanism is not only effective in reducing energy consumption but also adheres to the stringent latency requirements critical for maintaining high-quality network performance.

To better understand the operational benefits of predictive RRC management, Table 5 compares energy savings against the corresponding predicted latency delays for different modeling strategies. The baseline method, which relies on static 3GPP timers, exhibits no energy savings and incurs higher resume delays due to frequent fallbacks. In contrast, the proposed ensemble approach achieves a 26.3 % reduction in energy consumption while maintaining an average resume latency of 15 ms, demonstrating its ability to preserve responsiveness without sacrificing efficiency. Models like Random Forest and Logistic

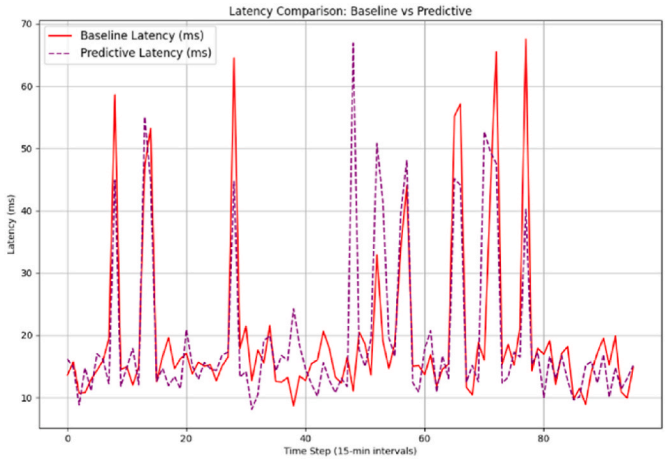


Fig. 14. Latency comparison of baseline and prediction.

Table 5
Energy savings vs Latency prediction trade-off.

Model	Energy Savings (%)	Predicted Resume Latency (ms)	Comments
Baseline	–	15	No prediction with static timers
LR	12.5 %	23.0	Weak prediction
RF	20 %	18.5	Less accurate
Proposed Ensemble	26.3 %	15	Optimal savings

Regression show moderate improvements but either underperform in latency prediction or lead to suboptimal fallback decisions. These results highlight that the ensemble model offers the best energy-latency trade-off, ensuring timely resume transitions while minimizing gNB processing overhead.

Overall, the proposed framework delivers key benefits by integrating predictive modeling into gNB operations. The ability to dynamically manage UE contexts based on predicted RRC state transitions reduces unnecessary processing and energy consumption, achieving measurable energy savings. Simultaneously, the mechanism ensures that critical latency requirements are preserved, making it a practical and scalable solution for real-world 5G SA networks. The results highlight the potential of leveraging machine learning to enhance operational efficiency and sustainability in modern cellular networks. Additionally, this framework provides a foundation for future innovations, such as incorporating advanced prediction models or integrating edge computing capabilities to further enhance network adaptability and performance.

4.4. Ablation study

To evaluate the effectiveness of the proposed ensemble learning framework, an ablation study was conducted by isolating and comparing individual base classifiers. Table 6 summarizes the performance of four baseline models—Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB), when trained independently to predict RRC Resume transitions. Among them, GB and XGB achieved the highest accuracy (91.85 % and 91.22 %) and AUC scores (0.935 and 0.934), indicating strong discriminative capabilities. RF also performed competitively with 91.68 % accuracy and an F1 Score of 64.89 %, making it a valuable contributor to ensemble diversity. In contrast, LR trailed with notably lower Recall (15.83 %) and F1 Score (25.67 %), reflecting its limited sensitivity to the minority class in this imbalanced classification scenario. Despite its weaker standalone performance, LR adds unique decision boundaries that marginally benefit the ensemble through diversity.

To further illustrate these findings, Fig. 15 visualizes the Recall and F1 Score across all models. These metrics are emphasized over Accuracy and AUC, as they better capture the ability to generalize to rare but critical outcomes such as RRC Resume transitions. The ensemble model achieves the most balanced and robust performance, with a Recall of 98.37 % and F1 Score of 99.08 %, substantially outperforming all individual models. While GB and XGB deliver strong standalone results, their integration within the ensemble architecture enhances generalization by consolidating complementary strengths and mitigating

Table 6
Ablation study - base model classification of RRC resume.

Metric	LR	RF	GB	XGB
Accuracy	87.38 %	91.68 %	91.85 %	91.22 %
Precision	67.85 %	77.45 %	82.23 %	71.42 %
Recall	15.83 %	55.83 %	52.08 %	60.41 %
F1 Score	25.67 %	64.89 %	63.77 %	65.46 %
AUC	0.828	0.931	0.935	0.934

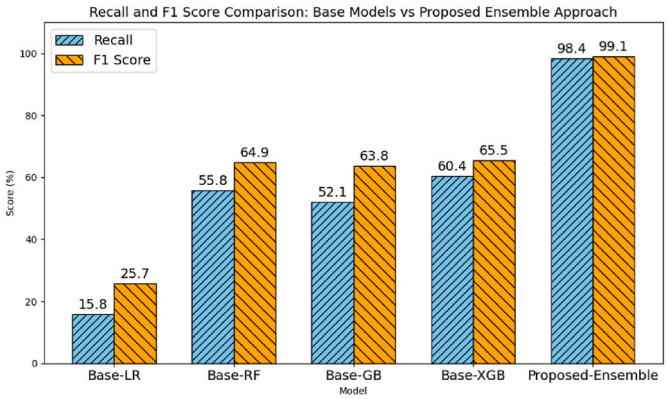


Fig. 15. Model Comparison across Base vs Proposed Ensemble Approach.

model-specific weaknesses—particularly in controlling false negatives. This superior balance is also reflected in Table 3, where the ensemble consistently leads across all evaluation metrics. Together, these quantitative and visual comparisons validate the ensemble's suitability for high-stakes, latency-sensitive classification tasks in 5G RRC state prediction.

5. Conclusion and future work

This study presents a predictive framework for optimizing energy efficiency in 5G NR SA networks by modeling RRC Inactive state transitions using real-world data and ensemble learning. Achieving a classification accuracy of 98.57 %, the system distinguishes RRC Resume from Fallback events and incorporates latency-aware regression to predict signaling delays, enabling intelligent gNB context management. The proposed approach yields 26.3 % energy savings while maintaining resume latency near 15 ms. SHAP-based feature analysis highlights the importance of uplink volume, device model, and signal quality, while an ablation study confirms the ensemble's superiority over individual models and static 3GPP timer baselines. Complementary learning curves and heatmaps offer additional transparency into model behavior. Together, these contributions support a scalable, explainable, and deployment-ready solution for energy-efficient and latency-aware RRC state management in 5G and future 6G networks.

While this study focused on ensemble-based learning using a combination of tree-based and neural network models, exploring alternative machine learning architectures remains a promising direction. Future work could investigate the applicability of transformer-based models or temporal sequence models (e.g., LSTM or GRU) to capture more nuanced temporal patterns in RRC state transitions. Additionally, lightweight models such as decision rules or quantized neural networks could be evaluated for deployment in highly resource-constrained environments, allowing further trade-offs between model complexity, interpretability, and latency. We could also extend this approach by exploring advanced reinforcement techniques to dynamically adapt network configurations based on predicted state transitions. Additionally, integrating edge computing frameworks can reduce inference latency, particularly in dense urban environments where network complexity and load are higher. Incorporating diverse data sources, such as IoT and vehicular networks, may further enhance predictive accuracy and enable broader applicability of the framework.

CRediT authorship contribution statement

Roopesh Kumar Polaganga: Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Qilian Liang:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Qilian Liang reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the U.S. National Science Foundation under Grant CCF-2219753.

Data availability

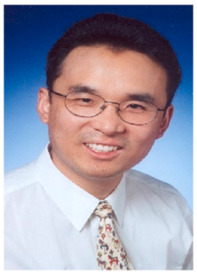
The data that has been used is confidential.

References

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.
Brezov, Danail, Burov, Angel, 2023. Ensemble learning traffic model for Sofia: a case study. Appl. Sci. https://doi.org/10.3390/app13084678.
Chen, W., et al., 2023. 5G-Advanced toward 6G: past, present, and future. IEEE J. Sel. Area. Commun. 41 (6), 1592–1619. https://doi.org/10.1109/JSAC.2023.3274037.
Da Silva, I.L., Mildh, G., Säily, M., Hailu, S., 2016. A novel state model for 5G radio access networks. In: 2016 IEEE International Conference on Communications Workshops (ICC), Kuala Lumpur, Malaysia, pp. 632–637. https://doi.org/10.1109/ICCW.2016.7503858.
Dagiuklas, T., 2023. The journey from 5G towards 6G. In: 2023 8th International Symposium on Electrical and Electronics Engineering (ISEEE), Galati, Romania, pp. 14–18. https://doi.org/10.1109/ISEEE58596.2023.10310418.
Dietterich, T.G., 2000. Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. Springer, pp. 1–15.
Dorogush, Anna, Ershov, Vasily, Gulin, Andrey, 2018. Catboost: Gradient Boosting with Categorical Features Support.
Erickson, Nick, Mueller, Jonas, Shirkov, Alexander, Zhang, Hang, Larroy, Pedro, Li, Mu, Smola, Alexander, 2020. AutoGluon-Tabular: Robust and Accurate Automl for Structured Data.
Fa-Long Luo, 2020. Machine learning in energy efficiency optimization. In: Machine Learning for Future Wireless Communications. IEEE, pp. 105–117. https://doi.org/10.1002/9781119562306.ch6.
Freund, Y., Schapire, R.E., et al., 1996. Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning, vol. 96, pp. 148–156. Citeseer.
Gao, Chong, Zhang, Hongtao, Liang, Yachao, Hao, Peng, 2019. Efficient Uplink Multi-Beam Initial Access Scheme for Inactive Users in Mmwave Networks, pp. 1–6. https://doi.org/10.1109/PIMRC.2019.8904136.
Giordani, M., Polese, M., Mezzavilla, M., Rangan, S., Zorzi, M., 2020. Toward 6G networks: use cases and technologies. IEEE Commun. Mag. 58 (3), 55–61. https://doi.org/10.1109/MCOM.001.1900411.
Guo, Cheng, Berkhahn, Felix, 2016. Entity Embeddings of Categorical Variables.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, pp. 3146–3154.
Khlass, Ahlem, Laselva, Daniela, 2021. Efficient Handling of Small Data Transmission for RRC Inactive Ues in 5G Networks, pp. 1–7. https://doi.org/10.1109/VTC2021-Spring51267.2021.9448945.
Li, X., Wang, Y., Zhang, Z., 2021. Energy-efficient resource allocation for 5G networks using deep reinforcement learning. J. Netw. Comput. Appl. 180, 102973.
Liang, Fei, Shen, Cong, Yu, Wei, 2019. Towards optimal power control via ensembling deep neural networks. IEEE Trans. Commun. https://doi.org/10.1109/TCOMM.2019.2957482, 1-1.
Liu, Y.H., Kung, B.-C., 2023. Energy saving in 5G cellular networks using machine learning based cell sleep strategy. In: 2023 24st Asia-Pacific Network Operations and Management Symposium (APNOMS). Sejong, Korea, Republic of, pp. 154–159.
Lundberg, Scott, Lee, Su-, 2017a. A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/arXiv.1705.07874.
Lundberg, S.M., Lee, S.-I., 2017b. A Unified Approach to Interpreting Model Predictions. NeurIPS, pp. 4765–4774.
Luo, Chen, Zeng, Jia, Yuan, Mingxuan, Dai, Wenyuan, Yang, Qiang, 2016. Telco user activity level prediction with massive Mobile broadband data. ACM Transactions on Intelligent Systems and Technology 7, 1–30. https://doi.org/10.1145/2856057.
Maheshwari, Isha, Gupta, Raju, 2021. Method for Optimal Resource Allocation During RRC Connection Establishment in 5GNR, pp. 42–47. https://doi.org/10.1109/ANTS52808.2021.9936954.
Mendoza, H., Klein, A., Feurer, M., Springenberg, J.T., Hutter, F., 2016. Towards automatically tuned neural networks. In: Workshop on Automatic Machine Learning, pp. 58–65.
Mohr, F., Wever, M., Hüllermeier, E., 2018. ML-Plan: automated machine learning via hierarchical planning. Mach. Learn. 107 (8), 1495–1515.
Nikaein, N., Marina, M.K., Bonnet, C., 2015. Towards a cloud-native radio access network. In: Proceedings of the ACM Workshop on Cloud-Assisted Networking (CAN '15), pp. 21–26.
Parmanto, B., Munro, P.W., Doyle, H.R., 1996. Reducing variance of committee prediction with resampling techniques. Connect. Sci. 893–4, 405–425.
Polaganga, Roopesh, Liang, Qilian, 2024a. Ensemble prediction of RRC session duration in real-world NR/LTE networks. Mach. Learn. Appl. 17, 100564. https://doi.org/10.1016/j.mlwa.2024.100564.
Polaganga, Roopesh, Liang, Qilian, 2024b. Extending causal discovery to live 5G NR network with novel proportional fair scheduler enhancements. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2024.3459798, 1-1.
Ryoo, Sunheui, Jung, Jungsoo, Ahn, RaYeon, 2018. Energy Efficiency Enhancement with RRC Connection Control for 5G New RAT, pp. 1–6. https://doi.org/10.1109/WCNC.2018.8377115.
Sagi, Omer, Rokach, Lior, 2021. Approximating XGBoost with an interpretable decision tree. Inf. Sci. 572. https://doi.org/10.1016/j.ins.2021.05.055.
Song, Yimeng, Xu, Yong, Chen, Bin, He, Qingqing, Tu, Ying, Wang, Fei, Cai, Jixuan, 2022. Dynamic Population Mapping with Autogluon, vol. 1, p. 13. https://doi.org/10.1007/s44212-022-00017-x.
Stojanović, M., Sekulović, N., Panajotović, A., Popović, P., Protić, M., 2021. Wireless channel prediction using ensemble of extreme learning machines. In: 2021 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), Sozopol, Bulgaria, pp. 167–170. https://doi.org/10.1109/ICEST52640.2021.9483464.
Sun, J., Liu, H., Chen, Q., 2022. Machine learning techniques for optimizing energy efficiency in 5G wireless networks. J. Netw. Comput. Appl. 200, 103301.
Upadhyay, Deepak, Tiwari, Pallavi, Mohd, Noor, Pant, Bhaskar, 2022. A Machine Learning Approach in 5G User Prediction. https://doi.org/10.1007/978-981-19-3571-8_59.
Vinutha, H.P., Poornima, B., Sagar, B., 2018. Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. https://doi.org/10.1007/978-981-10-7563-6_53.
Wang, X., Wei, W., Yu, X., Zheng, D., Kuma, N., Liu, L., 2024. Ensemble learning-based traffic classification with small-scale datasets for wireless networks. In: IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Vancouver, BC, Canada, pp. 1–6. https://doi.org/10.1109/INFOCOMWKSHPS61880.2024.10620836.
Wee, H Khoh, Pang, Ying, Ooi, Shih Yin, Wang, Lillian-Yee-Kiaw, Poh, Quan, 2023. Predictive churn modeling for sustainable business in the telecommunication industry: optimized weighted ensemble machine learning. Sustainability 15, 8631. https://doi.org/10.3390/su15118631.
Wilhelmi, Francesc, Carrascosa, Marc, Cano, Cristina, Jonsson, Anders, Ov, Vishnu, Bellalta, Boris, 2021. Usage of network simulators in machine-learning-assisted 5G/6G networks. IEEE Wireless Commun. 28, 160–166. https://doi.org/10.1109/MWC.001.2000206.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30, 79–82.
Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. CRC Press.

ROOPESH KUMAR POLAGANGA received his B.Tech degree in Electronics and Communication Engineering from Pondicherry Engineering College, Puducherry, India, in 2013. He earned his M.S. (Thesis) degree in Electrical Engineering from the University of Texas at Arlington (UTA), Arlington, TX, USA, in 2015, where he is currently pursuing his Ph.D. degree. Since 2015, he has been working as a Principal Systems Architect Engineer with T-Mobile US in Bellevue, WA, USA, and obtained his MBA degree from Capella University, Minneapolis, MN, USA, in 2019.

While at UTA, Mr. Polaganga served as a Graduate Research Assistant with the Communication and Networking Lab under the guidance of Dr. Liang, focusing his research on ultra-wideband and LTE technologies. At T-Mobile US Inc., he successfully designed several features and solutions in 5G-NR, LTE/LTE-Advanced, and IoT for everyday customer use. In addition to technology development, he contributed to multiple M&A projects to realize network synergies and improve overall customer experience. He has authored 5 impactful journal/conference papers and holds over 100 filed U.S. patents. Mr. Polaganga has received several corporate recognitions within T-Mobile US, including Peak nominations within the company. His technical areas of interest include wireless telecommunications, cloud networks, Internet of Things, and AI/ML in telecom networks.



QILIAN LIANG received the B.S. degree in electrical engineering from Wuhan University, Wuhan, China, in 1993, the M.S. degree in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 1996, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2000.

He was a member of Technical Staff with Hughes Network Systems Inc., San Diego, CA, USA. He is a Distinguished University Professor with the Department of Electrical Engineering, The University of Texas at Arlington (UTA), Arlington, TX, USA. He has authored or coauthored over 350 journals and conference papers, seven book chapters, and has six U.S. patents pending. His current research interests include machine learning, wireless sensor networks, wireless communications, smart grids, signal processing for communications, and fuzzy logic systems and applications. Dr. Liang was a recipient of the 2002 IEEE Transactions on Fuzzy Systems outstanding Paper Award, the 2003 U.S. Office of Naval Research Young Investigator Award, the 2005 UTA College of Engineering Outstanding Young Faculty Award, the 2007, 2009, and 2010 U.S. Air Force Summer Faculty Fellowship Program Award, the 2012 UTA College of Engineering Excellence in Research Award, and the 2013 UTA Outstanding Research Achievement Award. He was inducted into the UTA Academy of Distinguished Scholars, in 2015. He is a Fellow of the IEEE, AIIA, and AAIA.