National College of Ireland

Optimizing Cloud Power In An Open Radio Access Network Based on Subscriber Behavior

MSc Research Project
Cloud Computing

AbdulJalil Lotfi
Student ID: 22241388

School of Computing
National College of Ireland

Supervisor: Dr. Ahmed Makki

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	AbdulJalil Lotfi
Student ID:	22241388
Programme:	Cloud Computing
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Ahmed Makki
Submission Due Date:	03/01/2025
Project Title:	Optimizing Cloud Power In An Open Radio Access Network Based on Subscriber Behavior
Word Count:	4832
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	AbdulJalil Lotfi
Date:	3rd of January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing Cloud Power In An Open Radio Access Network Based on Subscriber Behavior

AbdulJalil Lotfi
22241388

Abstract

The fast adoption of 5G has significantly changed mobile communications, due to the high speed, very low latency, and the ability to serve billions of devices. However, this rapid expansion comes at the cost of increased energy consumption, caused by the need for more Radio Units (RU) deployment in high-traffic areas because of the use of higher-frequency radio waves, which offer shorter transmission ranges. This challenge has increased the computational load on cloud based Distributed units and Central units further compromising energy efficiency. Open Radio Access Network (O-RAN) architecture was a paradigm shift by introducing disaggregation, virtualization and open interfaces as those features enable more flexibility and interoperable networks. Additionally, O-RAN introduced RAN Intelligent Controllers which enables important features such as closed-loop control and AI-driven decision-making.

This thesis proposes a novel AI-driven solution to optimize energy consumption in O-RAN networks by predicting RU power states based on subscriber behavior. What makes our work different from the traditional approaches, static power management, is the use of AI to adapt in real-time to network changes and user mobility ensuring more efficient and responsive energy use in 5G networks. To achieve this goal a virtual network simulation was created and Markov process used to simulate realistic user mobility and traffic patterns across different regions and times. The generated dataset was used to train Random Forest Classifier in AWS SageMaker, resulting in predictive accuracy of 99.09%. Finally, the trained model will be deployed in a real-time cloud-based environment enabling prediction request and response through API Gateway and AWS Lambda integration.

1 Introduction

In today's modern world our dependence on internet usage has significantly increased due to the wide range of activities that can be done through the internet—from communication and social media interaction to managing IoT devices. Mobile network operators (MNOs) played a vital role in maintaining this connectivity especially after the introduction of 5G technology, with its high speed, low latency, and ability to support billions of devices, led to more reliance of MNOs. According to Chahar and Kaur (2023) the number of connected devices could reach 75 billion this year. Based on the foregoing facts, telecommunication vendors have been working hard in enhancing mobile network operators focusing on both Radio Access Network elements (RAN) and Core Network

elements (CN). However, in this research we will emphasize the RAN part because we believe in its importance and its role in connecting subscribers with the core network through radio connectivity Singh et al. (2020).

1.1 Background

RAN elements like any technology undergo several development phases due to world evolution. In this section we will go through this evolution cycle.

1.1.1 RAN Evolution

The RAN system started with its basic design while with the emergence of 4G technology, a new concept was introduced which is called distributed RAN (D-RAN) based on Alam et al. (2024). In this architecture the separation of remote radio head (RRH) and baseband unit (BBU) was applied, however both were located at the same place. After that at the late 4G era there was a move towards decoupling the RRH and BBU where the BBUs are relocated to a centralized data center (DC) to control several RRHs. This architecture was known as centralized RAN (C-RAN) based on Alam et al. (2024). While vendors weren't satisfied with this achievement and thriving towards distributed, programmable, RAN architecture. With the presence of virtualization and cloud computing concepts, 5G embraces these concepts which lead to a virtualized RAN(V-RAN) enabling the pooled BBUs of the C-RAN to be deployed as software on general purpose server not on a dedicated server as it used to be at C-RAN era according to Abubakar et al. (2023). With all this improvement, the MNOs continued to face challenges due to the lack of openness and RAN functioning as a black box according to Dryjański et al. (2021) so this encourages a group of researchers and industry leaders to establish ORAN Alliance introducing new concept known as O-RAN architecture.

1.1.2 Open RAN Concept

Open RAN concept was developed and built on four main key elements. Firstly, as shown in Figure 1 disaggregation divided the base station into different functional units resulting to have Central Unit (CU), a Distributed Unit (DU), and a Radio Unit (RU) (called O-CU, O-DU, and O- RU). This approach accepts deploying different functionalities to be deployed at different locations using different hardware Polese et al. (2022). Moving towards the most added value concept in O-RAN architecture: RAN Intelligent Controllers and Closed-Loop Control. This concept lays the foundation for Artificial Intelligence and machine learning capabilities. Moreover, this key element contains two logical controllers. The first one is Near-real-time RIC which is deployed at the edge of the network with response time range between 10ms and 1s. The second controller is called non-real-time with response time range more than 1s. The policies and functionalities of RICs are mainly dictated by applications called xApps within near-real-time RIC and rApps inside the non-real-time based on Polese et al. (2022). The third key element is the virtualization where all components of the O-RAN architecture can be delivered on cloud Polese et al. (2022) paving the way for network operators to collaborate with Cloud service providers. Finally, the last key element which complements virtualization is openness which ensures that those virtualized components can communicate smoothly regardless of the vendor who deployed the equipment.

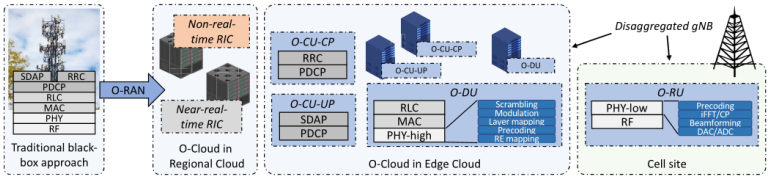

Figure 1: From Traditional RAN to ORAN Polese et al. (2022)

1.2 Motivation

One of the key features of the 5G network is its ability to operate using higher frequency radio waves compared to the previous networks. This high frequency provides advantages such as increased network capacity and reduced latency. On the other hand, this leads to have shorter wavelength of radio waves based on Cheng (2021). So, the solution was to deploy more RUs, particularly in high-traffic areas. However, this approach significantly increases the load on cloud-based CUs and DUs, where the majority of O-RAN's power consumption is already concentrated due to data processing and computing demands according to Abubakar et al. (2023). Such an approach conflicts with the mobile network operators' and the global focus on energy efficiency and sustainability.

1.3 Research Questions

This research is guided by the following question: How can subscriber behavior be utilized to predict and optimize Radio Unit power states in Open RAN networks to improve energy efficiency?

1.4 Research Objectives

1. **F**ind a solution to improve the energy efficiency of an ORAN network based on user behavior.

2. **I**mplement the proposed solution and evaluate its effectiveness.

1.5 Report Structure

1. **Related Work:** Reviews relevant work in power optimization in Open RAN networks focusing on AI/ML-based techniques, network slicing, and traffic steering.

2. **Methodology:** Explains in general the steps performed to create a virtual network environment, dataset generation, and preprocessing.

3. **Design Specification:** Details of the proposed architecture, including the client and business logic layers, virtual network simulation, dataset structure, and integration with AWS cloud services.

4. **Implementation:** The Steps under the implementation section includes dataset preparation, training of a Random Forest Classifier using AWS SageMaker, deploying of the model into an endpoint, and enabling real-time predictions through the integration of AWS lambda function and API Gateway.

5. **Evaluation:** Outline the test performed on the proposed solution and its results.

6. **Conclusion and Future Work:** Details around the directions of the work done and the future plans that can be implemented to further enhance power consumption.

2 Related Work

Numerous research studies have been conducted for telecommunication operators aiming to optimize the power usage especially at Radio side. The emergence of Open RAN standards paves the way for new opportunities in energy efficiency. As stated in our introduction O-RAN introduced more open and flexible architecture, which helps in managing power consumption effectively through RAN slicing, dynamic resource allocation and network function placement. Additionally, integration of xApps and rApps that leverage the artificial intelligence (AI) /Machine learning (ML) plays a vital role in predicting user load, mobility patterns and network load which further enhances power optimization. This section of the paper reviews state-of-the-art studies on power optimization in cloud-based Open RAN networks and the use of AI/ML algorithms for network efficiency.

2.1 AI-Driven RAN Slicing for Intelligent Power Optimization

Network slicing represents a huge leap forward with respect to Network optimization. This feature was first introduced in the 5G technology O-RAN Alliance (2021) which has gained researchers and telecommunication operators' attention due to its potential in enhancing resources efficiently. This methodology allows the creation of multiple virtual networks or "slices" on top of shared physical infrastructure to be tailored and meet specific service requirements, such as bandwidth, latency, and security, while ensuring logical separation and reliability Alam et al. (2024). The slicing concept was mainly implemented at core side due to the challenges that were under the standard RAN part which led to having most of the RAN slicing limited to research environments based on Cheng et al. (2024). However, with the emergence of O-RAN, which split the RAN components while introducing the RIC that leverage machine learning and artificial intelligence to dynamically mange optimize and orchestrate slicing concept that helped in having an E2E slicing implementation Polese et al. (2022) as we can see in our example below Figure 2. After this introduction, that covers the basic slicing concept and its importance, let us explore the work that has been done in this area by sharing some key studies and examples. To start with, this research paper Yeh et al. (2024) focuses on optimizing power usage under the 5G network that adopts an O-RAN architecture taking advantage of RIC presence to implement intelligent network slicing using a deep reinforcement learning-based framework. The proposed solution has introduced Network Slice Radio Resource Management (NSRRM) xApp supported by the near-Real-Time RAN Intelligent Controller (near-RT RIC). This xApp benefits from the real time RAN data such as current traffic load and Service Level Agreement (SLA) demands to dynamically allocate and prioritize radio resources across different network slices. Therefore, two modules were developed for this purpose: the first one was the prediction module, which uses deep learning techniques such as LSTM and TCN in forecasting traffic loads, while the second calculates resource allocation based on predicted loads and user-specific spectrum efficiency. These modules work together to dynamically adjust resource allocations at the MAC layer in a manner

that optimizes SLA compliance while minimizing power wastage.The authors who worked on Motalleb et al. (2019) proposed a method to optimize energy while maintaining the Quality-of-Service requirements across multiple services in Open RAN systems. To address this, they worked to solve the challenges of dynamic allocation and network slicing to grantee that resources Radio Units (RUs), Physical Resource Blocks (PRBs), and Virtual Network Functions (VNFs) are used efficiently by dynamically balancing user traffic across slices and allocating power resources proportionally to the traffic demand of each slice to ensure efficient energy usage. Mentioned road map was achieved by formulating the problem as a mixed-integer optimization challenge and heuristic methods were used to solve optimal power allocation across network slices while taking care of the limitations of fronthaul capacity, maximum power limits, and latency requirements. All what have been discussed is to allow the network to adapt power usage based on real-time user load, ensuring energy efficiency with QoS degradation. The work we will be discussing Nagib et al. (2023) doesn't specifically target ORAN architecture however, the approach is compatible with ORAN environment and can be applied with the RAN intelligent Controllers for AI-driven network management. The people who worked on this research proposed a reinforcement learning framework with predictive transfer learning to enhance dynamic resource allocation in RAN slicing. This is happening by reusing pre-trained RL policies from similar scenarios, which enable faster convergence and efficient adaptation to changes in user demand and SLA priorities. Which lead to optimal slice-level resource distribution while minimizing performance degradation during transitions between traffic states. The methodology mentioned indirectly improves energy efficiency by reallocating resource dynamically and allows underutilized Rus to function into low-power state.

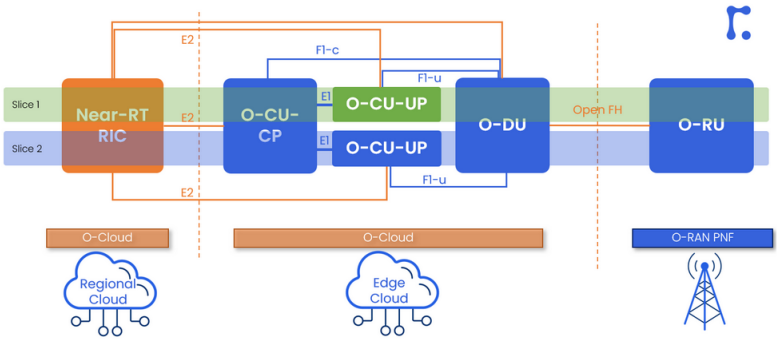

Figure 2: Example O-RAN Slicing Deployment (based on RIMEDO Labs, *Network Slicing in O-RAN*).

Table 1Provides a clearer overview between our paper and related work under section 2.1

2.2 Traffic Steering for Dynamic Resource and Power Management

Moving to a paper that shares similar concept as the one we are writing, Kavehmadavani et al. (2023) proposes an adaptive structure for optimizing resource allocation and enabling predictive management of bandwidth and power by focusing on traffic steering (TS), flow-split distribution, and radio resource management for a load balance of

Table 1: Comparison of Approaches

Aspect	Proposed Approach	Motalleb et al. (2019)	Nagib et al. (2023)	Yeh et al. (2024)
Focus	AI-driven power optimization for RU states (full, reduced, idle) based on user behavior	Joint power allocation and network slicing in an ORAN system	Accelerating RL for RAN slicing through predictive policy transfer	Intelligent and automated network slicing in RAN using deep learning and xApps for ORAN compliance
Goal	Reduce power consumption dynamically	Maximize energy efficiency and minimize power consumption and resource costs simultaneously	Improve RL convergence in dynamic RAN slicing scenarios when SLA priorities change	Automate RAN slicing with AI to ensure SLA compliance for diverse services while minimizing costs
Technique Used	Machine Learning model trained on network behavior data to predict RU states	Heuristic algorithms for optimizing resource allocation and slicing, solving a mixed-integer problem	Reinforcement Learning with Transfer Learning to accelerate policy adaptation for changing SLAs	Deep learning for traffic prediction and radio resource management implemented as xApps for near-RT RIC
Scope	Dynamic RU power management in Open RAN systems integrated with AI to optimize real-world network states	ORAN slicing for multiple services and slices with a focus on downlink performance and physical resources	Adaptive RL-based RAN slicing in 6G networks with policy reuse for SLA-driven resource management	Automated RAN slicing at the MAC layer for SLA-aware scheduling across network slices
Power Management	AI-based control for RU power states depending on user load predictions	Power allocation for RUs using optimization-based heuristics	Focused on RL policy adaptation for energy-efficient and SLA-compliant RAN slicing	Focused on SLA adherence for network slices with efficient resource planning
Traffic Handling	Utilizes user movement patterns and predicted load to allocate resources efficiently	Services are mapped to slices based on traffic requirements and physical resource constraints	RL models dynamically adjust resources to balance SLAs and traffic demands	Traffic load prediction using deep learning for dynamic slice-aware scheduling at the MAC layer
Resource Allocation	Focused on optimizing RU usage and power states using ML predictions of user behavior	Maps UEs to services, services to slices, and slices to physical data center resources	Predictive transfer learning accelerates RL for resource allocation in RAN slicing	Dynamic allocation and prioritization of radio resources to ensure SLA compliance

the QoS requirements between Enhanced Mobile Broadband (eMBB) and Ultra-Reliable Low Latency Communications (uRLLC). The Kavehmadavani et al. (2023) authors used Long Short-Term Memory (LSTM)-based approach as predictive model to forecast user traffic patterns and dynamically allocate resources to ensure optimal performance. However, our work in this paper focuses on power consumption through precise predictions of Radio Unit (RU) power states by letting the trained machine that was fed a data set which contains user traffic behavior to automatically predict the power state idle, reduced, or full power modes, this address energy efficiency directly. As a summary, while researchers of Kavehmadavani et al. (2023) deals with broader network-scale issues, such as flow management in general and QoS for both heterogeneous traffic, our work narrows the focus to power state prediction for RUs. In the realm of telecommunication, each cell might be covered by multiple Radio Access Technologies (RATs) so the author of Erdol et al. (2022) introduced a Federated Meta-Learning (FML) framework, which utilizes distributed reinforcement learning to allocate multiple Radio Access Technologies (RATs) in an intelligent way by training RL agent on tasks like latency, throughput and caching rates to quickly adapt with the user shifting behavior and traffic load. The study results show that the framework offers better caching performance and faster adaptation compared to traditional reinforcement learning techniques. Moreover, the framework supports traffic steering by dynamically allocating RATs to users based on their specific needs, to improve overall network efficiency. Additionally, the paper takes into consideration the security aspects that might happen from transferring the data to a far data center

for analysis, so they emphasize using federated learning to implement localized learning without transferring data while efficiently managing network resources. In the end optimized resource allocation reduces energy consumption contributing to power-efficient RU control in ORAN network. Adamczyk and Kliks (2021) is a similar work to the previous work where it used a SARSA-based reinforcement learning algorithm integrated with an artificial neural network (ANN) to dynamically allocate radio resources in HetNets. The model can adapt to real time user behavior and network conditions to optimize resource allocation and balance the network load. And in the end, power efficiency is achieved through optimized resource usage. The last research under this topic, we would like to highlight is Dryjański et al. (2021). This paper presents an advanced way in dynamic traffic steering in O-RAN architecture, which takes advantage of deploying the xApp on the RAN Intelligent Controller implementing policies related to spectrum management, cell assignment, and resource allocation. The deployed xApps allow real-time adaptation to user behavior, by prioritizing high-bandwidth mobile broadband (MBB) users in small cells or balancing network load based on user and network conditions. Additionally, the RIC predicts and optimizes traffic flow using AI/ML models, achieving enhanced resource utilization and improved user satisfaction. All methodologies mentioned are not directly targeting power management; however, the optimization of spectrum and resources leads to energy efficiency by reducing underutilized resources.

2.3 Subscriber Behavior Prediction for Network Resource Optimization

As the users move frequently, they undergo a process called Handover (HO). Handover is the procedure of switching an ongoing call from one cell to another within a cellular network while maintaining a seamless connectivity during the switchover Zhou and Ai (2014). Based on this the author of Makai and Varga (2023) took advantage of real-world signaling data to predict mobility management demands especially focusing on HO and Tracking Area Update (TAU) signaling traffic. After that they used this data to train several machine learning models, including linear regression, convolutional neural networks (CNNs), and long short-term memory (LSTM) networks, to predict signaling traffic volumes over 15-minute intervals. LSTM model have scored the highest prediction accuracy leading to more efficient resource allocation in core network by dynamically scaling Virtual Network Functions (VNFs) to meet traffic demands. Although the paper focused on optimizing signaling resources in the core network, we have decided to include it as it is extremely related to our work. The last paper we would like to add to this literature review is Szostak et al. (2020) where the authors propose a machine learning-based framework for short-term traffic forecasting in optical networks using the Linear Discriminant Analysis (LDA) classifier. This approach is achieved by splitting traffic predictions as a classification task and achieving up to 93% accuracy, by simplifying resource allocation to predict bitrate levels instead of exact traffic volumes. This would sound like wasting some resources —such as allocating 200 Gbps capacity for actual traffic of 155 Gbps, leaving 45 Gbps unused. This is true but wasting 45Gbps will avoid bottlenecks and is a tradeoff for simplicity and speed in prediction and resource allocation. Although this work focuses on optical system, it was added here because the methodology can be applied to cellular networks that apply ORAN architecture for its RAN part, where accurate traffic predictions could guide dynamic RU power adjustments (e.g., transitioning RUs to reduced or idle states).

3 Methodology

The proposed solution we have worked on in this paper aims to optimize power usage in an Open RAN network by predicting Radio unit states based on user behavior via a machine learning approach through several steps. Due to the private and sensitive nature of mobile operator data, publicly available datasets were not accessible, so we have created a virtual network to generate the desired dataset. However, to get dataset similar to a real-world ones we have done two main things in our code. Firstly, user movement within the network was modeled using Markov process, allowing for realistic transitions between areas based on time of day and location probabilities. Moreover, residential areas user's load was reduced during nighttime, while other areas load wasn't changed. After that the generated dataset was used for building a machine learning model and integrating it with real-time cloud-based predictions. The proposed methodology is demonstrated in below Figure 3.

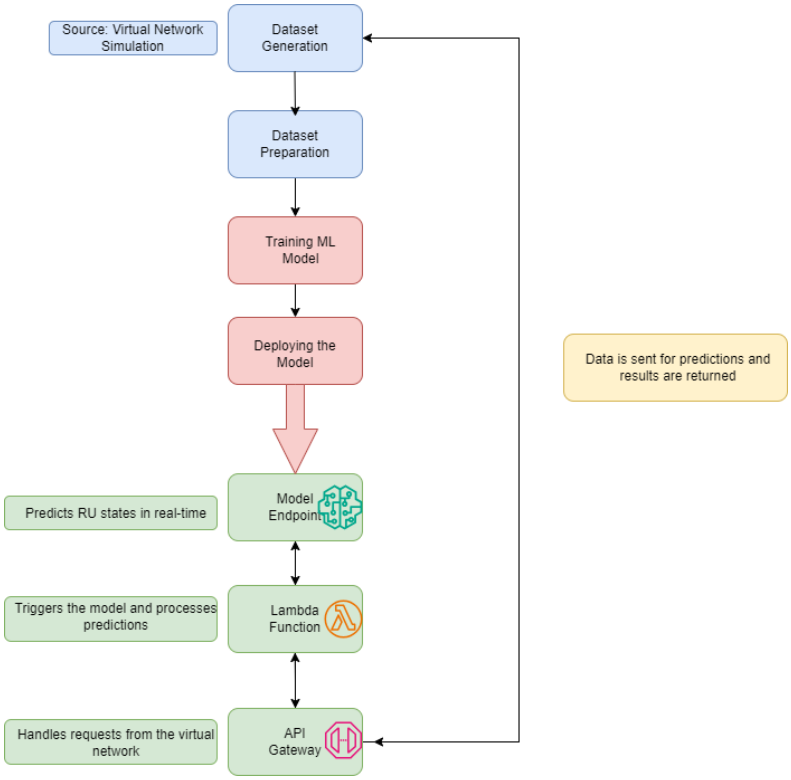

Figure 3: Methodology Workflow for Open RAN Optimization

- Dataset Generation: Virtual network was created to provide us with the needed data set presenting user and RU interactions. And the main thing Markov process was used for user movement which helps us concluding a realistic behavior.
- Dataset Preparation: The simulated data was processed to include Day Type, Hour, Area, RU type, Capacity, Current load, Load Percentage, and RU Power which were stored in a structured dataset. After that we have preprocessed it to enhance data quality through noise addition and balancing techniques.
- Training ML Model: A Random Forest Classifier was trained using AWS SageMaker in order to predict RU states (Idle, Reduced, Full) based on load and user behavior

- Deploying the Model: The trained model was deployed as an endpoint, so real-time predictions can be retrieved during network simulations.
- Lambda function: A Lambda function was created to call the ML endpoint, by sending the RU data, and receiving predictions in real time. It acts as the intermediary between the simulation and the model endpoint.
- API Gateway Integration: AWS API Gateway was configured to handle external triggers so there will be no need to include AWS security key in any call to the model endpoint. It provides an interface for the virtual network simulation to send data for prediction requests.

4 Design Specification

The proposed architecture of the Open RAN power optimization system built to predict Radio Unit states based on subscriber behavior which consists of two layers: the Client Layer and the Business Logic Layer, as illustrated in Figure 4. Under the client layer interaction between the ORAN network and prediction system is represented. This layer handles the communication of prediction requests and the delivery of results and predictions to the network. On the other hand, the core work is done under the Business logic layer where collected data are being processed, machine will be trained to have a model endpoint and finally integrate it with real-time cloud services.

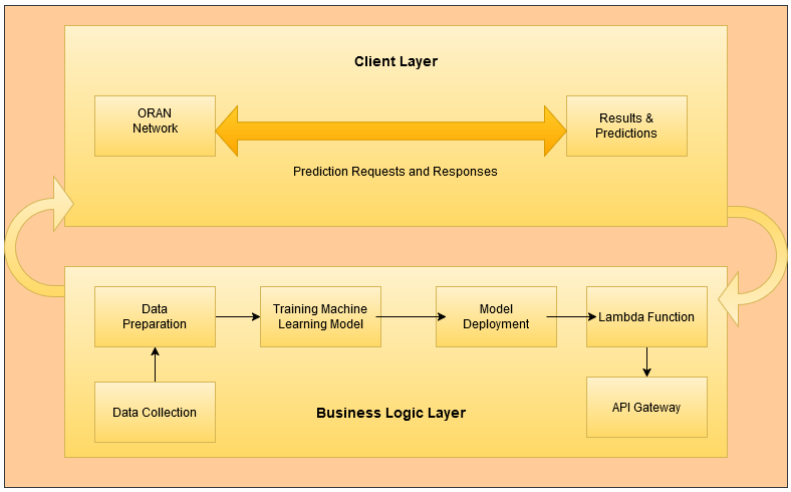

Figure 4: System Architecture: Client and Business Logic Layers for Open RAN Optimization

After collecting the structured dataset, some of the fields will be reformulated to be machine readable for example, Day Type was encoded as 0 for weekdays and 1 for weekends, Area as 0 for office areas, 1 for residential areas, and 2 for garden malls, and RU Power as 0 for Idle, 1 for Reduced, and 2 for Full. Columns that were not relevant for prediction, such as RU id, RU x, and RU y, were dropped to streamline the dataset. Then the dataset was fed to machine to be trained using Random Forest Classifier through AWS SageMaker to predict RU state. Based on IBM (n.d.) Random Forest was chosen

because it can handle both categorical and continuous data effectively. Also, it combines the output of multiple decision tree to improve accuracy and reduce the risk of overfitting. Next step was to deploy Random Forest model as an endpoint in AWS SageMaker. Using SageMaker was a plus in this project as it simplifies the process of building, training and deploying machine learning models at scale Shah (2023). The trained Random Forest Classifier achieved a testing accuracy of 99.09% proving its effectiveness in predicting RU states accurately. Furthermore, integrating AWS services was one of the features that we utilized including Amazon S3 for storing the training/testing datasets and AWS Lambda which was created to act as an intermediary role between the virtual network and the deployed model end point with the support of an API Gateway for secure and seamless communication. As API Gateway provides a public-facing interface that allows the virtual network simulation to send prediction requests securely without embedding AWS credentials directly.

5 Implementation

In Figure 5 the workflow of the ORAN power optimization system, shows the process of data collection to model predictions and resulting power states. As mentioned earlier we have worked with a dataset generation through a virtual network environment built using python code. The virtual network generates a dataset with several useful attributes such as Day Type, Current RU load, Area etc.... After that the dataset was enhanced by introducing noise to replicate a real-world condition. Gaussian noise was added to the Current load and Load Percentage attributes. Dataset enhancing has not stopped here as we have applied Synthetic Minority Over-sampling Technique (SMOTE) to make sure a balanced representation across the three classes in the dataset. After the foregoing modification, the total dataset records increased from 28080 to 31653 reflecting the addition of synthetic samples to address class imbalance. The 31653 records were split into 80% training data and 20% as test data. The implementation was made using a Python script for AWS SageMaker. Firstly, the data will be loaded from the S3 bucket. Then key features will be extracted, to identify the label column which is RU state. The Random Forest Classifier was initialized with specified hyperparameters, including n_estimators and random_state, to control the number of decision trees and reproducibility. Once training is finished with 99.09% accuracy the model was serialized and saved in SageMaker's designated model directory (model.joblib). Finally, the model was deployed as an endpoint using AWS SageMaker.
On the other hand, an AWS lambda function was created to be between the endpoint and any external request. On top of that an API Gateway was configured to expose the Lambda function as a RESTful API. This API gateway was designed to handle HTTP requests when the virtual network sends data via a POST request. Lambda function will be triggered through this request and will process the input then send it to the SageMaker endpoint for inference. Once the model process the request the results will be returned through API Gateway as a response to the client as per below Figure 7 .

6 Evaluation

The results are evaluated by efficiency of an AI-driven system to optimize energy consumption in an Open Radio network through predicting the RU power state based on

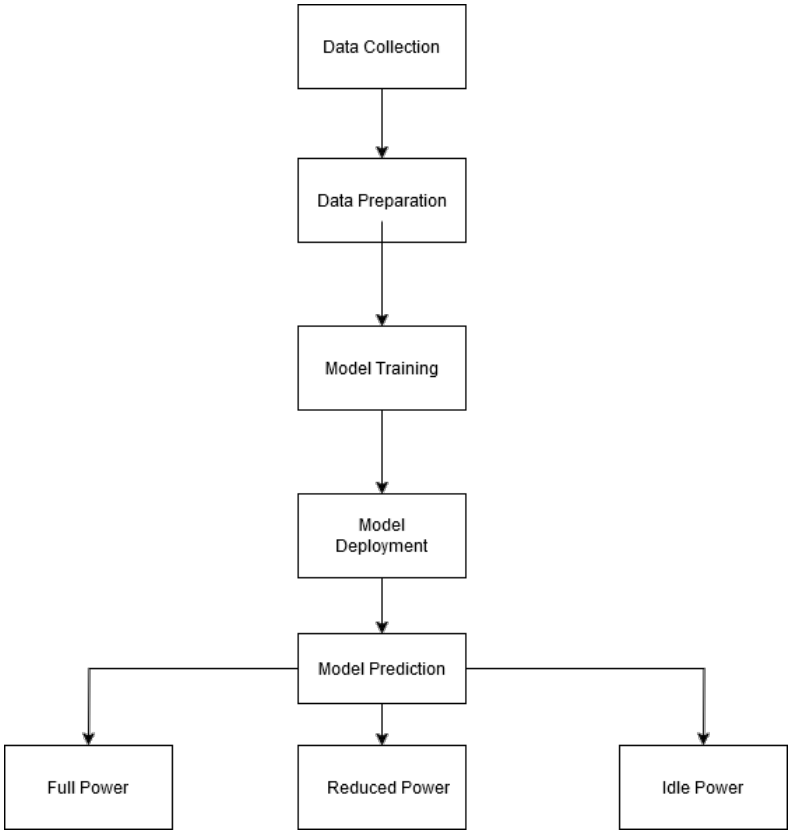

Figure 5: Workflow of the ORAN Power Optimization System

```
x['Current_load'] += np.random.normal(0, 0.5, size=len(x))
x['Load_Percentage'] += np.random.normal(0, 0.5, size=len(x))
```

```
smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)

trainX = pd.DataFrame(X_train_resampled, columns=features)
trainX[label] = y_train_resampled

testX = pd.DataFrame(X_test, columns=features)
testX[label] = y_test
```

Figure 6: Data Augmentation and Balancing using SMOTE

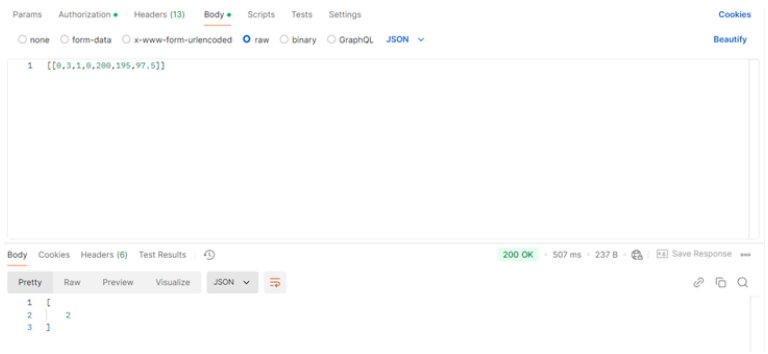

Figure 7: API Gateway POST Request and Response

user behavior. The work will be assessed from two points of view. The first concern is the performance comparison between AI-driven and non-AI systems, secondly the predictive accuracy of the machine learning model and its effect on energy efficiency.

6.1 Experiment /Evaluation of AI-Driven System and Comparison with Non-AI Baseline

The AI-driven system was developed with a Random Forest Classifier which shows excellence in the performance of an Open RAN network. The model achieved 99.09% accuracy in RU state prediction which is either Idle, Reduced or Full as detailed in Figure 8. For benchmarking purposes, the performance of a non-AI baseline system was compared where the RU states are statistically assigned. Unlikely the dynamic adaptability that the AI-driven system achieved was noticeable. This comparison underlines the fact that an AI system is more dynamic and can bring more energy efficiency.

```
Total Rows are:  5616
[TESTING] Model Accuracy is:  99.09188034188034
[TESTING] Testing Report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.99      1612
           1       0.99      0.98      0.99      2125
           2       1.00      1.00      1.00      1879

    accuracy                           0.99      5616
   macro avg       0.99      0.99      0.99      5616
weighted avg       0.99      0.99      0.99      5616
```

Figure 8: Classification Report

6.2 Discussion

Whereas the AI-based system occasionally leads in the top portions of RUs in the "full" and "reduced" power states relative to the non-AI system as shown in Figure 9 , this behavior has shown its ability to adapt to the changes in users' traffic. This is because the peak hour requires network performance from an AI system by keeping enough active RUs. On the other hand, the AI system are greatly compensated by the increased proportions of RUs in idle states during off-peak hours when the energy savings really make a difference as shown in Figure 10. This strategic optimization ensures a net reduction in energy consumption over time. Using predictive modeling, the AI system dynamically balances energy efficiency with network quality capability lacking in non-AI systems, which tend to be static and not responsive. This proves that sometimes the increase in the states "full" and "reduced" does not take away from an AI system's advantages, but rather means its strong sustainability by balance between the highest possible value of performance against energy efficiency.

6.2.1 Limitations

Although the proposed AI-driven solution for optimizing energy consumption in an Open RAN showed a significant improvement over other traditional approaches, there are specific limitations that we must highlight.

1. **R**eliance on Simulated Data
   The key limitation was the absence of real-world datasets, so we have generated the required dataset through a virtual network simulation using Markov process to replicate user behavior and network patterns. While this method provides a controlled environment for testing and ensures repeatability, it lacks the complexity and variability of real-world data. For example, in a real-world dataset usually includes unpredictable noise, different user mobility patterns which are difficult to simulate in our virtual network. In the end the simulation demonstrates its effectiveness under controlled conditions, however it might slightly differ in real-world scenarios.

2. **S**calability Challenges
   The deployed model was successfully tested/integrated with our virtual network. However, scaling the solution to large-scale, real-world deployments may introduce challenges. Computational overhead of real-time prediction is one of them, especially in areas with high traffic consumption, which could lead to system overload. Moreover, integrating our solution under different cloud platforms may require additional optimization and adaptation.

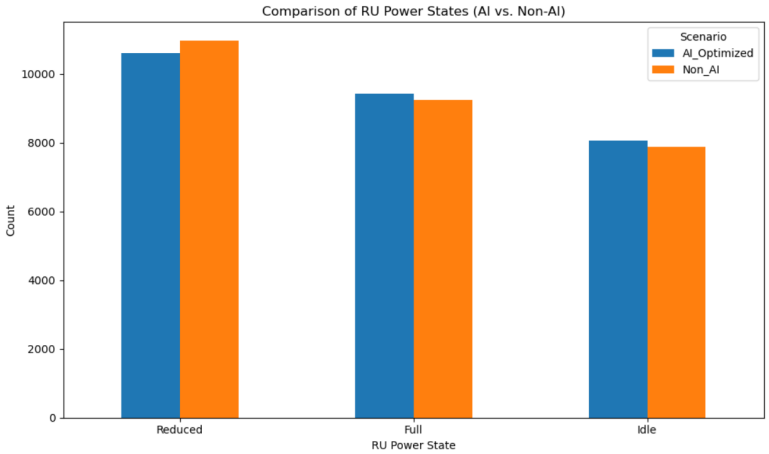

Figure 9: Comparison between an AI and non AI systems

7 Conclusion and Future Work

This research has proved its ability to optimize energy consumption in an Open Radio Access network through an AI-Driven approach by predicting Radio Unit power states based on user behavior. This AI-Driven approach achieved 99.09% prediction accuracy using Random Forest Classifier. The proposed solution showed a significant improvement over a non-AI method by dynamically transitioning RUs to idle state during the off-peak

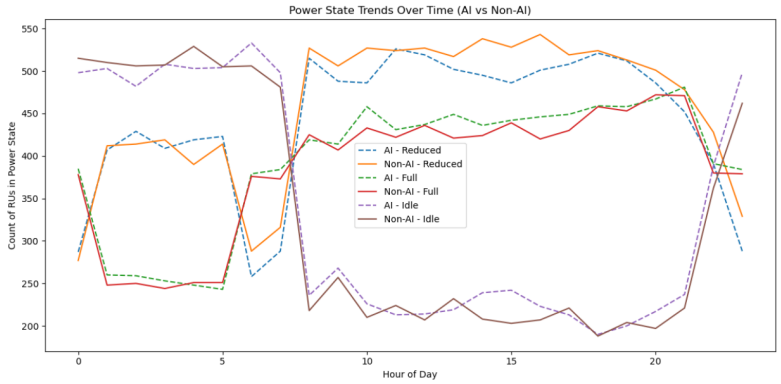

Figure 10: Comparison between an AI and non AI systems

hours while maintaining performance during rush hours, proving its ability to be used by real mobile network operators. There are several aspects that can be explored to advance this research. Firstly, developing a machine learning model that predicts user handovers would allow the system to proactively manage network resources by forecasting reduced load on specific RUs. Applying this approach the system will redirect users to the nearest active RU and dynamically adjust RU state to optimize energy efficiency. Secondly, the proposed solution could be implemented under multi-cloud or hybrid environments to test and evaluate its compatibility with different infrastructure setups by deploying the system on different cloud providers would further show its adaptability and readiness for real-world deployment scenarios. Finally, evaluating the proposed approach using real-world datasets and advanced simulation frameworks like OMNET++ would provide more realistic assessment of system's performance as it replicates real user behavior, protocols and conditions in real-world networks.

References

Abubakar, A. I., Onireti, O., Sambo, Y., Zhang, L., Ragesh, G. K. and Ali Imran, M. (2023). Energy efficiency of open radio access network: A survey, *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, pp. 1–7.

Adamczyk, C. and Kliks, A. (2021). Reinforcement learning algorithm for traffic steering in heterogeneous network, *IEEE Transactions on Wireless Communications* pp. 86–89.

Alam, K., Habibi, M. A., Tammen, M., Krummacker, D., Saad, W., Renzo, M. D., Melodia, T., Costa-Pérez, X., Debbah, M., Dutta, A. and Schotten, H. D. (2024). A comprehensive tutorial and survey of o-ran: Exploring slicing-aware architecture, deployment options, use cases, and challenges, *arXiv preprint arXiv:2405.03555* . **URL:** *https://arxiv.org/abs/2405.03555*

Chahar, S. and Kaur, K. (2023). Internet of things with 5g technology: A critical review, *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 1402–1406.

Cheng, H., D'Oro, S., Gangula, R., Velumani, S., Villa, D., Bonati, L., Polese, M., Arrobo, G., Maciocco, C. and Melodia, T. (2024). ORANSlice: An open-source 5g

network slicing platform for O-RAN, *arXiv preprint arXiv:2410.12978* .
URL: *https://arxiv.org/pdf/2410.12978*

Cheng, R. (2021). 5g: From galaxy s21 to new apps, here's what you need to know, *CNET* .
URL: *https://www.cnet.com/tech/mobile/5g-from-galaxy-s21-to-new-apps-heres-what-you-need-to-know/*

Dryjański, M., Kułacz, Ł. and Kliks, A. (2021). Toward modular and flexible open ran implementations in 6g networks: Traffic steering use case and o-ran xapps, *Sensors* **21**(24): 8173.
URL: *https://www.mdpi.com/1424-8220/21/24/8173*

Erdol, H., Wang, X., Li, P., Thomas, J. D., Piechocki, R., Oikonomou, G., Inacio, R., Ahmad, A., Briggs, K. and Kapoor, S. (2022). Federated meta-learning for traffic steering in o-ran, *IEEE Transactions on Wireless Communications* pp. 1–7.

IBM (n.d.). What is random forest?, *IBM Think Blog* .
URL: *https://www.ibm.com/think/topics/random-forest*

Kavehmadavani, F., Nguyen, V.-D., Vu, T. X. and Chatzinotas, S. (2023). Intelligent traffic steering in beyond 5g open ran based on lstm traffic prediction, *IEEE Transactions on Wireless Communications* **22**(11): 7727–7742.

Makai, L. B. and Varga, P. (2023). Predicting mobility management demands of cellular networks based on user behavior, *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–6.

Motalleb, M. K., Shah-Mansouri, V. and Naghadeh, S. N. (2019). Joint power allocation and network slicing in an open ran system, *arXiv preprint arXiv:1911.01904* .
URL: *https://arxiv.org/abs/1911.01904*

Nagib, A. M., Abou-Zeid, H. and Hassanein, H. S. (2023). Accelerating reinforcement learning via predictive policy transfer in 6g ran slicing, *IEEE Transactions on Network and Service Management* **20**(2): 1170–1183.

O-RAN Alliance (2021). O-RAN Minimum Viable Plan and Acceleration Towards Commercialization, *Technical report*, O-RAN Alliance.
URL: *https://mediastorage.o-ran.org/white-papers/O-RAN.Minimum-Viable-Plan-and-Acceleration-towards-Commercialization-white-paper-2021-06.pdf*

Polese, M., Bonati, L., D'Oro, S., Basagni, S. and Melodia, T. (2022). Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges, *IEEE Communications Surveys & Tutorials* **24**(4): 252–291.

Shah, C. (2023). Amazon sagemaker: An overview of amazon sagemaker's features, use cases, and benefits, *Medium* .
URL: *https://chirayushah7.medium.com/amazon-sagemaker-an-overview-of-amazon-sagemakers-features-use-cases-and-benefits-a4a029fc132c*

Singh, S. K., Singh, R. and Kumbhani, B. (2020). The evolution of radio access network towards open-ran: Challenges and opportunities, *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–6.

Szostak, D., Walkowiak, K. and Włodarczyk, A. (2020). Short-term traffic forecasting in optical network using linear discriminant analysis machine learning classifier, *2020 22nd International Conference on Transparent Optical Networks (ICTON)*, pp. 1–4.

Yeh, S.-P., Bhattacharya, S., Sharma, R. and Moustafa, H. (2024). Deep learning for intelligent and automated network slicing in 5g open ran (oran) deployment, *IEEE Open Journal of the Communications Society* **5**: 64–70.

Zhou, Y. and Ai, B. (2014). Handover schemes and algorithms of high-speed mobile environment: A survey, *Computer Communications* **47**: 1–15.
URL: *https://www.sciencedirect.com/science/article/pii/S0140366414001418*