AI/ML Model Training in O-RAN: Assessing Datasets for Hoarding and Choosing Strategies

Venkateswarlu Gudepu[$], Bhargav Chirumamilla[$], Venkatarami Reddy Chintapalli[†], Piero Castoldi[•],
Luca Valcarenghi[•], Koteswararao Kondepu[$]
[$]Indian Institute of Technology Dharwad, India,
[†]National Institute of Technology Calicut, Calicut, India,
[•]Scuola Superiore Sant'Anna, Pisa, Italy.
e-mail:212011003@iitdh.ac.in, venkataramireddyc@nitc.ac.in, k.kondepu@iitdh.ac.in

***Abstract*—The advent of Open Radio Access Network (O-RAN) is transforming the traditional cellular networks into flexible, interoperable, and innovative through open standard interfaces (i.e., O1, A1, E2) and RAN Intelligent Controllers (RICs) — Near-real time and Non-real time (Near- and Non-RT RIC). These RICs leverage AI/ML models for intelligent decisions, interacting with RAN components — centralized units (CUs), distributed units (DUs), and radio units (RUs). Choosing the appropriate data for AI/ML model training in O-RAN is critical, as training based on the nature of the data, whether homogeneous or heterogeneous, can significantly improve model accuracy and efficient resource utilization. This paper introduces an approach that determines the dataset homogeneity by employing the Kolmogorov–Smirnov test (KS Test) and also considers evaluating both real-time and synthetic datasets.**

***Index Terms*—Open RAN, AI/ML Model, Beyond fifth-generation (B5G) Networks, Choosing, Hoarding.**

I. Introduction

The Open Radio Access Network (O-RAN) paradigm represents a significant shift towards dynamic adaptation and optimization of RAN configurations. It introduces two RAN Intelligent Controllers (RICs) — Near-real time and Non-real time (Near- and Non-RT RIC) — that interact with 3rd Generation Partnership Project New Radio (3GPP NR) compliant base stations through the standard open interfaces — O1, A1, and E2. Both the Near- and Non-RT RIC receive telemetry and performance measurements of the RAN in order to make intelligent decisions through artificial intelligence/machine learning (AI/ML) algorithms about network, and apply new configurations to optimize radio resource management. The AI/ML models at both Near-RT and Non-RT RIC are called xApps and rApps, respectively. These RICs aim to enhance flexibility and support the diverse traffic demands of the beyond fifth-generation (B5G) network [1].

The advent of the O-RAN brings a large collection of data from the RAN components — Radio Units (RUs), Distributed Units (DUs), and Centralized Units (CUs) — inside the Non-RT RIC through open standard interfaces to train the AI/ML models (i.e., xApps or rApps) at the RICs. However, the challenge lies in selecting the most relevant and valuable data from the hugely available data for offline AI/ML model training. The process of data selection becomes critical in ensuring that AI/ML models are effectively trained and can deliver meaningful insights and solutions.

A work in [2], trains and validates AI/ML models offline in the Non-RT RIC with all the available data (i.e., hoarding strategy) and deploys its inference at the Near-RT RIC. However, the hoarding strategy can be resource (i.e., storage and vCPU) intensive and also results in longer delays, lesser improvement in model learning accuracy, longer training time, and often model overfitting. Another work in [3], investigates the importance of considering relevant data (i.e., choosing strategy) and its advantages over the hoarding strategy that only takes account of the relevant information from available data and brings higher model accuracy, efficient resource management, and even to create multiple instances of the AI/ML models depending on the nature of the dataset.

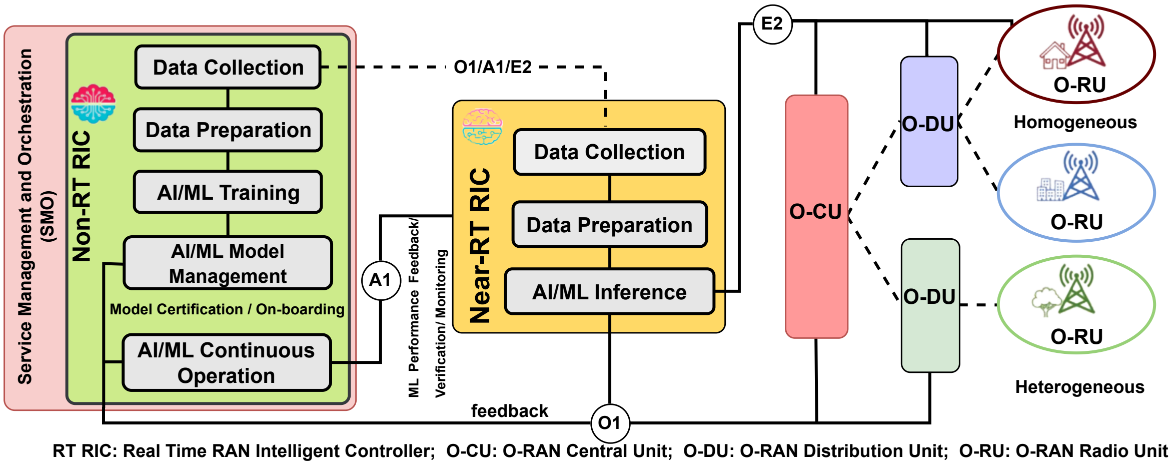
Works in [3], [4], emphasize that the hoarding strategy works well for homogeneous datasets (i.e., having similar traffic demands), whereas the choosing strategy can work for heterogeneous datasets (i.e., containing diverse traffic demands). However, approaches to determine the homogeneity or heterogeneity among the available datasets are not investigated in the context of O-RAN, which could help to decide whether to go for hoarding or choosing strategy. To fully realize the advantages of these strategies, it is essential to develop an approach for assessing dataset homogeneity or heterogeneity.

This work focuses on proposing an approach to determine the homogeneity or heterogeneity among all the available datasets in order to select either hoarding or choosing strategy to train the AI/ML models. We evaluate the proposed approach using both real-time and synthetic datasets.

II. problem statement and Syste model

The following problem statement is considered to address the dataset selection in the O-RAN context: choosing a dataset for AI/ML model training can impact the xApps/rApps preparation and their services. Thus, it is important to consider the proper dataset for model training, either hoarding or choosing based on the nature of the datasets. However, approaches to find the homogeneity or heterogeneity among the datasets are scarce in the literature in the context of O-RAN.

Figure 1 shows the O-RAN architecture that enables *intelligence* through two RICs: Non- and Near-RT, to operate at



RT RIC: Real Time RAN Intelligent Controller; O-CU: O-RAN Central Unit; O-DU: O-RAN Distribution Unit; O-RU: O-RAN Radio Unit

Figure 1: System Model

different components of the RAN with different time scales. The Non-RT RIC handles the use cases that have granularity of > 1 sec, whereas the Near-RT RIC deals with the applications that are between 10 ms and < 1 sec. The O-RAN architecture introduces interfaces such as O1, A1, and E2 to facilitate data collection and communication among the RAN components. The O1 interface manages all O-RAN components requiring orchestration and associated network functions. Additionally, the A1 interface supports policy-driven guidance and AI/ML feedback between Non-RT and Near-RT RIC, while the E2 interface controls RAN functions through E2 control messages.

The AI/ML model management block inside the Non-RT RIC is responsible for training and deploying an AI/ML model as an xApp or rApp. All collected data is stored inside the Non-RT RIC, and the AI/ML model management takes care of the AI/ML model training with the available data, which can be either homogeneous or heterogeneous traffic patterns. However, choosing the data for AI/ML model training from the Non-RT RIC database is important. The choice of data can significantly impact model generalization, accuracy, overfitting prevention, and resource utilization. To address the data selection issue, our proposed approach determines the homogeneity or heterogeneity of the datasets, which runs inside the AI/ML model management [5].

The proposed approach determines whether a dataset inside the Non-RT RIC database is homogeneous or heterogeneous in order to select either choosing or hoarding for training an AI/ML model. It employs a statistical measure as shown in Algorithm 1, and all the variables used in it are listed in Table I.

Algorithm 1 takes input parameters including $Data$, $WSize$, α, and $\#RUs$ and outputs whether the datasets are homogeneous or heterogeneous. The available data inside the Non-RT RIC database $Data$ is collected and loaded each RU data into its respective data stream (see lines 1-5). The datasets (i.e., $Data1$ and $Data2$) collected from different RUs are further divided into consecutive chunks of data with a length of $WSize$ (see lines 6-8). Each chunk of data produced at two different RUs is compared using the KS Test in order to determine whether the datasets are homogeneous or heterogeneous that are under consideration (see line 9). The KS Test calculates the difference between the empirical cumulative distribution (eCDF) functions of both datasets from RUs and represents it with a P_{value}. If the obtained P_{value} is less than α, it indicates heterogeneous traffic; otherwise, it signifies homogeneous traffic (see lines 10-13).

Table I: Description of variables used in the Algorithm.

Acronym	Referring to / Definition
$Data$	Data available inside the Non-RT RIC database.
WS	Window Size
α	Significance level.
$\#RUs$	Number of RUs.
$Data1$	Data from an RU that follows a traffic pattern 1 (for example, traffic at RU1).
$Data2$	Data from an RU that follows a traffic pattern 2 (for example, traffic at RU2).
$Window1$	Data chunks of $Data1$
$Window2$	Data chunks of $Data2$
KST	KS Test, which quantifies the distance between two distributions based on their empirical cumulative distribution functions (ECDF) and determines whether the traffic at different RUs is sampled from the same distribution or not.
P_{value}	Value determined by KS Test for each window of length $WSize$ by comparing both $Window1$ and $Window2$.

III. Experimental results

We experimentally evaluated our proposed approach using a real-time dataset from Rome [3]. This dataset includes traffic demands from specific RUs (i.e., RU_1, RU_2, RU_4, and RU_7), which exhibit similar traffic patterns (i.e., homogeneous traffic), while the remaining RUs (i.e., RU_3, RU_5, RU_6) have diverse traffic demands (i.e., heterogeneous traffic). Our approach evaluates the homogeneity of these RUs, as shown in Table II (here, ✓ indicates homogeneous and × for heterogeneous), utilizing the KS Test to measure similarity among the available datasets. Based on this evaluation, the AI/ML model management determines whether to employ the choosing or hoarding strategy for training AI/ML models.

In order to validate the outcomes of our proposed approach, we trained the AI/ML models either by choosing or hoarding

Algorithm 1: Data-set Selection for O-RAN

```
Input: $Data$, $WSize$, $\alpha$, $\#RUs$
Output: Homogeneous or Heterogeneous
while $data\_available$ do
    for $i \leftarrow 1$ to $\#RUs$ do
        for $j \leftarrow i+1$ to $\#RUs$ do
            $Data1 \leftarrow Data(RU_i)$
            $Data2 \leftarrow Data(RU_j)$
            for $n \leftarrow 0$ to $\lfloor \frac{length(data)}{WSize} \rfloor$ do
                $Window1 \leftarrow (Data1[(WSize * n)$ to $(WSize * (n+1) - 1)])$
                $Window2 \leftarrow (Data2[(WSize * n)$ to $(WSize * (n+1) - 1)])$
                $P_{value} \leftarrow KST[Window1, Window2]$
                if $P_{value} < \alpha$ then
                    $Heterogeneous$
                else
                    $Homogeneous$
```

RU_2	✓					
RU_3	×	×				
RU_4	✓	✓	×			
RU_5	×	×	×	×		
RU_6	×	×	×	×	×	
RU_7	✓	✓	×	✓	×	×
RUs	RU_1	RU_2	RU_3	RU_4	RU_5	RU_6

Table II: Data traffic similarity among the RUs

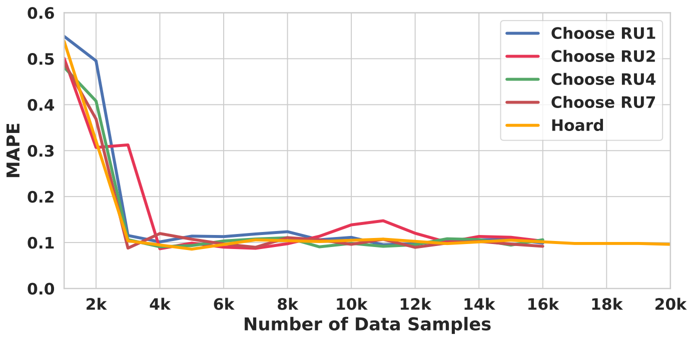


Figure 2: AI/ML model training with homogeneous datasets

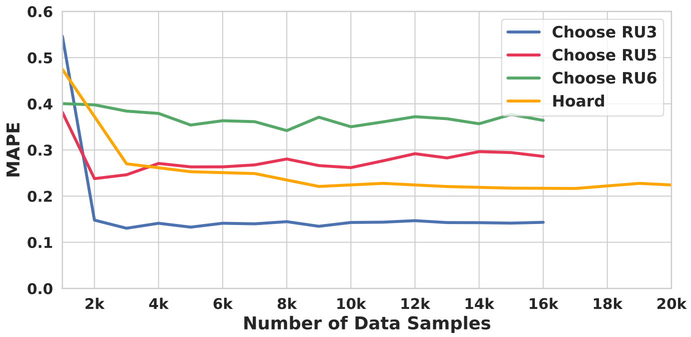


Figure 3: AI/ML model training with heterogeneous datasets

based on the RUs traffic pattern, as shown in Figure 2 and 3. We built an LSTM model with 3 hidden layers and 100 hidden units and trained it with the available RUs data. For datasets RU_1, RU_2, RU_4, and RU_7, the AI/ML model management determined that they are homogeneous based on the proposed approach and decided to go for hoarding. The Mean Absolute Percentage Error (MAPE), depicted in Figure 2, for these RUs is approximately 0.09. Notably, this MAPE remains consistent across these RUs, with a marginal reduction of only 0.004 achieved through the hoarding strategy. It can be observed that choosing individual datasets yields the MAPE closer to hoarding the datasets due to their homogeneity.

However, as shown in Figure 3, the AI/ML model management trains the RU_3, RU_5, and RU_6 datasets separately due to their heterogeneity. The best AI/ML model instance is obtained by choosing RU_3, as it yielded the lowest MAPE. Hoarding the data resulted in the lowest MAPE value of approximately 0.21. Considering the variance among the RUs' data, creating multiple model instances and deploying them as xApps or rApps, whether at the Near-RT or Non-RT RIC, becomes feasible.

IV. Conclusions and Future Work

In this paper, we present an approach to assess the homogeneity or heterogeneity of the datasets in the context of O-RAN, which could be useful in deciding whether to go for hoarding or choosing strategy for training an AI/ML model. This initial study focused on evaluating the traditional and widely adopted statistical test for a real-time and a synthetic dataset. Our potential future direction is to investigate other statistical tests, ensemble these tests, employ dynamic thresholds, utilize supervised learning techniques, and integrate with the O-RAN experimental setup.

Acknowledgment

This study is partly supported by a DST SERB Startup Research Grant (SRG-2021-001522). The European Union partially supported this work under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program"RESTART") and by the KDT-JU project Collaborative edge-cLoud continuum and Embedded AI for a Visionary industry of thE futuRe (CLEVER) (grant agreement no. 101097560). KDT-JU receives funding from the Horizon Europe Research Framework and the National Authorities.

References

[1] E. Moro, M. Polese *et al.*, "An open ran framework for the dynamic control of 5g service level agreements," *arXiv preprint arXiv:2309.07508*, 2023.

[2] L. Bonati, D'Oro *et al.*, "Intelligence and learning in o-ran for data-driven nextg cellular networks," *IEEE Communications Magazine*, vol. 59, no. 10, pp. 21–27, 2021.

[3] J. Martín-Pérez, Molner *et al.*, "Choose, not hoard: Information-to-model matching for artificial intelligence in o-ran," *IEEE Communications Magazine*, vol. 61, no. 4, pp. 58–63, 2022.

[4] X. De Luna and K. Skouras, "Choosing a model selection strategy," *Scandinavian Journal of Statistics*, vol. 30, no. 1, pp. 113–128, 2003.

[5] V. Gudepu, V. R. Chintapalli *et al.*, "Adaptive retraining of ai/ml model for beyond 5g networks: A predictive approach," *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, pp. 282–286, 2023.