Truong, Linh; Nhu Trang, Nguyen Ngoc

HAIVAN: a Holistic ML Analytics Infrastructure for a Variety of Radio Access Networks

Published in:
2022 IEEE International Conference on Big Data (IEEE BigData 2022)

DOI:
10.1109/BigData55660.2022.10020515

Published: 01/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

HAIVAN: a Holistic ML Analytics Infrastructure for a Variety of Radio Access Networks

Hong-Linh Truong
Aalto University, Finland
linh.truong@aalto.fi

Nguyen Ngoc Nhu Trang
Central MobiFone Network Centre, MobiFone Corporation, Vietnam
trang.ngocnhu@mobifone.vn

***Abstract*—This paper presents our approach for supporting machine learning (ML)-based analytics of quality of experience (QoE) related issues in a variety of Radio Access Networks (V-RAN). We focus on key problems in a holistic analytics infrastructure for engineers without strong ML skills and powerful computing infrastructures. We characterize types of relevant data and existing data systems to follow a specific data mesh approach suitable for engineers. The paper presents key steps in establishing the participation of engineers and the acquisition of domain knowledge. We introduce models for representing analytics subjects and their dependencies, and for managing relevant ML techniques and methods for analytics subjects. We explain our work through examples from a large-scale mobile network of approximately 4 million subscribers.**

I. Introduction

In practice, a mobile network provider must manage a variety of network infrastructures and equipment. Specially, the mobile network provider in our study must manage different types of Radio Access Networks (RANs), including 5G, 4G, 3G and even 2G. Furthermore, network equipment comes from different vendors and accordingly various systems are used to capture relevant data for managing RANs. Figure 1 outlines such a mobile network, called V-RAN in our work, which is quite common in many countries, such as in Vietnam, where several generations of network technologies are operated and maintained, while equipment and infrastructures are built from different vendors. Therefore, unlike silo machine learning (ML)-based analytics (even with very powerful ML ones) for a specific type of data [1], [2], our engineers must carry out different types of ML analytics using different methods for different types of data in a suitable context.

Specifically, the management of V-RAN requires us to use many techniques and algorithms for the ML-based analytics of the real network data. However, what we have found is that, to facilitate ML analytics, we need to prepare a holistic analytics infrastructure with various domain data and knowledge, together with manageable ML techniques and tools that would empower RAN engineers to quickly select and invoke suitable ML analytics for maintaining and improving customer quality of experience (QoE). The key challenge is that the engineers in our work *focus* only on specific ML analytics related to QoE issues in the network operations and, due to real-world, specific company conditions, engineers *do not possess* a deep skill in ML (mainly carrying out applied ML) as well as a powerful computing infrastructure for ML analytics, although our V-RAN is big and complex and serves a very large number of subscribers. Given existing algorithms and techniques introduced for different types of data collected from mobile networks, such as network measurements, customer feedback, and alarms, one can assume that these algorithms would be enough (such as, highlighted in the big data platform with Apache Hadoop, Hive, and SQL Spark for churn prediction in [3] or unified data model using random matrix [4]). However, with a variety of characteristics in V-RAN, we have different sources of data that cannot be easily integrated into a single system, while merged data through data integration cannot be analyzed by a single technique. Big data platforms, including datalake and data mesh technologies, help to speed up the aggregation of required data for ML analytics. Still, finding and reasoning root causes and relevant sources of problems requires us to incorporate different domain knowledge, algorithms, and techniques with different data in suitable analytics.

Our focus is to devise a holistic analytics infrastructure that provides required information and models for supporting ML analytics of V-RAN, given specific constraints on engineer teams and computing resources, in parallel with the testing and development of suitable ML analytics. This work presents HAIVAN – a holistic ML analytics infrastructure for supporting the analytics of root causes by using dependencies among analytics subjects to abstract different types of relevant entities in V-RAN. We will illustrate our infrastructure using data and examples in one V-RAN for approximately 4 million subscribers in the central part of Vietnam.

In the following, Section II characterizes relevant data. Section III presents key elements of our framework. Related work is discussed in Section IV. We conclude the paper and outline our future work in Section V.

II. Data for QoE Analytics in V-RAN

Deployment data: includes identification, geographical locations and configuration parameters of sites/cells, and operational and business contexts configuration parameters.

Network measurements: as partially shown in Figure 1, are collected from multiple Network Management Solutions (NMS).We have different types of network measurements data, including (i) *network key performance indicators (KPIs)*, such as setup success rate, drop rate, successful handover rate, and availability, (ii) *usage statistics*, such as number of connected users, uplink and downlink traffic, utilization of radio link

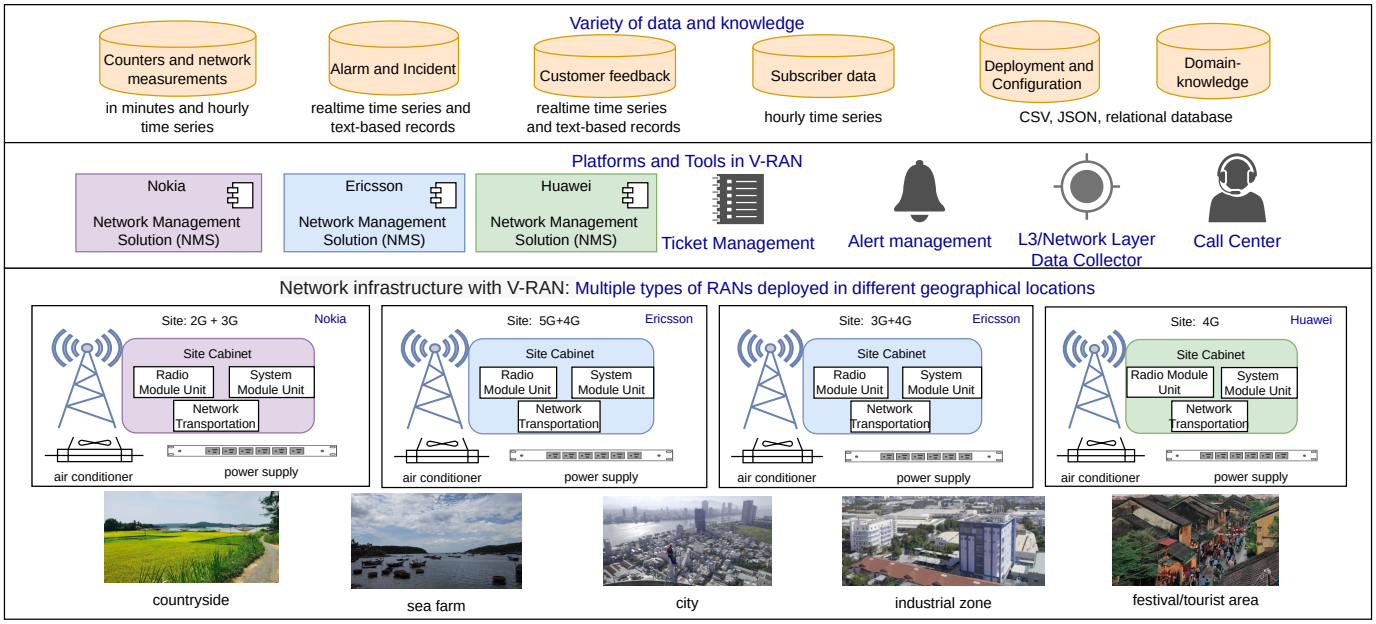

Fig. 1. A high-level view of a V-RAN – a variety of Radio Access Networks – and its platforms, tools, data, and domain knowledge

via Physical Resource Block (PRB), or transmission link, (iii) *network key quality indicators (KQIs)*, such as web response success rate, web download throughput, and video streaming success rate, and (iv) *subscribers*, such as the number of subscribers in each cell/site via Visitor Location Register (VLR), information about the type of user equipment.

Alarm data: is collected for network sites and cells. In a V-RAN, alarms about the network infrastructures are monitored by *different* alarm monitoring tools from different vendors. This leads to a variety of data formats for alarms, each is associated with a specific hardware vendor. For example, in our case, we have formats from Nokia, Ericsson, and Huawei.

Service incidents: capture various types of unplanned interruptions or deterioration of service quality that affect QoE.

Configuration changes: include a variety of changes made by operators (e.g. parameter changing, software upgrading, and hardware replacement). Configuration changes possibly create quality impacts on network resources and services.

Feedback data: reports customer experience that provide time and geographical contexts reflecting when and where the customer has reported the experience about services.

III. Holistic ML Analytics Infrastructure

A. Understanding constraints on holistic ML analytics

The specific characteristics of our V-RAN present a major problem affecting the way to carry out ML analytics. The engineers must leverage existing data sources and systems and will focus on new ML analytics in their scope. It is not a goal to establish a holistic infrastructure for all, which is not suitable in the current conditions of people, software and computing resources (*lack of* ML engineering competences and access to powerful computing resources).

A holistic approach will allow engineers to build a unified view on how to extract and obtain different types of data collected and stored in different systems for ML analytics of QoE. The holistic approach should also tackle the data quality problems and impacts spread in different components and data sources. The quality of data is different w.r.t. the completeness, accuracy, currency and granularity. It is easy to get wrong ML analytic results if the quality from different data types used for the same analytics cannot be synced. Finally, it is crucial to enable the participation of engineers at different levels to support QoE analytics where domain understanding is important. A holistic infrastructure must invoke these engineers to solve QoE issues.

B. Customizing data mesh approach for ML analytics

As shown in Figure 1, we have distributed data sources from different systems. These systems already support certain types of monitoring and analytics, but not ML ones. Our holistic infrastructure will focus only on pipelines and data components for preparing data suitable for new ML analytics based on a combination of different data and long historical data that cannot be supported by these existing systems. One possible technique to aggregate data is to define the data lake [5] and data ingestion pipelines. The ML analytics infrastructure is designed for one V-RAN deployment in the central provinces of Vietnam. But there exist similar V-RANs in other regions. All of studied V-RANs are belong to the same network provider but their analytics for operations are *separated* due to the business structure. Therefore, we utilize the data lake features only for providing and managing certain type of data products for selected ML analytics based on the need of individual teams in a specific V-RAN. The data products are centered around suitable ML analytics of the data in the datalake invoked by different workflows.

Figure 2 shows our data mesh layer, which includes various profiles (deployment information for correlations among

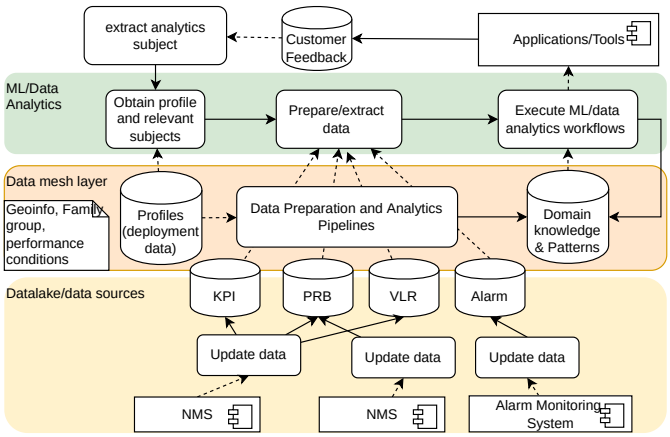

Fig. 2. The data mesh approach for correlating and analyzing data

systems), domain knowledge/patterns (cross system analytics results), and different data preparation and analytics pipelines. The pipelines extract and manage relevant data in the datalake, of which the data sources are updated from existing systems. The underlying data sources and data lake use the state-of-the-art, based on Apache Spark, Apache Hudi and other technologies, whereas data pipelines are based on Apache Airflow, Apache Spark, Pandas and other well-known technologies.

C. Identifying profiles of engineers carrying out ML analytics

The involvement of engineers is crucial due to the specific needs of V-RAN w.r.t. analytics customization and domain knowledge as inputs for improving analytics. Depending on the analytics needed by the engineers, there are many types of data (data from operation systems, human-made data, etc.) with different formats, storage times and granularity (minutes, hours, days, etc.). Such a variety of data and needs require different data analytics workflows. Second, different engineers focus on different types of analytics. Each type of analytics serves different decision-making for different requirements. For example, engineers as managers need overall analysis results (e.g., forecasting traffic/disruption trends) to be able to make decisions and strategies, while site engineers need analytic results at the site level (e.g., classification of causes influencing the operation of the site), and troubleshooting engineers need real-time and highly accurate analytic results to promptly handle errors and quickly restore services. When running basic analytics, the engineers often have to utilize domain knowledge to improve the accuracy of analytics.

To identify and build profiles of engineers for HAIVAN, first, we categorize engineers into a set of profiles based on the current work organization. Second, we carry out profile analysis (training, requirement engineering and survey). For each profile, we perform analysis of the profile to identify their requirements, scope of data analytics, and possible domain knowledge that the operators can contribute. Table I presents an example of our analysis for different profiles.

D. Identifying and incorporating domain knowledge

Data analytics methods, including ML ones, must take into account of domain knowledge to provide accurate and efficient analytics based on context-specific information of the network deployment and business. The accuracy of the analytics or ML/AI algorithms is highly dependent on the quality of the data used as well. In V-RAN, domain knowledge is an essential data input that is collected from operations and descriptive data analytics, such as (i) knowledge about relevant & important alarms and faults, (ii) knowledge about dependencies among site components and possible QoE effects, and (iii) knowledge about operational and business contexts in each region based on business phenomena. Since engineers' profiles are disparate, their knowledge inputs are diverse as well as operation ways are culture- and context-specific, we focus on building "micro" templates for acquiring knowledge.

E. Models of analytics subjects and dependencies

1) Analytics subjects: An entity under ML analytics is called *analytics subject*. We use it to represent a physical or abstract entity. We devise a common structure to describe properties of analytic subjects so that concrete analytic subjects can be extended. Each analytics subject is characterized by:

- *measurements*: measurements monitored, collected or gathered, or calculated for the subject. This can be based on existing raw/processed measurements (e.g., network KPIs), analytics results, or inputs from human.
- *conditions*: pre-defined conditions that help to determine the status of the entities. These conditions are established based on descriptive analytics and domain knowledge.
- *statuses*: based on measurements and conditions and other analytics, we can determine the status (e.g., a bad cell). A status can be predictive or descriptive, determined by a function via a (complex) analytics in the infrastructure.

Concretely, several analytics subjects are directly derived from existing data and network structures, for example, `Site`, `Cell`, `Unit` and `Zone`. Other analytics subjects are defined based on the goal of V-RAN operations. These subjects are partially based on existing data and mobile network structures.

ML analytics can be designed based on individual analytics subjects or on the couplings among different analytics subjects, such as `QoE` for `DataService` in a `Zone`. Furthermore, the couplings are based on specific views. For example, in the view of a subscriber, a `CustomerQoE` is associated with a service, e.g., `DataService` in a `Zone` where `Zone` can be established based on quite uncertain information from the feedback (e.g., based on Uber H3 distance and an approximate position). However, in the view of the engineer, a `CustomerQoE` could be defined for a service in a well-defined `Zone` or `Site` identified by deployment data. These couplings require a combination of different ML analytics or new analytics implemented with different capabilities.

2) Relationships: We define key different relationships among analytics subjects based on observations from the descriptive analytics of existing data:

- `consistsOf`: explains an entity consists of another entity. For example, a site consists of many cells.
- `withinZone`: explains geographical and business zone relationship. For example, a site is within a district.

Profile	Role	Scopes of data analytics	Provided domain knowledge
Site engineer	operating sites & handling alarms	single sites and zones	alarm relevancy factors and context for alarm analysis
Layer 1 (L1) RAN optimization engineer	optimizing site/zone coverage, handling customer feedback	single sites and zones	business and community events with a large number of subscribers; mappings between feedback and services
Layer 1 (L1) Transmission engineer	dealing with transmission management and configuration at the site	single sites and a set of sites	affected area/sites due to quality deterioration of a transmission line or transmission node
Layer 2 (L2) RAN optimization engineer	maintaining and improving network performance and customer QoE	radio network	conditions of network KPIs/KQIs for subscriber QoE analytics
Layer 2 (L2) Transmission engineer	managing network transmission (high-level design)	transmission network	affected service due to network transmission issues
RAN operation engineer	operating RAN network and implementing RAN configuration changes	RAN-equipment configuration	affected service due to deployment and configuration changes

TABLE I
EXAMPLE OF PROFILES IN A HOLISTIC FRAMEWORK

- `distanceFrom`: explains the distance between two entities whereas the distance can be defined based on physical or abstract measurement.
- `occurredAt`: explains the place or time at which an event, a behavior or an observation happens.
- `happensBefore`: explains the happen-before relationship between two events (e.g, between two faults).
- `overlapsWith`: explains the overlap of duration between two phenomenons.
- `hasEffects`: explains the cause relationship. For example a fault triggers another fault that can be detected from two alarms via data analytics.
- `hasChanges`: captures changes (e.g., based on measurement, anomaly detection or prediction).

Some of these relationships are static or unchanged during a long period of V-RAN operations (e.g., `consistsOf`, `withinZone` and `distanceFrom`). Other relationships are established within a time frame (e.g., `happensBefore`, `overlapsWith`, `hasEffects`, and `hasChanges`).

3) Model of composable analytics: Given a system with a large number of analytics subjects, it is extremely challenging to apply ML to individual subjects (e.g, we could have 7000 sites as individual analytics subjects). On the other hand, applying the same techniques for all analytics subjects of the same type often does not bring useful results due to the specific context for these subjects. In this view, we see that similar physical entities (e.g. cells from the same vendor) can have similar types of data (e.g., network measurements and alarms). However, the ML analytics can be similar or different due to *the context* of deployment conditions and business goals for the entity. This view provides information about how to obtain and extract data for analytics of a given analytics subject.

F. Managing feasible data analysis and ML tools/techniques

To support engineers to select and invoke suitable analytics, we manage the relationships between analytics, data types, and types of output, as follows:

- A domain problem (dp) associated with an analytics subject (as) is identified that needs to be analyzed.
- Given an input data in belonging to a dataset $ds \in DS = \{networkKPI, alarm, incident, \cdots\}$ we can apply a technique $tech$ in a set of techniques T. Each $tech \in T$ produces an output out.
- An output $out \in O$ is associated with an output category (OC). We have defined the following categories $OC = \{dd, cp, fo, ad, ce\}$ where dd is drift detection, cp is change point, fo is forecasting, ad is anomaly detection, and ce is causal effect.
- There are different ML algorithms $ALG = \{alg_i\}$ that can be used within $tech \in T$ for producing the output $out \in O$ belonging to $oc \in OC$.

Table II shows examples of such information. The information is used for selecting and invoking suitable analytics in different workflows for the engineers and explaining analytics results.

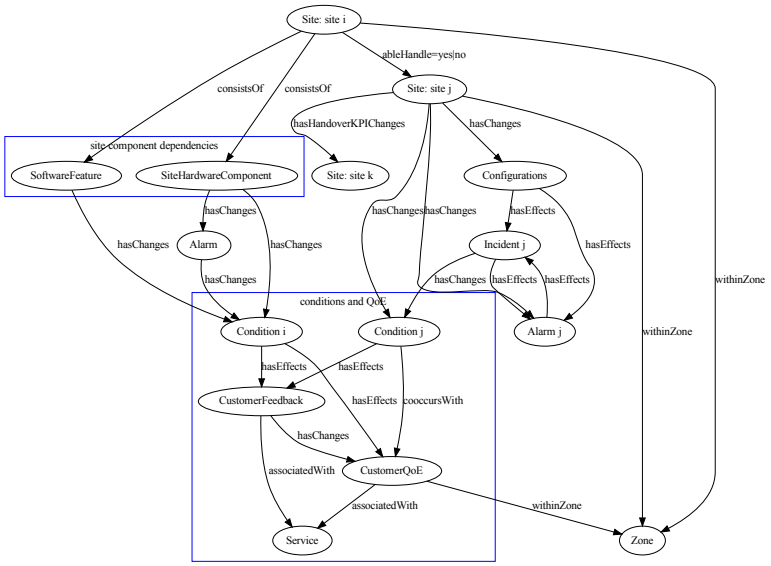

Fig. 3. Example of relationships between different analytics subjects (`Site`, `Service`, `CustomerFeedback`, `CustomerQoE`, and `Zone`) and other relevant components and data. Based on the data, we can build such a graph for a selected window of time. For each site it is possible to have anomaly detection, forecasting and change point detection analytics based on the data.

G. Putting things together: Examples

Figure 3 shows an example of dependencies based on analytics subjects, relationships, and domain knowledge. Such dependencies are built from monitoring data, domain knowledge, and existing analytics results. They can be used as input for guiding workflows of ML analytics. Figure 4 shows an example of using domain knowledge, engineer-in-the-loop, and ML algorithms to determine and label important patterns in traffic/utilization anomalies.

IV. RELATED WORK

Currently, big data analytics in telcos is usually carried out for a single type of (big) data, such as alarms, network traffic,

Domain problems: analytics subjects	Datasets: Input data elements	ML methods/techniques	ML tools/algorithms	Output categories
Quality of Experience: `Zone`, `VoiceService`, `DataService`	Deployment data: site/cell identification, operational & business contexts; Feedback data: serving cell, feedback category/causes; Network measurements: KPIs/KQIs, subscribers, usage statistics	MLP, SVM, regression, clustering, t-digest	Tensorflow-based MLP, Apache Spark KMeans	anomaly detection, forecasting, change point, causal effect
Anomaly analytics: `Cell`, `Site`, `Zone`	Deployment data: site/cell identification, operational and business contexts; Alarm data: alarm severity, faultID, equipment unit, business day and hour; Network measurements: affected network KPIs/KQIs, service traffic; Incident data: incidents causes	Density-based techniques, clustering, regression, t-digest, adaptive windowing, PCA	Anomaly Detection Toolkit (ADTK), Luminaire WindowDensityModel, BOCD, ADWIN	anomaly detection, change point
Network behavior optimization: `Zone`, `VoiceService`, `DataService`	Deployment data: site/cell identification, operational and business contexts; Network measurements: KPIs/KQIs, usage statistics; Configuration changes: service affected, configuration change severity, impact layer (core, RAN, transmission), impact type	Density-based techniques, clustering, regression, ensemble, LSTM	ARIMA; Prophet/NeuralProphet; Kats Global Model, CUSUM & LSTM; evidently	drift detection, change point, forecasting, anomaly detection, causal effect

TABLE II
EXAMPLES OF DOMAIN PROBLEMS, TYPES OF DATA, POSSIBLE ML TECHNIQUES AND TOOLS, AND TYPE OF OUTPUTS

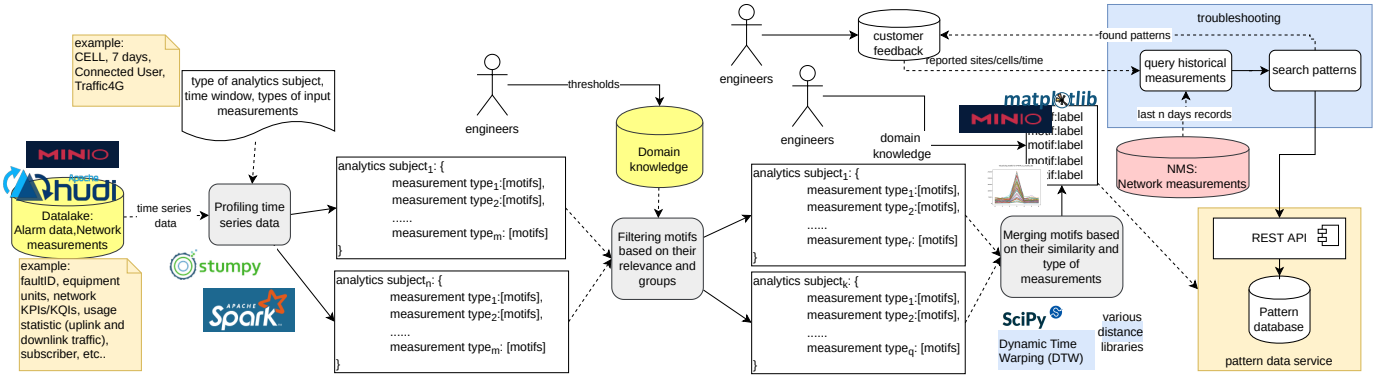

Fig. 4. Example of a composite analytics for finding and applying anomaly patterns for network traffics and utilization that leverages analytics subjects profiles, knowledge about configuration, machine learning and other types of data analytics, and engineer-in-the-loop to provide new knowledge for patterns. Data pipelines extract data from NMS and store the extracted data into the data lake and the team manages the quality of data in the datalake (not shown in the figure). From the datalake, given inputs about analytics subjects, we have run Apache Spark and stumpy to find motifs of traffic and utilization for different analytics subjects. Then, with the domain knowledge by the engineer, we filter irrelevant motifs (based on various conditions). The remaining motifs are then compared and merged based on their similarity and groups. Merged motifs are shown to the engineer so that the engineer can evaluate and assign labels to create reusable patterns. The patterns are used for supporting QoE analytics, given customer feedback at the operation time.

or customer churn. In terms of analytics outputs, we can see a wide availability of industrial analytics, such as statistics of alarms, anomaly detection, and traffic forecasting. In parallel, many ML algorithms have been applied and developed.

In terms of holistic approaches, challenges for dealing with various types of analytics have been presented and discussed [6], [7]. We do not analyze the challenges in this paper but we present our methods to support solving part of these challenges. From the abstraction viewpoint, the paper [7] advocates integration approach for data analytics, but it looks at the very high-level and does not introduce a practical approach like ours. The work in [8] discusses possible architectures and use cases, but does not focus on concrete solutions for domain knowledge and management of analytics like ours.

V. CONCLUSIONS AND FUTURE WORK

Real-world mobile networks with a *variety* of RAN technologies introduce complex data and systems as well as requirements for coherent analytics on complex subjects based on different types of data. Our experiences lead to the focus of building a *practical, holistic* ML analytics infrastructure for specific needs of engineers *under the lack* of ML skills and computing resources in V-RAN. In this paper, we have presented key aspects of our holistic ML analytics infrastructure by focusing on characterizing requirements, data and knowledge, specifying analytics subjects, and managing algorithms.

REFERENCES

[1] G. Kathareios, A. Anghel, A. Mate, R. Clauberg, and M. Gusat, "Catch it if you can: Real-time network anomaly detection with low false alarm rates," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 924–929.

[2] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, 2022.

[3] Y. Huang, F. Zhu, M. Yuan, K. Deng, Y. Li, B. Ni, W. Dai, Q. Yang, and J. Zeng, "Telco churn prediction with big data," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '15, 2015, p. 607–618.

[4] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, "Big data analytics in mobile cellular networks," *IEEE Access*, vol. 4, pp. 1985–1996, 2016.

[5] T. John and P. Misra, *Data Lake for Enterprises: Lambda Architecture for Building Enterprise Data Systems*. Packt Publishing, 2017.

[6] J. Li and H. Liu, "Challenges of feature selection for big data analytics," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 9–15, 2017.

[7] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE Access*, vol. 6, pp. 32 328–32 338, 2018.

[8] M. Z. Kastouni and A. Ait Lahcen, "Big data analytics in telecommunications: Governance, architecture and use cases," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, Part A, pp. 2758–2770, 2022.