CURATE: On-Demand Orchestration of Services for Health Emergencies Prediction and Mitigation

Luis Sanabria-Russo, Jordi Serra, David Pubill, Christos Verikoukis
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA)
{lsanabria, jserra, dpubill, cveri}@cttc.es

***Abstract*—Telemedicine, or the ability granted to doctors to remotely assist patients has been greatly benefited by advances in IoT, network communications, Machine Learning and Edge/Cloud computing. With the impeding arrival of 5G, virtualized infrastructures and cloud-native approaches enable the execution of unprecedented procedures during such patient/doctor interactions, allowing medical professionals to e.g. request higher granularity metrics from patients' telemetry equipment, or perform on-demand data mining/processing of patient's stored data in order to provide a more educated diagnostic or prediction.**

In this work we coalesce the virtues of virtualized infrastructures and IoT into a solution able to satisfy increasing data processing demands for eHealth, e.g. for telemedicine applications, remote assistance or patient pre-screening procedures. The proposed platform, CURATE, leverages Network Functions Virtualisation Management and Orchestration (NFV MANO) for the on-demand instantiation of the required virtual resources on the operator's infrastructure, as well as the concept of 5G Network Slices to guarantee efficient resource allocation and tenant isolation. Results show the proposed platform is able to efficiently make use of the available hardware resources via Network Slices, as well as provide cost-effective service guarantees employing dynamic scaling operations.

***Index Terms*—NFV, IoT, eHealth, Slicing, 5G**

I. Introduction

The 5G Public-Private Partnership Infrastructure Association (5GPPP IA) has identified the growing proportion of elderly population as a key factor of the increase in expenditures on the healthcare sector [1]. It is then no surprise to realise that much of the 5GPPP efforts have shifted towards assessing its root causes, such as lifestyle monitoring, active ageing and wellness. One of the main trends focuses on *out of hospital* medical assistance (i.e. telemedicine) leveraging IoT for metrics collection and actuation, as well as cloud-native approaches and Machine Learning techniques for remote on-demand preventive/reactive analysis or pre-screening procedures (e.g. while on route to hospitals).

To circumvent the CAPital and OPErational expenditures (CAPEX, OPEX, respectively) that would be required in traditional network infrastructures, e.g. to provide lifestyle monitoring applications over large geographical areas, Network Functions Virtualization (NFV) [2] has been proposed. In the NFV vision, most of the network functions supporting the communication infrastructure are placed inside virtualization containers, e.g. Virtual Machines (VM), which are then instantiated on demand on top of fairly generic pools of compute, network and storage resources, i.e. data centers. Whilst adapting its infrastructure to accommodate NFV and other 5G-enabling technologies (e.g. SDN), Mobile Network Operators (MNO) are finding value on sharing its infrastructure with third-party applications developers, which in turn see in the MNO platform's unique services (i.e. ultra low latency via MEC) a key business enabler.

This paper coalesces into an end-to-end 5G service orchestration platform the virtues of constant health metrics monitoring provided by IoT, the prediction and alert capabilities enabled by Machine Learning techniques, the infrastructure dynamism offered by NFV, and the centralized control of resources realized via the standardized NFV Management and Orchestration (MANO) framework [3]. Said platform, referred to as CURATE, leverages the resource heterogeneity characteristic of two-tiered edge/cloud architectures to efficiently satisfy application requirements in terms of computation and latency.

The following Section II provides an outlook of similar proposals in the literature, as well briefly describes Machine Learning techniques leveraged in eHealth applications. The proposed end-to-end eHealth services orchestration platform, CURATE, is presented in Section III, alongside descriptions of the enabling technologies, e.g. NFV. The evaluation platform and scenarios are described in Section IV, while results and conclusions are presented in Sections V and VI, respectively.

II. Related Work

Even though our previous work proposed a platform able to instantiate computation agents tailored to the detection of anomalies on patient data [4], it did not consider Machine Learning techniques nor the standards proposed by ETSI regarding Network Functions Virtualization (NFV). Furthermore, it disregarded the role MNO could play as providers of eHealth services. This section discusses relevant learning techniques used in eHealth applications, as well as how NFV would enable them on top of a dynamic MNO infrastructure.

Deep learning (DL) is a subset of machine learning (ML) methods. They represent a major advance to uncover useful information from raw data [5], which paves the way to build better models than classical ML methods and thereby to improve data analytics tasks

such as classification, prediction or pattern recognition. DL has been recently considered for several healthcare applications, particularly, to analyze time series of health data. For instance, in [6] authors consider Long Short-Term Memory Networks (LSTM), a type of recurrent neural network, to develop a predictive model for healthy Electrocardiogram (ECG) signals. Such predictive model is then used to detect anomalies in ECG signals as deviations. In this regard, the advantage of LSTM, and DL in general, is that they avoid elaborate preprocessing of the raw data. Another example on the applicability of DL to time series of health data is [7]. In that paper authors detect arrhythmias in ECG signals by using LSTM methods. Also, they tackle the class imbalance problem of anomalous ECG signals by using data augmentation techniques. Namely, DL methods known as Generative Adversarial Networks (GAN).

Several works have considered NFV, network slicing and a hybrid IoT Edge/Cloud architecture to pave the way for health data analytics. In [8] authors consider a hybrid Edge/Cloud IoT architecture for the problem of ECG classification. They consider IoT devices that generate health data, then such raw data is sent to edge servers that implement the ECG classification based on a DL algorithm consisting of a Convolutional Neural Network (CNN). They claim that their solution offers lower data analytics delay compared to having the algorithm located at the cloud and that it preserves the user privacy as the data is kept at the edge of the network. The cloud is considered as an interface with doctors and for more accurate inference, though no further details are provided. In [9], a media-centric eHealth use case is considered within the context of 5G network slicing. The scenario is based on a connected ambulance for pre-hospital care in e.g. stroke prevention. The ambulance streams video to a Multi-access Edge Computing (MEC) server. Then, the applications at the MEC analyze the health data and send the result to doctors at a remote location. Their contribution relies on proposing end-to-end QoS network slices to support this mission-critical media service use case and other vertical industries with diverse QoS requirements.

Effort towards an infrastructure-and-specifications agnostic platform as a service (PaaS) for Vertical Service Providers (VSP), like [10], are getting attention mainly due to the abstraction of the complexities of infrastructure management, but also because the services it provides follow a cloud-native approach (e.g. deployment of micro services) which will play a key role into enabling 5G and beyond [11]. Nevertheless, it proposes too big a disruption to the current MNO environment, disregarding valuable specifications (e.g. [2], [3], [12]) that have promoted both industry and community involvement (e.g. [13]).

Contributions

The contributions of this work can be summarized as follows:

- Herein a hybrid Edge-Cloud IoT architecture is considered as in [8]. However, unlike in [8] herein NFV technology, an cloud-centric paradigms for the development of applications is leveraged for health data analytics. Thereby, our resulting architecture is more flexible, programmable, and efficient in terms of computing and storage resources. Moreover, the role of the cloud in health data analytics is described more clearly, as we highlight that it carries out the training of DL algorithms using a large dataset of patients. Consequently, the pre-trained DL methods can be deployed at the edge, thereby still adhering to the low latency and privacy requirements of such type of applications.
- Show the benefits of designing any eHealth application with dynamic scaling-out operations in mind. Results show the number of simultaneous processing request can doubled with the selection of appropriate scaling-out triggers.
- In this work, as in [9] NFV and network slicing are considered for health data analytics. Nonetheless, unlike in [9] a hybrid Edge/Cloud architecture is proposed. Namely, Cloud is used for the training of the DL prediction algorithm and for protecting the privacy of the patients in the training dataset. Then, the pre-trained DL method can be deployed at the Edge, thus achieving low-latency and also complying with privacy regulations applying to patients being monitored. Moreover, unlike [9], this work proposes an scenario where the ML algorithm and frameworks used are thoroughly described.

III. CURATE

This section covers the different elements that compose the CURATE platform. First, a high level description of ETSI Network Functions Virtualization (NFV) is presented. Secondly, CURATE platform is described alongside its three main layers: Field, Edge and Cloud. Thirdly, common ETSI NFV and CURATE interfaces which administrators employ to customize running services or gather telemetry information are overviewed. At the end of the section readers should have a complete picture of the infrastructure and be well aware of its capabilities.

A. NFV Overview

The 5G vision proposes a flexible infrastructure, where a pool[1] of white servers (i.e. data centers) provide virtual compute, network and storage resources to be provisioned on demand in isolated partitions referred to as Network Slices. Such slices host Network Services in the form of Virtual Network Functions (VNFs), which are connected together via Software Defined Network (SDN) overlays.

The ETSI Network Functions Virtualization (NFV) Architectural Framework [14] is illustrated in Figure 1. In it, the NFV Management and Orchestra-

[1]Or geographically distributed pools e.g., data centers in different locations, or the multi-tiered cloud model.

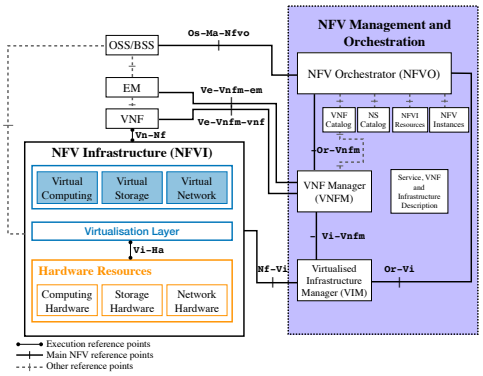

Fig. 1. ETSI NFV Architectural Framework

tion (MANO) block is in charge of determining the availability of virtual resources in the data centers - referred to as NFV Infrastructure (NFVI) - for the orchestration of a network slice, as well as taking care of the lifecycle management of each VNF, provide telemetry information on the state of the NFVI and VNFs, the termination of slices and the release of the virtual resources. NFV MANO is composed of the NFV Orchestrator (NFVO), VNF Manager (VNFM) and Virtualized Infrastructure Manager (VIM). A brief description of their role in the NFV framework is provided below.

1) Virtualized Infrastructure Manager (VIM): Is the responsible for the control and management of the interaction between VNFs and the NFVI hardware resources, such as compute, storage, and network, as well as their virtualization. It takes care of exposing a pool of virtualized resources derived from the NFVI, as well as the allocation of such resources to each VNF.

2) VNF Manager (VNFM): Takes care of VNF lifecycle management. That implies the instantiation, scaling, and termination of one or several VNFs.

3) NFV Orchestrator (NFVO): NFVO is able to gather information about the NFVI from one or several VIMs through standardized reference points or APIs (see Figure 1), and then determine the suitable place in the NFVI to instantiate a VNF. As services are often provided via network slices, NFVO is in charge of satisfying all the slice's VNFs requirements prior orchestration. All in all, NFVO works as an automation tool for instantiating and terminating network slices from a centralized control position. Furthermore, it enables unprecedented infrastructure reutilization by allowing scaling out VNFs at runtime (e.g. for preserving KPIs), or freeing resources at low-demand periods for energy savings.

Scaling out refers to the replication of an existing VNF, conversely, scale in eliminates such replicas. In CURATE, as shown in the proposed Scenario-B in Section IV, any scaling operation is triggered by the NFVO according to a set of user-defined policies. These policies in turn are based on telemetry information (e.g. percentage of CPU usage, available memory, etc.) of the VNFs. Reference VIMs, such as OpenStack [15], provide such telemetry services via projects like Gnocchi [16], which exposes APIs for the NFVO.

B. CURATE description

The proposed CURATE platform is rooted in the aforementioned NFV architecture. It is segmented in three layers, namely: Field, Edge, and Cloud. As names suggest, Field is where data is generated, whereas Edge and Cloud refer to two different Point of Presence (PoP) of the network operator, usually considered part of a NFVI. The former is closest to the Field, while the latter refers to powerful data centers far removed from the users. This so-called two-tiered cloud architecture allows for delay sensitive applications to be orchestrated in VNFs at the Edge, while computational-heavy operations can be orchestrated in the Cloud.

In CURATE, the characteristics of the Cloud and Edge layers are exploited in two main ways. First, as mentioned above computational heavy operations (i.e. model training) are performed with large training sets in the Cloud, while a pre-trained model within a Network Slice is orchestrated at the Edge in order to provide faster reaction times. Second, to enforce data privacy, testing data sets (i.e. data from the Field layer) are not stored in the Cloud but directly tested against the pre-trained model at the Edge.

An overview of the CURATE platform is presented in Figure 2. What follows details the components of each layer.

1) Field layer: In CURATE, the Field layer refers to the user domain. Here, data from sensors are gathered and transmitted to the CURATE platform via LTE. As shown in Figure 2, CURATE is not limited to patients, but is envisioned to satisfy many m/e-Health use cases' computational needs, such as patient's pre-screening while still in the ambulance, or serve as support infrastructure for processing data from homes or hospitals.

2) Edge layer: This layer holds the required virtual compute, network and storage resources to:

- Receive and temporarily store patients' metrics from the Field layer.
- Instantiate the services as network slices required to process patients' data.
- Provides NFV telemetry information to the Cloud layer for management purposes.

More relevantly, the Edge layer supports the network slices containing pre-trained services that guarantee faster anomaly detection in the data received from the Field layer.

3) Cloud: Is composed of powerful compute nodes and large training sets, better suited for faster model training than the Edge. It also holds the controllers of the whole infrastructure, that is: NFV MANO, and Radio Access Network (RAN) Slice Manager (i.e. FlexRAN [17]); which makes it the entry point for Operations/Business Support Systems (OSS and BSS, respectively) to the infrastructure. That is, Network Slices (i.e. services) are requested and orchestrated from the Cloud.

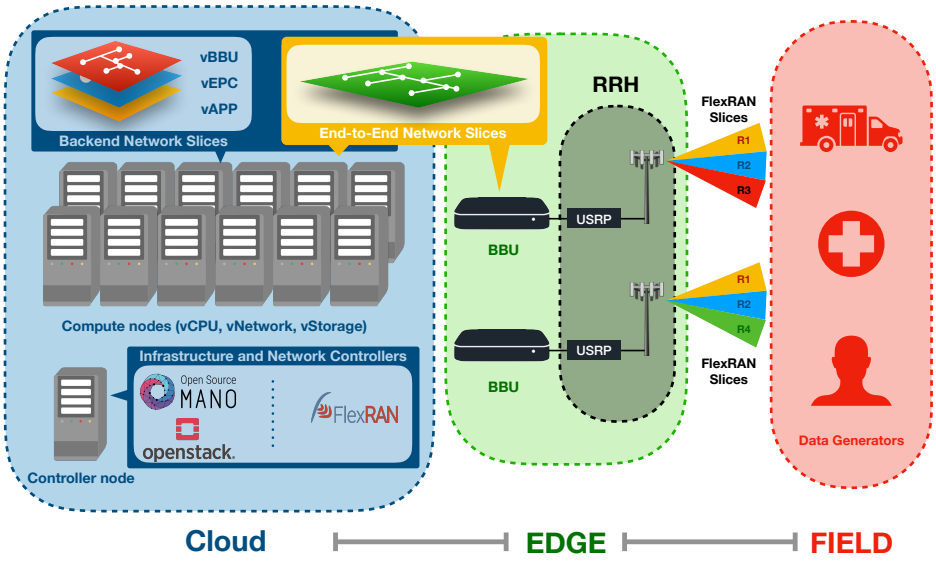

Fig. 2. CURATE platform overview, from left to right: Cloud (holding powerful compute nodes, ample storage and the MANO controllers of the whole infrastructure), Edge (also part of the NFV Infrastructure but composed of smaller compute nodes and Software Defined Radios (SDR) to support LTE links towards the Field), and Field (user domain, data generators).

C. Domains

With such a high collection of elements, this section attempts to assign ownership or lack thereof to the different layers of the CURATE topology, and also its administrative domains. As shown in Figure 2, CURATE infrastructure has three layers, each one corresponding to an infrastructure domain of the same name. Without loss of generality, it can be safely assumed that Edge and Cloud layer are typically owned by a MNO, while the Field can be composed of Commercial Off-The-Shelf (COTS) hardware[2]. This does not restrict the event of MNO outsourcing Cloud functions (e.g. AWS, Azure), but indeed stresses the relevance of the Edge and the competitive advantage it offers MNO, e.g. to develop cloud services offering low latency or edge computation [12], [18].

As MNO may lease portions of their NFVI (i.e. as Network Slices), a description of different administrative domains is required. The common scenario assumes a MNO owns the NFVI, but a Vertical Service Provider (VSP) owns the application, i.e. eHealth provider. That is, the eHealth provider is responsible for the correct operation of the application, while MNO should ensure Lifecycle Management (LCM) of the slices and elements therein (e.g. VNFs, virtual links). CURATE, via the scenarios proposed in Section IV demonstrates the deployment of eHealth services assuming complete control of infrastructure and administrative domains. Therefore the procedures detailed in Section IV cover both perspectives, i.e. MNO and VSP.

D. CURATE Interfaces

CURATE allows authorized third-parties (e.g. OSS) to request the instantiation of a predefined Network Slice to the NFV Orchestrator. Furthermore, these same authorized entities may query the VIM in order to gather relevant performance metrics from the Cloud or Edge layers[3]. In the evaluations that follow, medical professionals will trigger the instantiation of services in Network Slices from the position of OSS/BSS in Figure 1, that is, through NFVO Application Programming Interfaces (APIs) provided via the `Os-Ma-Nfvo` reference point.

IV. System Evaluation

This section evaluates the capabilities of the CURATE platform through two different but complimentary scenarios.

In these evaluations, patients or data generators are emulated via software. Specifically, patients' data generators run on top of Raspberry Pi 3 Model B devices. Metrics from such generators are transmitted via LTE towards the virtual Evolved Packet Core (vEPC), which are then routed towards the corresponding receiver within a network slice. The aforementioned LTE link is composed of three elements: 1) the User Equipment (UE), 2) the Remote Radio Head (RRH) and Baseband Unit (BBU), and the 3) vEPC. The first element is a commercial USB stick connected to the data generators, such as the Huawei E3372. The second is an USRP software-defined radio, specifically National Instruments B210 as RRH, and as BBU OpenAirInterface (OAI) [19] in run on top of an Intel NUC i7 3.5GHz. The vEPC is OpenAir Core Network (OAI CN), instantiated as a VNF at the Edge.

The Edge and Cloud layers of CURATE are considered a single NFVI. Edge is composed of 3 Intel NUC i7 3.5GHz, while Cloud holds 2 Intel Xeon Skylake Gold servers, each one with 12 cores at 2.3 GHz. As

[2]With the exception of the SIM card.

[3]Both Cloud and Edge are considered part of the NFV Infrastructure (NFVI), and are managed by a single VIM located at the Cloud layer.

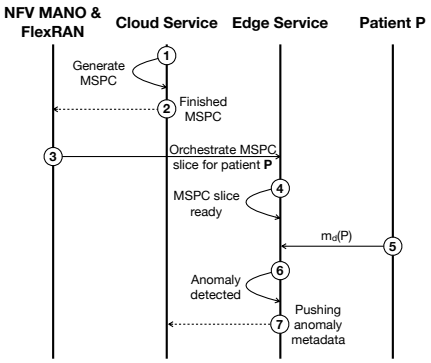

Fig. 3. Scenario A: message sequence diagram

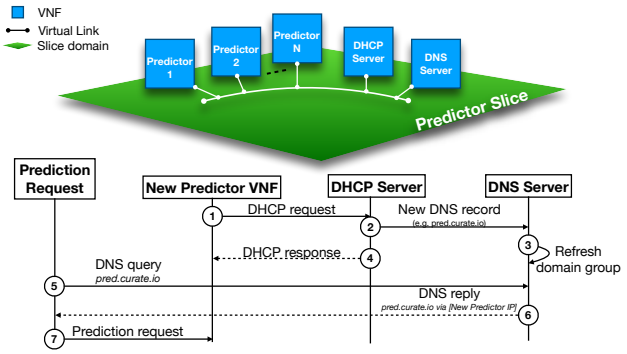

Fig. 4. Components of a Prediction Slice (PS) and message sequence diagram for the registration of a newly created VNF (e.g. predictor replica resulting from a scaling-out operation) in the server pool for load balancing.

shown in Figure 2, ETSI OSM release SIX [20] is used as NFVO, while OpenStack Queens [15] fills the VIM and VNFM role.

Scenarios description

Across all scenarios, patients, P are assumed to be equipped with telemetry devices, i.e. sensors. Each sensor gathers a collection of metrics, m, every pre-defined reporting interval d (in seconds). Metrics include: heart rate (hr), breathing rate (br), blood oxygen (sato), temperature (t), diastolic and systolic blood pressure (bpd and bps, respectively). Therefore, $m_d(P)$, refers to the batch upload of metrics m, from patient P which is programmed with a reporting interval d.

1) Scenario A - Emergency notification: This base scenario attempts to showcase an example eHealth application. It assumes that a day-worth of data from a patient P is available, and therefore a Multivariate Statistical Process Control (MSPC) could be computed at the Cloud. Such model is then orchestrated in a so-called *MSPC slice* per patient at the Edge. Subsequent metric receptions from patient P will then be compared to such model in the corresponding slice in search for breaks in the correlation of metrics (according to a pre-defined threshold). If an anomaly is detected, the health professional is notified immediately. This scenario is used to exemplify all the elements and procedures commonly expected of an eHealth application on top of CURATE. Figure 3 provides a message sequence diagram for this generic scenario.

2) Scenario B - Metrics' trend prediction: This scenario envision the on demand creation of a *Prediction Slice*, PS, at the Edge. PS holds the required *predictor* VNFs to predict the trend of all metrics for a patient. The medical professional may then evaluate the result and take appropriate action. More specifically, in this scenario we consider the prediction of time series of health data, e.g. ECG, in a future time horizon that is defined by the medical professional.

In this scenario, a given time series excerpt is associated to a given patient. It is also assumed that such time series are sent to the PS located at the Edge layer. Namely, we assume that N forecast requests are sent to the PS, and that in order to comply with such requests the so-called predictor VNF is designed to scale-out.

A service that effectively scales should be stateless and be provided with networking services. The former means that it should be able to perform its function without persistent storage, much like a *black box* function. The latter relates to the enabling of load balancing among active instances of the service. The predictor VNF for the eHealth service in CURATE is stateless, that is, it processes incoming time series according to an already-defined model[4] without relying on persistent storage. Furthermore, to allow load balancing among active replicas of the predictor VNF, DHCP and DNS services were configured within PS leveraging ISC [21] and Bind9 [22], respectively. The composition of a given PS as well as an example message sequence diagram for the registration of a newly created VNF are shown in Figure 4. In the aforementioned figure, the algorithm selecting the appropriate reply (i.e. predictor VNF address to return) is not reflected. In Scenario-B, the DNS server in a PS performs a round-robin of the available addresses in the domain group, e.g. *pred.curate.io* in Figure 4.

Regarding the scaling-out of the predictor VNF, it is setup based on percentage of CPU usage ($\%_{\text{CPU}}$). Precisely, the predictor VNF will be replicated if $\%_{\text{CPU}} > \alpha$ during $t > 10$ s; where $\alpha = 60\%$ is a CPU usage threshold, and t is referred to a threshold time. This ensures a relatively fast trigger of scaling operations given sudden surges in CPU usage resulting from the prediction process. Scaling in is also subject to similar parameters. NFVO will trigger the termination of replicas when $\%_{\text{CPU}} < \beta$ during $t_c > 60$ s; where $\beta = 20\%$ and t_c is referred to as cooldown time.

The selection of these parameters is heuristic, nevertheless it does not fall far from default values used by Container Orchestration Engines (COE) such as Kubernetes ($\alpha = 50\%$, $t > 30$) [23]. To provide the NFVO with performance metrics of sufficient resolution, the default polling intervals at VIM telemetry services where modified. Different research directions may be derived attempting to find optimal telemetry

[4]The one previously trained at the Cloud.

and scaling out parameters for particular traffic loads, or to reduce the application footprint on the infrastructure.

The main benefit of locating the prediction algorithm at the Edge layer, compared to the traditional approach of locating it at the Cloud, is that lower latency can be provided. Moreover, we address the privacy concerns of locating patient's health data at a remote backend cloud, as the data is analyzed closer to its generation source and network slicing guarantees isolation among users (i.e. among PS).

Regarding the prediction algorithm, herein we consider LSTM, a type of recurrent neural network (RNN). The rationale is that LSTM are among the most effective and widely applied DL methods to analyze time series, see e.g. [5, Ch. 10] and references therein. In fact, compared to classical RNN, LSTM are able to learn long-term dependencies of the sequential data more easily. Also, they avoid convergence problems of classical RNN, which arise during the training of the algorithm due to the vanishing gradient issue.

Finally, we assume that the LSTM is trained at the backend cloud. This permits the use of a larger number of time series excerpts for training purposes. Further, as Cloud holds more computational resources than the Edge, more exhaustive training can be done, which leads to a more accurate LSTM model. Afterwards, the pre-trained LSTM algorithm is deployed to the PS located at the Edge. Note that the PS is ready for predicting the time series of health data and no further training is required at the Edge, this allows for a reduction on the footprint scaling out operation would leave on the NFVI, as replicas of predictor VNFs are relatively lightweight[5]. This two-tiered Edge/Cloud approach guarantees low latency, preserves the privacy of the health data of the patients, and lowers the footprint on NFVI resources.

V. Results

This section discusses the outputs from the previously defined scenarios. Specifically, an example of the output from Scenario-A is provided, while the aggregated number of processed requests employing scaling-out strategies on top of CURATE are shown for Scenario-B.

A. Output from MSPC workers

As shown in Figure 3, the anomaly detection triggers the push of metadata to a Cloud service. An example of such output is shown in Figure 5. The figure highlights the fact that the observed value of the patient's temperature (t) deviated too much from past trends. With such information the medical professional may decide to contact the patient in order to check her status.

Scenario-A attempts to highlight the elements and the procedure followed by applications running on top of CURATE. Namely, the training of a model in the Cloud, and then the orchestration of pre-trained processing VNFs within a slice at the Edge. The benefits of this approach include: greater computational capacity at the Cloud, which may handle larger data sets and therefore yield a more precise model; and a faster anomaly detection given pre-trained VNFs in each *MSPC slice*. Notice that for exemplification purposes, Scenario-A can be considered agnostic of the type of model being generated in the Cloud, i.e. sequence diagram in Figure 3 holds regardless of the type of processing done in Step 1.

[5]1 vCPU, 1 GB RAM, no persistent storage.

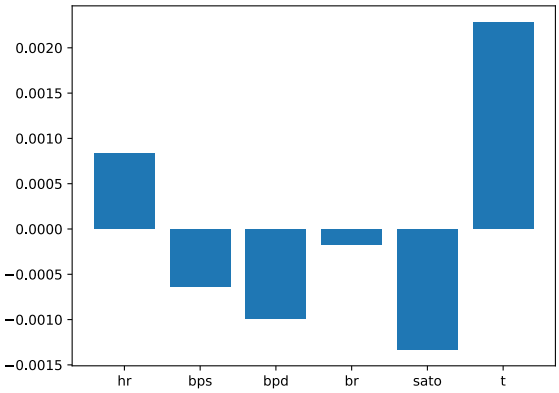

Fig. 5. Sample of a temperature (t) anomaly sent to the Cloud service from Scenario-A. X-axis show the name of other monitored metrics, while Y-axis shows a normalized measure of error between the observed and the expected correlations between metrics.

B. Leveraging Edge and dynamic scaling

In Scenario-B a LSTM neural network is trained at the Cloud for the task of predicting a time-series of health data, namely ECG time series.

For the experiment purposes we use the MIT-BIH database, which contains ECG time series [24]. More specifically, it contains two-channel ECG recordings for 48 patients. Moreover, we have considered a stacked LSTM neural network with an input layer, followed by two LSTM layers and then a fully connected layer as the output layer. This architecture is shown in Figure 6. Note that the proposed LSTM architecture belongs to the family of DL algorithms.

The dimensionality of the input layer is defined by the history length of the time series considered for training purposes and the number of features. We consider 120 past points of the time series and 2 features, which correspond to the two possible ECG channels associated to each patient in the MIT-BIH database. The first LSTM layer has a dimensionality of 32 hidden units and a hyperbolic tangent activation function. The second layer, has 16 hidden units and we consider a ReLU activation function. The dimensionality of the output layer is defined by the number of time-series data points that we forecast. For Scenario-B, a future horizon prediction of 72 points is considered. A RMSprop optimizer was used for the training of the stacked LSTM method. This is a widely used variant of the stochastic gradient method and in

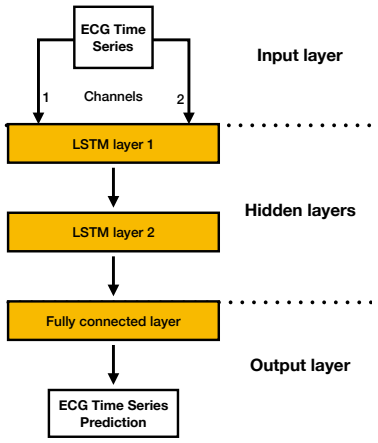

Fig. 6. Architecture of the proposed stacked LSTM predictor.

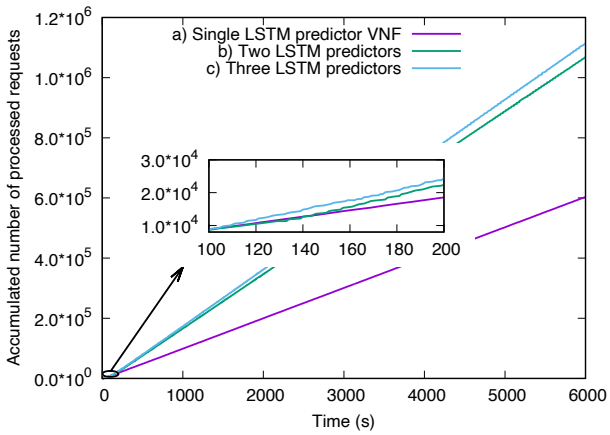

Fig. 7. Accumulated number of forecast requests using different dynamic scaling strategies. a) No dynamic scaling, b) Scaling-out a single predictor VNF, c) Scaling-out two predictor VNFs.

general provides excellent adaptation of the learning rate in the training process. Also, it permits to work with mini-batches of training data to ensure a good tradeoff between convergence speed and stability. The loss function considered for the training is the mean absolute error between the LSTM output and the target.

A training set of $3 \cdot 10^5$ points and 2 features is used for the stacked LSTM network described above. The data set used for validation has $1 \cdot 10^5$ points and 2 features. This corresponds to the ECG time series length and the two ECG channels, respectively. Also, 10 epochs and 200 steps per epoch are used during the training phase. Furthermore, mini-batches of 256 training examples are used. In the validation, 50 steps were considered. Also, it is worth mentioning that Keras, Tensorflow release 2.0 [25] and Python are used for the implementation of the LSTM and for the data preparation of the MIT-BIH database.

The training at the Cloud paves the way to deploy a pre-trained LSTM algorithm at the Edge. That is to say, the LSTM is already trained and it is devoted to predict the time series sent from the Field. This allows low latency and circumvents any privacy concerns as data is not stored. Note that our approach leverages the benefits of a hybrid Edge/Cloud architecture more properly than the related work [8]. As they do both training and deployment at the Edge and do not consider the Cloud for training purposes.

In Scenario-B the proposed evaluation attempted to measure the number of concurrent forecast requests. Taking advantage of the Edge layer, the PS was setup so the predictor VNFs could scale-out if the $\%_{\text{CPU}}$ usage exceeded a predefined threshold. This effectively makes the PS grow on demand and then shrink to its original components when no longer stressed. Figure 7 shows the accumulated number of forecast requests handled by the PS with different scale-out configurations, namely: none, single and up to two additional predictor VNFs per PS.

Figure 7 shows that by taking advantage of the Edge layer and dynamic scaling-out operations for predictor VNFs within the PS, a positive impact can be obtained on the number of forecast requests served. Namely, for a given processing time, more requests can be served by scaling the virtual computing resources, i.e. instantiating more of the same VNF instances. Note that this is more flexible than a traditional approach where no virtualization of the infrastructure resources is used, as in [8]. That is, we can adapt to the number of requests by scaling-out the virtual resources and consequently maintain the quality of service, i.e. the processing time. Furthermore, were a sudden surge on prediction request occur, these would be handled automatically: requests will be handled by available predictor VNFs[6], or new replicas will be created as part of a scale out operation orchestrated by the NFVO.

Conventional approaches, such as [8], are not able to adapt to an increasing number of requests while maintaining the processing time. Thereby, CURATE's approach is able to reduce CAPEX/OPEX compared to traditional non-virtualized approaches. Another interpretation of Figure 7 is that for a given number of requests, the scaling-out of resources allows the reduction of the requests' processing time, leading to better quality of service than [8].

C. General comments

In the process of comparing CURATE's NFV approach versus a traditional deployment of physical network functions, focus should be put at the alleged virtues of virtualization. CURATE would impose less operational expenses if solely due to the reduced amount of hardware components, but also thanks to the ability to centralize control. As applications and network functions are realized in software, a joint Management and Orchestration (MANO) of the infrastructure is possible. This fact greatly simplifies traditionally complicated tasks, such as updating

[6]Load balanced by the DHCP and DNS services at each PS, see Figure 4.

applications and/or communications infrastructure's services on demand.

CURATE emulates a business trend being explored by MNO and enabled by 5G, which consist on offering Platform as a Service (PaaS) instances to application developers who seek to reap the benefits of multi-tiered architectures (e.g. MEC for lower latency), such MNO's. This poses new challenges when MNO attempt to share their infrastructure, but standardization groups like ETSI NFV ISG are providing guidelines to enable this kind of PaaS for cloud-native applications [26], [10]. All in all, multi-tiered clouds offer enough resources for massive data processing as well as sufficient flexibility for faster processing and response. Such benefits are valuable in selected emergency management scenarios, e.g. ambulances [27]; but may also become the default paradigm for more common medical services [28].

VI. Conclusions

In the face of increased expenditures in the *centralized* healthcare system and an increasing proportion of elderly population, 5GPPP IA proposes the decentralization of medical assistance (e.g., telemedicine or at-home healthcare) in order to target lifestyle monitoring, active ageing and wellness leveraging pervasive monitoring from IoT, state of the art prediction techniques provided by Machine Learning or Artificial Intelligences techniques, and a flexible virtual communications infrastructure under the NFV umbrella of specifications.

In this work we present CURATE, an NFV-rooted two-tiered Edge/Cloud platform able to support eHealth applications leveraging the aforementioned technologies. CURATE emulates a MNO's infrastructure as a Edge/Cloud NFVI, where computational heavy operations (e.g. ML model training) are delegated to VNFs at the more powerful Cloud layer, while pre-trained models are orchestrated as separate network slices at the Edge for resource isolation, and faster prediction or anomaly detection. CURATE is implemented with Commercial Off The-Shelve (COTS) hardware and open source tools.

Results from the evaluation highlight the virtues of a dynamic NFVI, specifically, coupled with a paradigm shift towards cloud-native applications, CURATE is able to support automatic scaling-out operations that guarantee a response to an increasing number of simultaneous prediction request at the Edge. Furthermore, efficiency is enforced by automatically releasing the NFVI resources during scaling-in operations.

Acknowledgements

This work has been funded by the following research projects: 5G-SOLUTIONS (856691), SEMIOTICS (780315), SPOT5G (TEC2017- 87456-P), and Generalitat de Catalunya under grant 2017 SGR 891.

References

[1] Anastasius Gavras, et al, "5G and eHealth," https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf.

[2] ETSI NFV ISG, "Network Functions Virtualisation (NFV); Infrastructure Overview," https://www.etsi.org.

[3] ——, "Network Functions Virtualisation (NFV); Management and Orchestration," https://www.etsi.org.

[4] L. Sanabria-Russo, D. Pubill, J. Serra, and C. Verikoukis, "IoT Data Analytics as a Network Edge Service," in *IEEE INFOCOM WKSHPS 2019*, April 2019, pp. 969–970.

[5] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, http://www.deeplearningbook.org.

[6] S. Chauhan and L. Vig, "Anomaly Detection in ECG Time signals via Deep Long Short-Term Memory Networks," in *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2015.

[7] P. Wang, B. Hou, S. Shao, and R. Yan, "ECG Arrhythmias Detection Using Auxiliary Classifier Generative Adversarial Network and Residual Network," *IEEE Access*, vol. 7, pp. 100 910–100 922, July 2019.

[8] J. Yu et al., "EdgeCNN: A Hybrid Architecture for Agile Learning of Healthcare Data from IoT Devices," in *IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, Dec. 2018.

[9] Q. Wang, N. Nikaein et al., "Enable Advanced QoS-Aware Network Slicing in 5G Networks for Slice-Based Media Use Cases," *IEEE Trans. on Broadcasting*, vol. 65, no. 2, pp. 444–453, June 2019.

[10] A. Mimidis-Kentis, J. Soler, P. Veitch, A. Broadbent, M. Mobilio, O. Riganelli, S. Van Rossem, W. Tavernier, and B. Sayadi, "The Next Generation Platform as a Service: Composition and Deployment of Platforms and Services," *Future Internet*, vol. 11, no. 5, 2019.

[11] 5G-PPP Software Network Working Group, "Cloud-Native and Verticals' services 5G-PPP projects analysis," https://5g-ppp.eu.

[12] ETSI MEC ISG, "Multi-access Edge Computing (MEC) Framework and Reference Architecture," https://www.etsi.org.

[13] Huawei, et al, "5G Service-guaranteed network slicing white paper," https://www-file.huawei.com/-/media/corporate/pdf/white%20paper/5g-service-guaranteed-network-slicing-whitepaper.pdf?la=en.

[14] ETSI NFV ISG, "Network Functions Virtualisation (NFV); Architectural Framework," https://www.etsi.org.

[15] OpenStack, "OpenStack Queens," https://www.openstack.org/software/queens/.

[16] Gnocchi, "Gnocchi," https://gnocchi.osci.io.

[17] X. Foukas and N. e. a. Nikaein, "FlexRAN: A flexible and programmable platform for software-defined radio access networks," in *Proceedings of the 12th International Conference on emerging Networking EXperiments and Technologies*. ACM, 2016, pp. 427–441.

[18] Huawei, "Telco Cloud: Business value and Solutions," https://carrier.huawei.com/en/products/service-and-software/Telco-Cloud.

[19] N. Nikaein et al., "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.

[20] ETSI, "Open Source MANO," https://osm.etsi.org.

[21] I. S. Consortium, "ISC DHCP Server," https://www.isc.org/dhcp/.

[22] ——, "BIND 9 DNS Server," https://www.isc.org/bind/.

[23] Kubernetes, "Kubernetes Horizontal Pod Autoscaler algorithm," https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/#algorithm-details.

[24] G.B. Moody and R.G. Mark, "MIT-BIH Arrhythmia Database," https://www.physionet.org/content/mitdb/1.0.0/.

[25] Tensorflow, "Tensorflow r2.0," https://www.tensorflow.org/.

[26] ETSI NFV ISG, "Network Functions Virtualisation (NFV) Report on the Enhancements of the NFV architecture towards "Cloud-native" and "PaaS"," https://www.etsi.org.

[27] R. Guedes, V. Furtado, T. Pequeno, and J. J. Rodrigues, "Pareto set as a model for dispatching resources in emergency centres," *Peer-to-Peer Networking and Applications*, vol. 12, no. 4, pp. 865–880, 2019.

[28] M. W. Moreira, J. J. Rodrigues, N. Kumar, K. Saleem, and I. V. Illin, "Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems," *Information Fusion*, vol. 47, pp. 23–31, 2019.