




A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology

## Deliverable D5.1: Virtual environment for building flexible AI solutions

<b>Reference</b>	D5.1_ EuCanImage_UB_v1
<b>Lead Beneficiary</b>	UB
<b>Author(s)</b>	Josep Ll. Gelpí, Laia Codó
<b>Dissemination level</b>	Report
<b>Type</b>	Public
<b>Official Delivery Date</b>	30 September 2021
<b>Date of validation of the WP leader</b>	30 September 2021
<b>Date of validation by the Project Coordinator</b>	30 September 2021
<b>Project Coordinator Signature</b>	

EuCanImage is funded by the European Union's H2020 Framework  
Under Grant Agreement No 952103



## 1. Version log

Issue Date	Version	Involved	Comments
31/08/2021	v 0.1	Josep Ll. Gelpí, Laia Codó	First Draft
29/09/2021	v 0.2	Kaisar Kushibar	Completing Cincal use-case section
30/09/2021	V1	Karim Lekadir, Isabell Tributsch	Revised and corrected final version.

## 2. Executive Summary

The AI Research Environment (AI-VRE) is part of EuCanImage platform and focuses on providing a flexible computational baseline for supporting the development and assessment of radiomics methods, distributed learning and interpretable AI algorithms. Here we present the layout of the AI development platform, which exposes an integrative and intuitive research environment for data analysis on top of a cloud-based computational backend. The document details how the environment links to data infrastructures leveraged within EuCanImage, how modularity permits the continuous and rapid integration of tools from AI scientists, and how virtualization provides computational elasticity and portability to the system for evolving towards a distributed scheme. The document also describes the first prototype installation of AI-VRE, hosted in a private cloud at ELIXIR-ES (ELIXIR Spain) facilities and accessible at <https://vre.eucaimage.eu>.



## Table of Contents

1.	Version log	2
2.	Executive Summary	2
1	Introduction	4
1.1	Motivation and strategy	4
1.2	Background	5
2	Conceptual design	6
2.1	Requirements	7
3	Platform components	7
3.1	Cloud infrastructure	7
3.2	Research Environment	8
3.3	Job process management	11
3.4	Tools integration	12
4	AI Toolbox	13
5	Towards a Federated infrastructure	14
5.1	Roadmap	14

## Acronyms

Name	Abbreviation
Application Programming Interface	API
Artificial Intelligence	AI
Biobanking and Biomolecular Resources Infrastructure	BBMRI-ERIC
EuroBioImaging	EuBI
European Genome-Phenome Archive	EGA
Magnetic Resonance imaging	MRI
Mammography	MG
Open Cloud Computing Interface	OCCI
Open Grid Scheduler	OGS
OpenID Connect	OIDC
Programming Model Enactment Service	PMES
Spanish National Institute of Bioinformatics	INB
Virtual Machine	VM
Virtual Research Environment	VRE



## 1 Introduction

One of the goals of EuCanImage is to build a highly secure, federated and large-scale cancer imaging platform for enhancing artificial intelligence (AI) in oncology and radiology. Designed as a set of interconnected infrastructures and services, the EuCanImage platform is built upon well-established frameworks and platforms accomplishing complementary tasks of the overall biomedical imaging data flow. On that account, EuCanImage data infrastructures are in charge of coordinating consortium's data deposition of imaging and non-imaging datasets for enhanced data discovery as part of WP2. Data-privacy preserving platforms are being put in place for data anonymization, curation and collaborative annotation by WP3. Novel AI toolboxes for radiomics, distributed learning and interpretable AI algorithms will be compiled in WP5 and transparently assessed by an AI benchmarking platform as defined by WP6 metrics and procedures. The newly developed AI applications will be integrated in the **AI Virtual Research Environment<sup>1</sup> (AI-VRE)** for being tested, benchmarked and published across and beyond EuCanImage researchers and data scientists.

The service architecture and the software components behind AI-VRE are described in detail in this document, and demonstrated with a pilot installation live at <https://vre.eucanimage.eu/>.

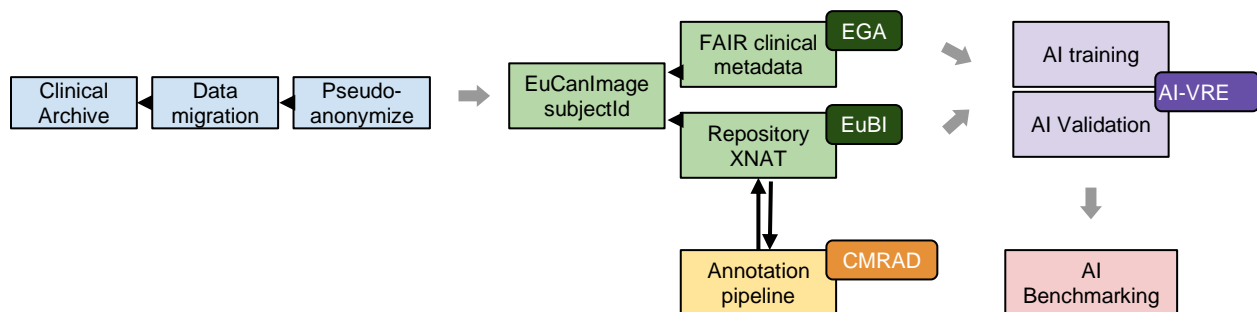


Figure 1: Illustration of the general EuCanImage data flow

### 1.1 Motivation and strategy

The past decade has become the great rise of Artificial Intelligence, moving rapidly from the fundamental research domain towards the translational biomedicine and bioinformatics field. Accordingly, end-users have increased in number and heterogeneity, and biomedical researchers, data analysts and clinicians are demanding advanced and user-friendly applications for real use in practice. In response, integrative platforms supporting most common AI operations in controlled research environments are becoming essential for spreading the use and development of AI applications in an unbiased and reproducible manner. These environments aim to **reduce entry barriers for individuals** to start experimenting with AI through the automation of machine learning and data science tasks. The platform should provide functionalities like intuitive user interfaces, logging and provenance facilities, experiment repeatability and traceability, tools for data life-cycle management, etc. EuCanImage's tools integrated in the computational platform will permit

<sup>1</sup> AI virtual research environment, <https://gitlab.bsc.es/inb/eucanimage/vre>



the use, development and sharing of algorithms as well as the end-to-end training, development and deployment process of models.

Through the research platform, researchers should also be able to identify and **access relevant data** from the unified EuCanImage catalogue of multiple data types, including cancer imaging, biological data and health records. Stored in associated repositories like the European Genome-Phenome Archive<sup>2</sup> (EGA), EuroBioImaging<sup>3</sup> (EuBI), or Biobanking and Biomolecular Resources Infrastructure<sup>4</sup> (BBMRI-ERIC), datasets are required to be accessible for being imported to the user's workspace at the platform for further analysis. Hence, specific connectors with EuCanImage data infrastructures are required, either to central servers or to in-house distributed nodes, if a federated scenario was proposed. Data connectors should consider the authentication and authorization protocol of the primary data repository, an efficient and location-aware data transfer, and a secure management of access credentials.

Another key aspect when providing a computational framework for AI is to offer the **maximum flexibility** in the combination of heterogeneous software deployments that might be eventually integrated. Leveraged AI tools should be modular and autonomous parts compiled and maintained by AI developers, who might easily modify them, for instance, when new data is released. If lightweight wrapping adaptors are used to make tools pluggable into the platform, the bundled AI software will remain unaltered after the integration, enabling a rapid and easy deployment phase and promoting components' reuse. Furthermore, adaptors should permit the virtualization of the AI applications as software containers, increasing in this way the isolation and portability of tool boxes. Analogously, offering a fully virtualized computational platform is a good strategy that permits building a portable infrastructure ready to be deployed in additional computational facilities. Like so, the platform could be installed next to data providers locations following a federated-like scenario, where data transfer needs are minimized and data-privacy regulations easily reconciled.

And last but not least, **a scalable and adaptable computational backend** is required to satisfy the increasing heterogeneity of hardware architectures, accelerators, and decentralized programming models supported by the AI tools ecosystem. Without aiming for a specific AI-accelerated hardware cluster, the flexibility provided by a cloud-based infrastructure could fulfill the demands of machine learning and deep learning workloads, especially when the platform maximizes cloud computing capabilities, for instance, making use of auto-scalable cloud provisioners and schedulers for an optimal resource exploitation.

## 1.2 Background

AI-VRE is based on already existing software components that were adapted to the specific needs of EuCanImage. The infrastructure layout is designed as an evolution of the EuCanSHare<sup>5</sup> computational platform, a cloud-based analysis platform for multi-center cardiovascular data analysis.

The basis of both research infrastructures is **openVRE**<sup>6</sup>, an open-source analysis workbench for the rapid deployment of customized analysis infrastructures. Designed according to the

---

<sup>2</sup> EGA, [www.ega-archive.org](http://www.ega-archive.org)

<sup>3</sup> EMC, [www.eurobioimaging.eu](http://www.eurobioimaging.eu)

<sup>4</sup> BBMRI, [www.bbmri-eric.eu](http://www.bbmri-eric.eu)

<sup>5</sup> EuCanSHare, [www.eucanshare.eu](http://www.eucanshare.eu)

<sup>6</sup> openVRE, <https://github.com/inab/openVRE>



one-stop-shop philosophy, the framework includes a complete web-based virtual research environment with integrated access to local and remote datasets, data visualization applications, and the necessary adaptors and cloud resource schedulers to execute any of the plugged-in domain-specific analysis tools as native software-as-a-service applications in one click.

## 2 Conceptual design

The main goal of AI-VRE is to offer a flexible **computational baseline** on top of which the AI-driven tools developed by EuCanImage consortium members will be integrated. In this way, the AI platform is going to support the development of novel AI models from large-scale imaging and non-imaging data.

The components of the platforms are orchestrated as virtual machines working in the protected network of an *in-premises* cloud infrastructure. Figure 2 depicts the overall AI-VRE layout. User authentication service is delegated to an external identity provider (openID-connect2 compliant), permitting when configured, user federation or single-sign-on with other EuCanImage services. Users interact with the platform through two complementary interfaces, a web-based virtual research environment (VRE), and a REST API for programmatic access. The computational backend is in charge of processing analysis job requests initiated from user's interfaces. It is composed of software schedulers (PMES/COMPSs and OGS) that launch and manage job executions on the cloud resources, which in turn are on-demand provisioned either enacting on-the-fly a new virtual machine via OCCI API, or autoscaling pre-allocated virtual machines instances using a dynamic provisioner service (oneFlow). Either way, the execution takes place in a single or a cluster of virtual machines where the AI application has been installed, system-wide or using software containers, and where openVRE has enabled the access to the datasets relevant for the execution. [Section 3](#) describes in detail the functioning of individual components.

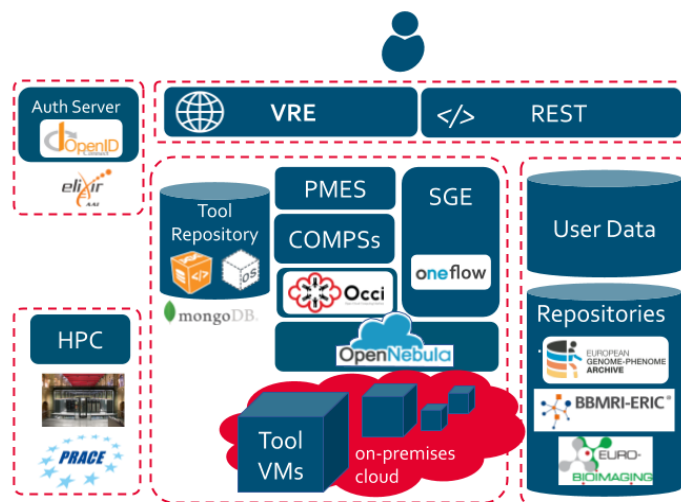


Figure 2: AI-VRE general layout



## 2.1 Requirements

The proposed layout is the result of adopting a selection of technical solutions that fulfill the design requirements of an AI development platform fitting the EuCanImage ecosystem. The following table summarizes them:

Table 1: AI-VRE requirements and adopted solutions.

Design Requirements	Technical Solution
Flexible and modular design to guarantee an easy incorporation of new services and data sources	All components will be assembled/developed as independent modules, implemented in <b>virtualized systems</b> , using the appropriated data communication channels
Integration of advanced authentication and authorization services able to centrally manage fine-grain data access control on federated EuCanShare resources.	The central authentication services will be based on <b>OpenID connect (OIDC)</b> servers identity providers. All data communications and API accesses will be controlled using the OAuth2 protocol. Authorization services will be initially kept at their original sites
Portable and versatile cloud-based computational infrastructure that permit to conveniently integrate heterogeneous software components while providing scalable compute resources on-demand.	The platform will be built on top of a openVRE system, relying on <b>OCCI compliant</b> cloud managers like OpenNebula or OpenStack.
Software scheduler(s), able to manage analysis workflows, and computational resources in a transparent and adaptable manner. This will be an elastic infrastructure with automatic adaptation to user loads.	Two initial software schedulers, available at openVRE will be used. Open Grid Engine will be used for applications requiring stable computational needs, while PyCOMPS/PMES will be used when computational needs may depend on the specific analyses.
Data storage solutions able to grow on demand, with a fast data mobilization infrastructure, and providing the necessary data synchronization capabilities among data providers and analysis	Lower level storage system will be based on NoSQL <b>MongoDB</b> database manager. It provides the necessary characteristics of stability, ability of growth, and horizontal scalability.

## 3 Platform components

The following section describes individually the software components used in the initial installation and their specific function.

### 3.1 Cloud infrastructure

AI-VRE is natively designed to be hosted on top of a cloud infrastructure, as the platform makes use of cloud or hybrid computing for executing analysis tools in a flexible and scalable manner. Compatible cloud middlewares are those compliant with the **Open Cloud Computing Interface<sup>7</sup> (OCCI)**, a popular open cloud standard adopted but open-source stack like OpenNebula<sup>8</sup> or OpenStack<sup>9</sup>, or private vendors like Amazon EC. It permits the

<sup>7</sup> OCCI, <http://occi-wg.org>

<sup>8</sup> OpenNebula, <http://www.opennebula.org>

<sup>9</sup> OpenStack, <https://www.openstack.org/>



remote management of virtual machines (VMs), and other cloud resources like virtual networks or storages via cloud orchestration services. openVRE uses them to boot the VMs where AI toolboxes are installed (see below [PMES: Programming Model Enacting Service](#)).

Alternatively, openVRE can be distributed fully virtualized through KVM-enabled virtual machines. Thus, the framework is easily deployable on top of other high-end computing infrastructures able to meet the computational capabilities to power up AI applications. However, under this scenario, the elasticity provided by cloud-based applications would be lost, yet the regular cluster-based execution provided by [OGS: queuing system](#) would be available.

Currently, the pilot installation of AI-VRE runs on top of an ELIXIR-ES 'on-premise' cloud, hosted at the Barcelona Supercomputing Center (BSC) and based on OpenNebula.

Table 2: Pilot installation dedicated resources.

CPUs	RAM	Storage
18 CPUs - CPU model: QEMU/KVM virtualization Host passthrough - Host CPU: Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz	32 Gb	500Gb GPFS network storage

### 3.2 Research Environment

AI-VRE exposes the users to a **fully intuitive** Virtual Research Environment (VRE) while job-based executions are transparently managed on the underlying cloud infrastructure. The VRE is an integrative and user-friendly web-based interface for bioinformatics data analysis that gives support to the common analysis flow: (i) users connect to the environment, (ii) upload or import the datasets, which (iii) are feed into the selected analysis tool or pipeline, (iii) send to the backend for batch execution, and (iv) when results are ready, they become available to the user for being visualized or further processed using another tool. This section describes the components required to achieve it.

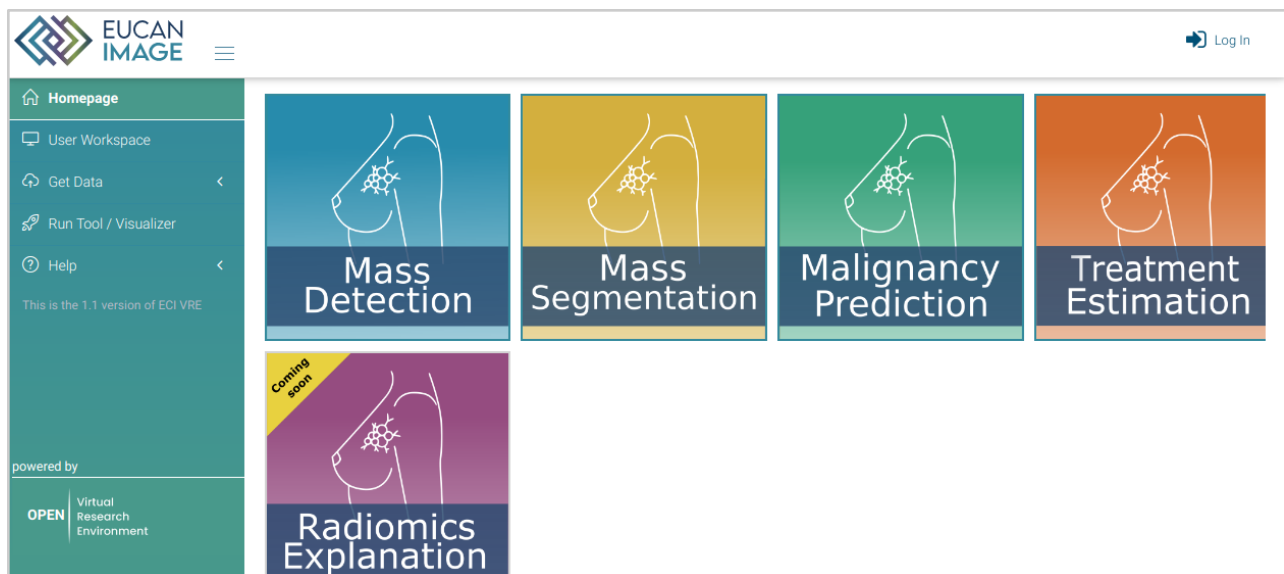


Figure 3: AI-VRE home page. Boxes represent the available analyses





## Authentication

AI-VRE should assure complete data privacy with respect to users' data and activities. To this end, interactive or programmatic access is made using an encrypted channel (https, ssh), and users are authenticated on every access.

AI-VRE connects to the authentication service of the Spanish National Institute of Bioinformatics (INB - ELIXIR-ES) for managing its users. The server is based on Keycloak, an open-source identity manager that implements **OpenID Connect 1.0 (OIDC)**. It supports the OAuth2 authentication flow for web access (based on username/password), and token-based authentication for the remaining services. The web portal displays and refreshes access tokens, so that the user is able to authorize himself to the publicly available services via REST. If configured, additional external OIDC identity providers (idPs) like Google, ORCID or ELIXIR AAI, might be accepted.

## Getting data

Users can populate the workspace in several ways:

- Data upload:
  - Direct Upload: File(s) from the user's local computer can be uploaded directly to the workspace through an HTTPS protocol.
  - Create files: A text editor is available to create simple plain text files. This is intended for data or metadata of reduced format that can be simply typed in.
  - External URL: given a URL (FTP/HTTP), the platform downloads the data into the workspace. This is the recommended procedure to include bulky data, as the procedure is performed in the background and no limit in size applies, being only limited by the user's quota available in their workspace

### - Data Import from Reference Data Repositories:

The platform is leveraging the access to the well-established **Euro-Bioimaging** infrastructure, as well as the **European Genome-phenome Archive**, storing biological and clinical phenotypic records. It will allow the development of AI solutions that integrate organ-level, molecular and other clinical predictors. From the platform, the user browses the remote content and imports it into the workspace with one click. The option is available for both, public and protected datasets, as AI-VRE establishes the connection on behalf of the logged-in user. The necessary credentials are configured in the user's profile section (figure 4) and securely stored in the platform, unless the repository permits a federated authentication system based on OIDC (see section [Authentication](#)), which could avoid the need of credentials' sharing.

Table 3: AI-VRE data downloaders from external repositories.

Repository	Content	Data transfer protocol	Access to protected data?	Authentication method
EGA	- biological and clinical phenotypic records - genomic datasets	SFTP (FUSE enabled)	Yes	SSH ed25519 key
EuBI	- bioimaging data	HTTPS (REST API)	Yes	EuBI key token



Both data infrastructures will be queried from AI-VRE using the common subject EuCanImage identification system for datasets being developed as part of T3.3. Adequately cross-linking cancer imaging to biological and health data is essential for building integrated AI models.

Personal Info Keys

LINKED ACCOUNTS

European Genome-phenome Archive (EGA)

User Name  
laia.codo@bsc.es

Crypt4GH Public key  
ssh-ed25519 AAAAC3NzaC11ZDI1NTE5AAAAI8FkkGewJqL1xq53dokymP9T1kz1Sja

Crypt4GH Private key

euro-Biolmaging

Do you have an euroBioImaging account? Link it and you'll have one-click access for all your euBI protected datasets, under [Get Data](#) • EuroBioImaging

[+ Link your account](#) [How to apply to euBI access?](#)

(a)

Central EGA outbox — Datasets

EGA datasets accessible through your EGA account

10 records

Dataset ID	Title	Creation time	Avail
EGAD5000000024	Dataset of Fastq files of three trio members	2020-04-08T14:45.920Z	false
EGAD5000000025	Dataset of VCF files of three trio members	2020-04-08T14:33.799Z	false

(b)

EuroBioImaging Your available biomaging Projects

Public projects

10 records

ID	Description	Project Name
stwstrategydr	This collection contains thorax CT scans of the "Rider Test-Retest" series. There are 32 unique cases denoted by a 10-digit identifier. The CT examinati	STW_STRATEGY_RIDER
stwstrategyps4	This collection contains phantom scans of a Gammex 467 CT phantom (Middleton WI USA) for radiomics intra-scanner testing due to X-ray tube	STW-STRATEGY-Phantom_Seri

(c)

Figure 4. Downloading imaging and non-imaging data from external data repositories. (a) User's profile view. (b) AI-VRE screenshot of the project selector page for EGA. (c) AI-VRE screenshot of the project selector page for EuroBioImaging

Regardless of the data source, when new data is uploaded or imported, AI-VRE gathers some metadata items. These include descriptor fields like data types (e.g. 'MRI Bioimage') and formats (e.g. 'DICOM') selected from a predefined list. Data types and formats enable the system to select the appropriate set of tools and visualizers usable with the uploaded files. Metadata for files obtained from installed tools are automatically obtained from the tools metadata manifest.

## Personal workspace

The personal workspace concentrates most of the user's activity. It contains a central table displaying the **user's datasets** and the operations available for each of them: eligible analyses, visualization options, file toolkit, etc. The workspace shows a file system layout where uploaded data and the analysis results of each execution are available under separated directories. Files can be filtered by any name, format, data type, or project. The page also displays the status of jobs in progress, recovering the results as soon as they are available.



File	File type	Data type	Execution	Date	Size	Actions
File	All	All	All			Clear filters
uploads			uploads	2021/09/20 15:37	118.00 M	cc
<input checked="" type="checkbox"/> P10_DATA.nii	NIFTI	Bioimage	uploads	2021/09/20 15:37	29.00 M	cc
<input type="checkbox"/> P10_MASK.nii	NIFTI	Bioimage Mask	uploads	2021/09/20 15:37	29.00 M	cc
<input checked="" type="checkbox"/> P1_DATA.nii	NIFTI	Bioimage	uploads	2021/09/20 15:36	30.00 M	cc
<input type="checkbox"/> P1_MASK.nii	NIFTI	Bioimage Mask	uploads	2021/09/20 15:36	30.00 M	cc
repository			repository	2021/09/23 12:00	179.58 M	cc
<input type="checkbox"/> SUBJECT001.zip	DICOM	Bioimage	repository	2021/09/23 12:00	179.58 M	cc
run007			run007	2021/09/21 10:09	72.00 B	cc
<input type="checkbox"/> patient_prediction.csv	CSV	Sample Info File	run007	2021/09/21 10:09	72.00 B	cc

Figure 5. Screenshot of a user's private workspace

The selection of a specific tool triggers the analysis configuration screen where users can assign the selected data files to the appropriate input parameters and arguments. After configuration, the job is sent to the VRE backend who orchestrates the cloud services in a transparent way.

### 3.3 Job process management

The openVRE backend is able to **process batch jobs** using (i) an auto-scalable queuing system based on OGS/Oneflow, or (ii) short-lived compute instances controlled by PMES/COMPSSs. The elastic queuing system successfully manages web application backends where the major requirement is to be able to deal with pick demands. Workflows that show a more complex structure where, for instance, computational resources should be adjusted at run time, are recommended to be configured using PMES/COMPSSs.

#### **OGS/Oneflow: auto-scalable queuing system**

Open Grid Scheduler (OGS) (former Sun Grid Engine, SGE) is designed to manage distributed software executions in heterogeneous computational environments. OGS is used normally in **cluster-based** infrastructures as a general process scheduler. To better exploit OGS features into a cloud-based environment like openVRE, an OpenNebula self-provisioning tool called oneFlow is added to the equation. oneFlow is able to automatically trigger the deployment/undeployment of VMs in front of monitored parameters, like the CPU workload of these VMs, the I/O stress or a certain network, etc.

Analysis tools are implemented as VMs under the control of OGS. Each VM packs an application with the corresponding OGS queue configuration. In the case of increased demand on a certain application (i.e. certain VM), oneFlow instructs OpenNebula to replicate such VM, which is translated into an increased number of hosts available in the queue system for such an application. Once the workload for such group of VMs decreases, OneFlow undeploys them one by one, always keeping at least one instance ready to accept new job petitions. In this way, allocated resources are dynamically and transparently adjusted on-demand.



## PMES/COMPs: short-lived compute instances

The Programming Model Enacting Service<sup>10</sup> (PMES) allows the platform to run a batch execution in a transient VM, even if the VM belongs to a remote cloud. When a job petition is initiated, the platform instructs PMES to (i) boot and contextualize the image packing the selected application, (ii) stage-in infiles there, (iii) run the application in the remote instance, (iv) stage-out results, and (v) stops the virtual machine. For accomplishing these steps, PMES interacts with the cloud provider (e.g. OpenNebula) through the use of the OCCI connector.

Optionally, further elasticity can be given to the system if the launched application is a pipeline parallelized using COMP Superscalar<sup>11</sup> (COMPSSs). COMPSSs is a programming model designed to ease the development of applications for distributed infrastructures (e.g. Clusters, Grids, and Clouds). COMPSSs runtime is able to discover parallelisms during the execution time and dynamically distribute the tasks. When a task requires extra resources, COMPSSs provisions new VMs that act as transient workers of a compute virtual cluster. In this way, pipelines are elastically executed, demanding the resources according to the specific needs of the execution.

### 3.4 Tools integration

The progressive introduction of AI tools over the time and the continuous update and redeployment of the same are part of the AI development platform life cycle. That's why the role of tool developers, *i.e.* AI researchers, is fundamental. The platform provides means to ease the integration protocol of new tools: (i) the adapter to connect openVRE with new AI applications requires very basic programmatic skills, and (ii), a specific developers' workspace is available at the web interface to register the adapted application with few clicks.

The integration basis it that all openVRE tools must respond to a standardized command line, *i.e.*, a uniform interface. To this end, an openVRE template tool<sup>12</sup> is provided, and developers only need to adapt it to execute the application or pipeline of interest. Hence, the openVRE Tool API acts as an adapter, that provides the common access interface and wraps the actual execution of the application, that remains untouched. It is formalized as a simple Python library, hence, the bundled application could correspond to any non-interactive batch execution: a second python subprocess, an R script run, a set of container-based executions (*i.e.* dockers, singularities), a full workflow execution (CWL, Nextflow), etc.

---

<sup>10</sup> F. Lordan et al., "ServiceSs: An Interoperable Programming Framework for the Cloud," J. Grid Comput., vol. 12, no. 1, pp. 67–91, Mar. 2014

<sup>11</sup> COMP Superscalar, an interoperable programming framework, SoftwareX, Volumes 3–4, December 2015, Pages 32–36, Badia, R. M., J. Conejero, C. Diaz, J. Ejarque, D. Lezzi, F. Lordan, C. Ramon-Cortes, and R. Sirvent, DOI: 10.1016/j.softx.2015.10.004

<sup>12</sup> openVRE tool API, <https://github.com/inab/openvre-tool-api>

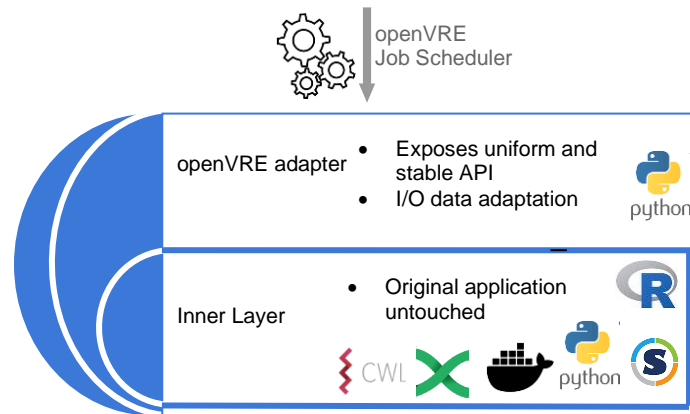


Figure 6. Layered schema of openVRE tools

Once the application is wrapped, the code ought to be delivered to the platform’s administrators, together with some extra information like tool’s descriptive metadata (i.e. title, keywords, etc), deployment details (i.e. CPUs, memory, tool main script path, etc), etc. This submit process is available through the web interface.

## 4 AI Toolbox

The computational framework presented here is meant to integrate the selection of AI solutions proposed by EuCanImage members. As a first step, a range of existing tools for radiomics-based analysis is being integrated. Being driven by the clinical use cases, the current toolbox is focused on the detection, segmentation and evaluation of deep learning and machine learning pre-trained models for breast cancer in mammography (MG) or Magnetic resonance images (MRIs). More details on the radiomics library being developed are going to be delivered as part of D5.2.

Table 3 summarizes the current content of the toolbox, which is planned to grow and evolve at the pace of new AI applications and case studies. Due to the availability of public open-access datasets on breast cancer screening, the first three tools were initially implemented for (i) mass detection (tool #1), i.e., localisation of masses from high-resolution MG images; then (ii), the detected regions of interests are subsequently passed to the segmentation application (tool #2) that delineates the mass boundaries; and (iii) malignancy prediction model that classifies the regions of interest into benign or malignant categories (tool #3). These tools correspond to the crucial steps towards accomplishing the tasks highlighted by the clinical uses-cases, in particular, for early detection, diagnosis, and radiomics analysis in screening MGs. Furthermore, a treatment response estimation tool (tool #4) has also been integrated within the VRE. In contrast to the MG based imaging approach, this predictor assesses radiomics features (texture and shape) based on statistical information derived from volumetric MRI scans of patients diagnosed with cancer. The goal of developing such a tool consists of predicting whether a patient will reach a partial, complete, or no pathological response after neoadjuvant chemotherapy. An accurate treatment response model will guide the clinicians if a particular patient requires a special treatment (e.g., more aggressive treatment, or bigger dose) or additional clinical assessment.



Table 3: List of available tools at AI-VRE

	Analysis Tool	Author	Status
#1	Breast MG Mass Detection	UB	Public
#2	Breast MG Mass Segmentation	UB	Public
#3	Breast MG Benign/Malign Classification	UB	Public
#4	Treatment response estimation from MRI cancer tumors	FORTH Institute	Public

Next steps is the inclusion of novel integrative machine learning techniques to build AI predictive models that integrate imaging and non-imaging information, including support vector machine, multiple kernel learning and deep learning networks (T5.2, T5.3).

## 5 Towards a Federated infrastructure

According to the EuCanImage Data Management Plan (D3.1) both, centralized and distributed approaches for data management are going to be used within EuCanImage. EuCanImage facilitates centralized platforms for storage, curation and AI support. Alternatively, and depending on the regulations and availability of local (IT) support of data providers, *i.e.* clinical centers, a distributed approach is going to be required. Under this scenario, protected datasets do not trespass data providers' premises. The full virtualization of AI-VRE and the use of open cloud standards for interoperability (OCCI) largely ease the installation of the platform in multiple cloud infrastructures. However, a major portability could be achieved if the framework is containerized in the same way the individual analysis tools are.

One step further is not only moving analyses where data is located but building an overall federated network for avoiding data scarcity and fragmentation. Privacy-preserving federated learning (FL) frameworks are becoming of extensive use, particularly when dealing with biomedical data. The integration of one of these libraries as part of the AI-VRE could be of great interest as data scientists and clinicians would keep using the same intuitive and familiar research environment when participating in FL studies, than when running any other AI tool in EuCanImage. Some open-source libraries like pySyft<sup>13</sup>, Vantage6<sup>14</sup>, or Flower<sup>15</sup> are being explored to this end as part of a clinical use case for multi-image classification on Breast Cancer MG.

### 5.1 Roadmap

The following points summarize the development plan until task completion (M18).

---

<sup>13</sup> Ziller A. et al. (2021) PySyft: A Library for Easy Federated Learning. Federated Learning Systems. Studies in Computational Intelligence, vol 965. Springer, Cham. [https://doi.org/10.1007/978-3-030-70604-3\\_5](https://doi.org/10.1007/978-3-030-70604-3_5). <https://github.com/OpenMined/PySyft>

<sup>14</sup> Moncada-Torres A. et al. (2020) VANTAGE6: an open source priVAcY preserviNg federaTed leArninG infrastruCTurE for Secure Insight eXchange. AMIA Annual Symposium Proceedings, 2020, p. 870-877.

<sup>15</sup> Daniel J. Beutel, et al. (2021) Flower: A Friendly Federated Learning Research Framework. <https://arxiv.org/abs/2007.14390>



- Adaptation of AI-VRE to a container-based system
  - During an initial design phase, the individual modules of the containerized platform are going to be defined, as well as the communication requirement among them, and the necessary data volumes. At the same time, different container engines and the corresponding orchestrators are going to be considered.
  - According to the designed layout, a set of images implementing the openVRE core are going to be built. These will include a web server, the VRE front-end, a mongoDB server, and the selected VRE job scheduler
  - Analogously, a set of software container images containing AI-VRE tools are going to be built. For reusability, a base VRE tool image will be first implemented, to later be adapted to the distinct AI-VRE tools.
- Exploration of a federated learning framework as part of AI-VRE
  - A comparative analysis of a selection of open-source FL solutions will be conducted taking into account the requirements imposed by the clinical use-case under study, and needs of the cloud-based computational infrastructure.
  - The selected federated learning library will be deployed in the ELIXIR-ES cloud infrastructure as part of the AI platform.