Release of MIP8.0 (SC5)
(D4.16- SGA3)

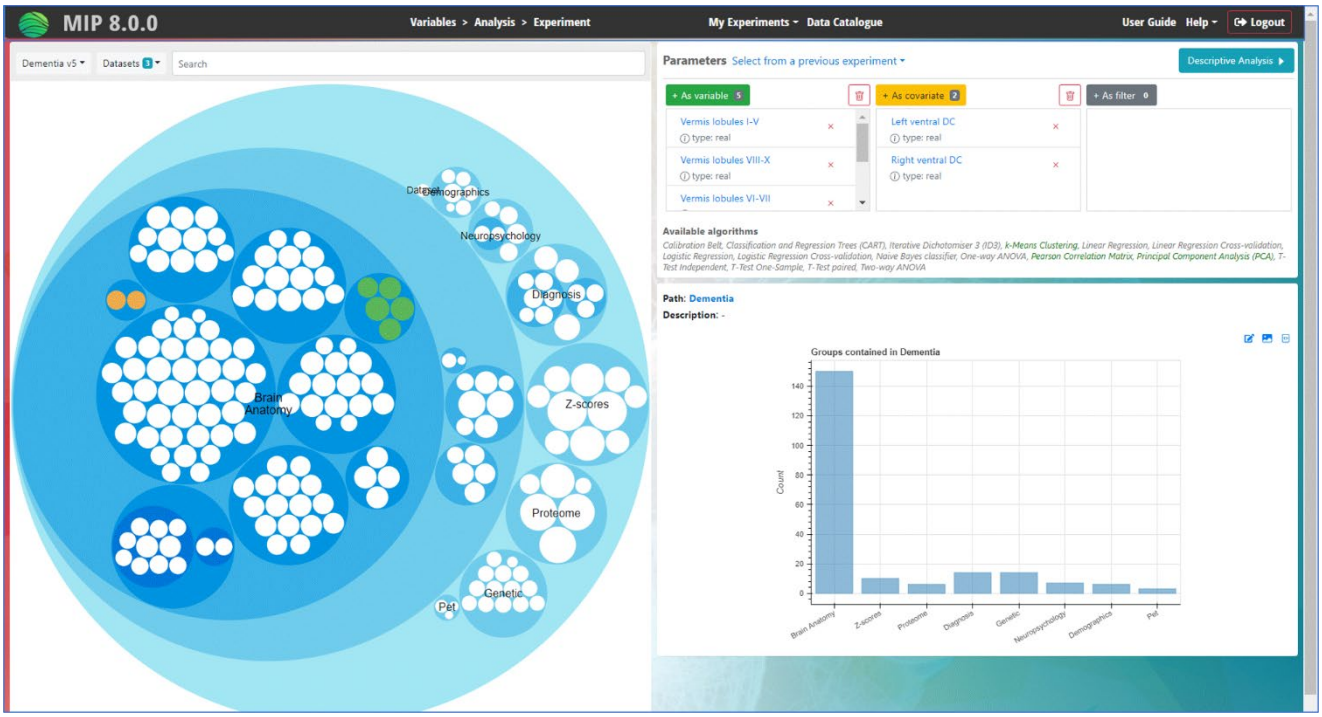

Figure 1: MIP 8.0.0 User Interface

MIP 8.0.0 User Interface, reflecting the current representation of data model, experimental parameters, and visualisation of data.

Project Number:	945539	Project Title:	HBP SGA3

Document Title:	D4.16 release of MIP 8.0.0 (SC5)
Document Filename:	D4.16 (D47) SGA3 M42 SUBMITTED 230927.docx
Deliverable Number:	SGA3 D4.16 (D47)
Deliverable Type:	Report
Dissemination Level:	PU = Public
Planned Delivery Date:	SGA3 M42 / 30 SEP 2023
Actual Delivery Date:	SGA3 M42 / 27 SEP 2023
Author(s):	Birgit SCHAFFHAUSER, Georges MELISSARGOS, CHUV (P27), Thanasis KARABATSIS, ATHENA RC, (P133)
Compiled by:	Birgit SCHAFFHAUSER, CHUV (P27)
Contributor(s):	Manuel SPUHLER, Viktoria BASTIC-SCHMIDT, Pauline DUCOURET, CHUV (P27), contributed to Sections 4 and 5 Evita MAILLI, Maria - Olympia KATSOULI, ATHENA (P133) contributed to Sections 4 and 5 Iosif SPARTALIS, AUEB (P4), contributed to Section 4 and 5
WP QC Review:	Martha Elisabeth BRIGG, UIO (P81)
WP Leader / Deputy Leader Sign Off:	Jan BJAALIE, UIO (P81)
T7.4 QC Review:	N/A
Description in GA:	New release of MIP with strengthened data privacy and safety, optimised data factory and front-end, and novel federated algorithms.
Abstract:	Report complementing Medical Informatics Platform Release 8.0.0
Keywords:	Federated infrastructure, release 8.0.0, software, data model, data management, medical data, data privacy, Kubernetes, federated analysis, Secure multi-party computation (SMPC), Differential Privacy (DP)
Target Users/Readers:	Clinical researchers, developers, data managers, technical deployment and support staff, computational neuroscientists, HBP/EBRAINS users

Table of Contents

Table of Figures

1. Summary

This deliverable describes the final release of the Medical Informatics Platform (MIP), MIP8.0, within the Human Brain Project (HBP), SGA3. It is a document accompanying the actual release of the product and summarises where the developments stand, what the final offer entails and indicates, what is foreseen for its future beyond the HBP.

2. Introduction

The Medical Informatics Platform was developed over several periods of the Human Brain Project, in the last phase of the HBP being part of the Medical Data Analytics service category. The MIP is providing sophisticated technical solution for federated medical data analysis and clinical research to the project. In the context of the EBRAINS research infrastructure, it is one of three GDPR-compliant health research platforms (https://www.ebrains.eu/health-research-platforms/health-platforms/work-with-health-data-2) with specific functional scopes tailored to diverse user communities. In the case of the MIP this is clinicians and clinical scientists.

The MIP represents a software bundle that is distributed as a secure virtual appliance and functions as a managed service. It places a strong emphasis on safeguarding privacy and provides solutions for federation of harmonised data between hospitals or research institution. The solution encompasses a holistic approach to addressing legal, ethical, and data security concerns while preserving data privacy.

The MIP is particularly esteemed for its network of established partners and its substantial potential in unlocking data that was previously confined within the boundaries of national or hospital systems. This liberated data can be harnessed for new research endeavours and collaborative data analytics, thus valorising existing data through data re-use.

Moreover, the MIP offers a diverse array of algorithms, including algorithms that support machine learning, which are readily accessible as pre-configured tools in this open-source solution.

3. Preamble

This deliverable serves as a comprehensive index, housing an array of documents and resources pertinent to the final version of the MIP.

Designed to cater to developers, technical deployment and support teams, and individuals with a profound technical interest in the MIP's inner workings, our primary objective is to streamline access to crucial information essential for understanding the current state of the MIP.

Our main MIP repository is geared to provide ample materials for qualified personnel to grasp the intricacies of the MIP, facilitating their ability to develop, deploy, and operate this platform effectively.

4. MIP 8.0.0 Release Major Updates

4.1 Introduction of Differential Privacy

In the pursuit of enhanced data security and confidentiality, MIP 8.0.0 provides a first integration of differential privacy, a feature facilitated by the Secure Multiparty Computation (SMPC, https://en.wikipedia.org/wiki/Secure_multi-party_computation) cluster, providing a global differential privacy guarantee for federated analytics, which gives a better utility for the same privacy budget compared to the local differential privacy alternative. This pioneering addition assures cryptographic guarantees that local, partial results remain obfuscated. This pioneering addition assures cryptographic guarantees that local, partial results remain obfuscated. This feature

is available as an opt-in deployment configuration option, permitting meticulous control over the enforcement of differential privacy within the SMPC pipeline.

4.2 Enhancements to the Analysis Engine Exareme2

In line with our commitment to keep improving the platform`s capabilities, MIP 8.0.0 brings significant enhancements to the Exareme2 engine. Consequently, the Galaxy workflow engine is now deprecated and has been entirely removed from the MIP ecosystem.

4.3 New Algorithmic Approach

MIP 8.0.0 adopts a new algorithmic approach, where each node autonomously generates its own model. These models are subsequently aggregated using the federated averaging technique (https://www.educative.io/answers/what-is-federated-averaging-fedavg), promising optimised performance and efficiency.

4.4 Algorithmic Advancements

This release introduces new analysis algorithms, including Naive Bayes (https://en.wikipedia.org/wiki/Naive_Bayes_classifier) with integrated cross-validation for nominal and numerical Categorical Data Elements (CDEs) k-MEANS clustering (https://en.wikipedia.org/wiki/K-means_clustering). Additionally, Support Vector Machine (SVM) is now available, leveraging the federated averaging technique to further enhance model accuracy and efficiency. A full catalogue of all the algorithms implemented in MIP 8.0.0 and their related documentation can be found here: https://github.com/HBPMedical/mip-docs/blob/master/algorithms.md. Finally, one new capability is added in the form of now handling and analysing Longitudinal datasets.

4.5 Kubernetes Monitoring, Log Search, Grafana Visualisation

MIP 8.0.0 extends its system monitoring and reporting capabilities by employing the use of related Kubernetes (microk8s, https://microk8s.io/) packages. More specifically, Prometheus is configured for monitoring the containers across the MIP Federations, Grafana for visualising the resources used from the container pods and nodes in a dashboard and finally, Loki for aggregating and searching the platform's logs of events.

4.6 Further Enhancements and Bug Fixes:

MIP 8.0 incorporates a series of additional improvements, spanning aspects related to security fortification, system stability, operational efficiency, and advanced monitoring capabilities.

5. MIP 8.0.0 Components

Overview of relevant components and essential tools within the MIP ecosystem.

5.1 Core Platform Components:

- The "Deployment" component of the MIP enables the preparation of the hosting environment and the actual installation of the MIP, both for a local MIP instance, not connected to a federation

and the federation set-up with a central node and its remote nodes hosted by the member Hospitals of that Federation. https://github.com/HBPMedical/mip-deployment An overview, depicting the high-level architecture of the local versus federated MIP set-up can be seen in Figure 2: MIP 8.0.0 high-level architecture.

- The "**Frontend**" is a web-based application that offers the graphical user interface (GUI) to the users of the MIP and provides user access to most of the functionalities of the platform. https://github.com/HBPMedical/portal-frontend
- The "**Backend**" component implements functionalities supportive to the "Frontend" like a local user authorisation/authentication service and implementing analysis workflow elements like the "Experiments". https://github.com/HBPMedical/portal-backend
- The "**Gateway**" is the component that acts as the functional interface (via its associated API) between the "Frontend" and the federated analytics engine. https://github.com/HBPMedical/gateway
- The "**Exareme2 Engine**" (release 0.19.0) is the latest federated analysis engine implementing statistical and ML algorithms and is supporting the extra security layer of SMPC and DP. https://github.com/madgik/Exareme2

5.2 Additional Stand-alone Tools of the MIP Ecosystem:

- The **Data Quality Control Tool** (MIP-DQC): a software for exploring, validating, and transforming datasets based on CDEs before uploading to the MIP. It has both a Command Line Interface (CLI) and a Graphical User Interface (GUI). https://github.com/HBPMedical/DataQualityControlTool/tree/v.4.1
- **Data Catalogue**: A web-based application to explore and manage MIP Data Models and pre-defined Common Data Elements (CDEs). It handles the transformation from .csv to JSON format (metadata file) , the required data format for data upload on the MIP. https://github.com/HBPMedical/DataCatalogue
- **MIPMap engine**: a tool that provides Schema mapping https://github.com/HBPMedical/MIPMap
- **MRI neuromorphometric pipeline:** Pre-processing pipelines for MRI neuroimaging. https://github.com/HBPMedical/mri-parallel-nmm-pipeline
- **mip_schema**: Open-source Python package to manipulate Common Data Elements Metadata Schema for the MIP. https://github.com/HBPMedical/mip_schema
- **mip-dmp**: Python tool with Graphical User Interface to map datasets to a specific Common Data Elements (CDEs) metadata schema of a federation. https://github.com/HBPMedical/mip-dmp
- **pytest-mip**: Tool designed to assess the operational status of active MIP federations. https://github.com/HBPMedical/pytest-mip

5.3 Federated Analysis Algorithms

Within the MIP, privacy-compliant capabilities stand as a principal concern, resonating uniformly across the entire spectrum of analytical functionalities offered by the platform. Each distinct analytical methodology may necessitate specialised handling to methodically ensure the attainment of the essential level of compliance. This requisite arises due to nuanced differentiations in the operational workings inherent to each algorithm. The scientific and statistical validation procedures applied to these analytical methodologies are subjected to rigorous analysis. These validation measures were conducted through the implementation of tailored, internally designed tests, executed over controlled datasets. Furthermore, compliance validation extended across a diverse array of parametrisation settings and scenario simulations.

For convenience, documentation on the existing algorithm federation approach, unit tests, as well as information related to creating a new algorithm can be accessed at the following location: https://github.com/madgik/exareme/tree/24.3.0/Exareme-Docker/src/mip-algorithms.

5.4 Documentation

- **Technical documentation**: The main repository provides a broad overview of MIP releases and associated components. https://github.com/HBPMedical/mip-docs
- **User documentation:** The MIP provides a built-in interactive user guide that can be accessed by users at any step of the of their data analysis. A static version of the user guide is provided in the home repository of the MIP at https://github.com/HBPMedical/mip-docs/blob/master/docs/mip_guide_updated.pdf
- **Data Management documentation**: To enable users to prepare and harmonise their data for the use in the MIP, three essential guides are provided to support users.
- **Data Management Guideline**: General guidelines for data management on the MIP. https://github.com/HBPMedical/mip-docs/blob/master/MIP_Data_management_documentation.md
- **Data Quality Control Tool Guide**: provides detailed explanations and technical details about the Data Quality Control features available for the MIP. https://github.com/HBPMedical/DataQualityControlTool/wiki
- **Data Catalogue Guide:** provides detailed explanations and technical details about the Data catalogue features available for the MIP, like visualisation of Common Data Elements (CDEs or data models) with version control. https://github.com/HBPMedical/DataCatalogue

6. MIP8.0.0 High Level Architecture

The schematic architecture illustration Figure 2: MIP 8.0.0 high-level architecture Figure 2: MIP 8.0.0 high-level architecture, showcases the deployment stack of the MIP, enabling a versatile dual mode of operation for hospitals. In the MIP Local mode, hospitals can harness analytical capabilities within their individual boundaries, while the MIP Federated mode empowers federated analysis across anonymized data, transcending multiple hospital settings.

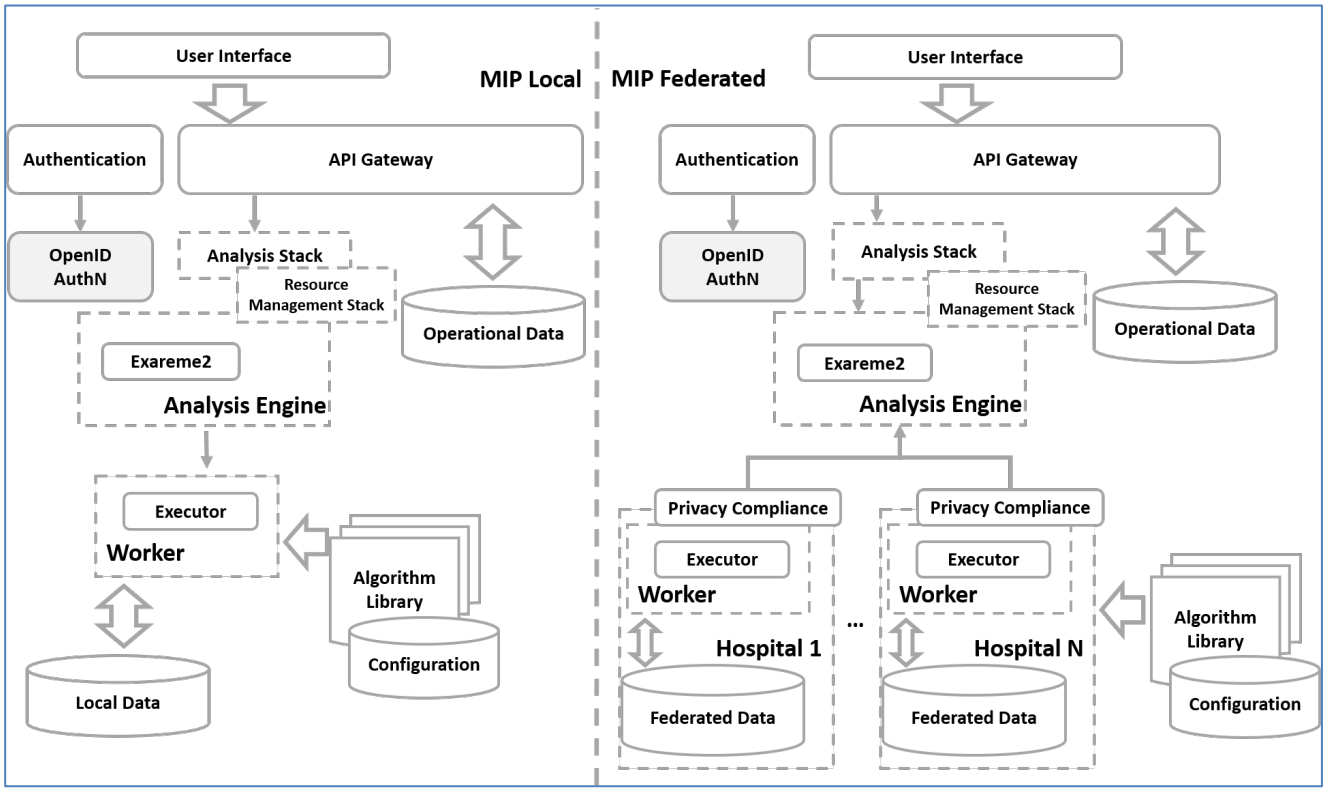

Figure 2: MIP 8.0.0 high-level architecture

High-level architecture of the MIP 8.0.0 provides a visual representation of the architectural framework and the processing flow, elucidating the distinct modes of analysis (local versus federated) available within the MIP ecosystem.

7. Looking Forward

- Review the communication architecture between the federated nodes and the central node of a MIP Federation for improving the efficiency of the deployment and operation of the MIP. Investigate direct container to container communication via a Kubernetes VPN.
- Update and redeploy the MIP 8.0.0 version in existing Federations.
- Expand the use of the MIP by organising new Federations in more medical domains, beyond Dementia, Epilepsy, Stroke, etc. Invite more members for existing and future Federations and thus enlarge the user community.
- Integrate more analytical algorithms, including more machine learning methods as per feedback from user community, especially as the MIP expands into more Federations.
- Improve/add visualisation methods both for depicting the results of the applied analytical algorithms but also for exploring the datasets for patterns.
- Implement the capability to create more complex "Experiments" on the datasets, applying a series of algorithms.