SAFEXPLAIN

Safe and Explainable
Critical Embedded Systems based on AI

PhIMG0003 Inference Management Guideline

Version 2.0

Documentation Information

Contract Number	101069595
Project Website	www.safexplain.eu
Contratual Deadline	31.03.2024
Dissemination Level	SEN
Nature	R
Author	Javier Fernández
Modified by	Lorea Belategi
Reviewed by	Irune Agirre
Approved by	Irune Agirre
Keywords	DL, Functional Safety, Inference Management

This project has received funding from the European Union's Horizon Europe programme under grant agreement number 101069595.

Table of Contents

1 Review / Modification History

Version	Date	Description Change
V2.0	15/02/2024	Changes Applied as a result of TÜV Review 2024-01-19
V1.0	04/12/2023	First version after complete internal review
V0.2	01/12/2023	Modifications and improvements
V0.1	21/11/2023	First draft

Note: Since AI-FSM utilizes templates from both the traditional FSM and its own templates, this annex distinguishes the AI-FSM templates by color-coding them in orange and the traditional FSM templates in green. Additionally, the files' names created from the templates are written in italics and underlined. It is worth mentioning that all the templates' names are preceded by "REF_" which should be changed to reflect the specific safety project reference.

2 Objective

The purpose of this document is to provide guidance for the Inference Management phase. This phase can be broken down into five primary steps, as illustrated in Figure 1. In that figure, the blue rhombuses represent input from the Data Management phase, corresponding to the verification dataset (Rhombus labelled with the number 2).

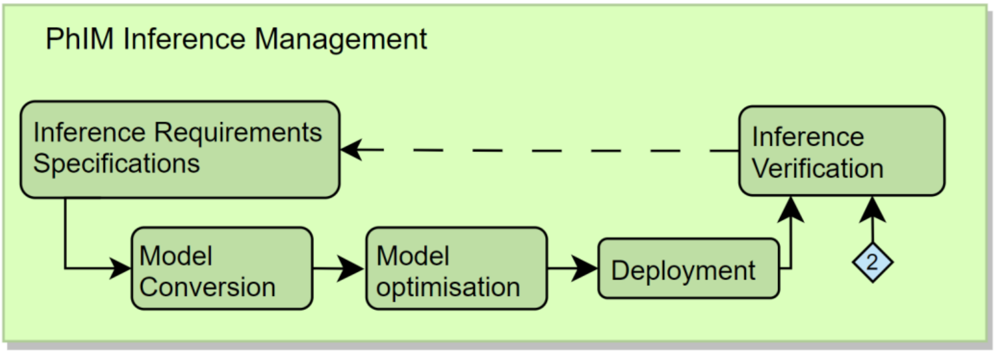

Figure 1. Learning Management phase

3 Scope

This guideline applies to all information relative to the Inference Management phase.

4 Introduction

During the inference management phase, the following documents are generated:

1. *REF_PhIMD0001_Inference_Requirements_Specifications.docx.* This document collects the data requirements specifications refined from the Deep Learning (DL) requirements specifications previously defined in phase 2.
2. *REF_PhIMD0002_Inference_Requirements_Specifications_IR.xlsx*. Internal review document to be checked after completing *REF_PhIMD0001_Inference_Requirements_Specifications.docx.*
3. *REF_PhIMD0003_Model_Conversion_Log.docx.* Document collecting the information relative to the process of converting the model from training to inference.
4. *REF_PhIMD0004_Model_Conversion_Log_IR.xlsx*. Internal review document to be checked after completing *REF_PhIMD0003_Model_Conversion_Log.xlsx.*
5. *REF_PhIMD0005_Model_Optimization_Log.docx.* Document collecting the information relative to the process of optimizing the model.
6. *REF_PhIMD0006_Model_Optimization_Log.xlsx*. Internal review document to be checked after completing *REF_PhIMD0005_Model_Optimization_Log.xlsx.*
7. *REF_PhIMD0007_Inference_Requirements_Verification_Tests.docx*. Inference requirements tests encompass a set of metrics to assess whether the inference requirements specifications have been fulfilled, the test definitions, and their corresponding outcomes.
8. *REF_PhDMD0008_Inference_Requirements_Verification_Tests_IR.xlsx.* Internal review document to be checked after completing *REF_PhIMD0007_Inference_Requirements_Verification_Tests.xlsx.*

Additionally, the following artifacts must be generated and stored:

1. Converted Model. The initial model undergoes a conversion process to transform it into a format suitable for deployment or compatibility with a specific target inference platform.
2. Optimized Model. Following the conversion, the model may undergo optimization to enhance its performance, reduce its size, or adapt it for resource-constrained environments. Optimization aims to maintain or improve the model's accuracy while making it more efficient for deployment.
3. Verified Inference Model. The final outcome is the verified inference model, which has undergone a comprehensive verification process. This involves checking the optimized model (or the converted model in cases where the optimization step is not performed) against specified criteria to ensure that the model adheres to the inference requirements specifications.

Table 1 presents the inputs and outputs associated with each step of Inference Management phase, which will be elaborated in the subsequent sections: Section 5 guides the development of the inference requirements specifications. Sections 6 and 7 guide model conversion and optimization steps, respectively. Sections 8 and 9 relates to the deployment and the inference verification steps. Finally, Sections 10 and 11 collect the acronyms and bibliography relevant to this document, respectively.

Table 1. Inputs and outputs of each step of the inference stage

Phase	Step	Inputs	Outputs	Corresponding templates
PhIM Inference Management	*Inference Requirements Specifications*	*REF_Ph2D0001_DL_Requirements_Specifications* *REF_PhLMD0001_Learning_Requirements_Specifications*	*REF_PhIMD0001_Inference_Requirements_Specifications* *REF_PhIMD0007_Inference_Requirements_Verification_Tests*	*PhIMT0001_Inference_Requirements_Specifications* *Ph0T0009_Test_definition_and_results_template*
		REF_PhIMD0001_Inference_Requirements_Specifications *REF_PhIMD0007_Inference_Requirements_Verification_Tests*	*REF_PhIMD0002_Inference_Requirements_Specifications_IR* *REF_PhIMD0008_Inference_Requirements_Verification_Tests_IR*	*REF_PhIMD0002_Inference_Requirements_Specifications_IR* *Ph0T0009_Test_definition_and_results_template_IR*
	Model Conversion	*REF_PhIMD0001_Inference_Requirements_Specifications* Verified Learning Model	*REF_PhIMD0003_Model_Conversion_Log* Converted Model	*PhIMT0002_Model_Conversion_Log*
		REF_PhIMD0003_Model_Conversion_Log	*REF_PhIMD0004_Model_Conversion_Log_IR*	*PhIMT0002_Model_Conversion_Log_IR*
	Model Optimization	*REF_PhIMD0001_Inference_Requirements_Specifications* Converted Model	*REF_PhIMD0005_Model_Optimization_Log* Optimized Model	*PhIMT0003_Model_Optimization_Log*
		REF_PhIMD0005_Model_Optimization_Log	*REF_PhIMD0006_Model_Optimization_Log_IR*	*PhIMT0003_Model_Optimization_Log_IR*
	Inference Model Verification	*REF_PhIMD0007_Inference_Requirements_Verification_Tests* Optimized Model or Converted Model [1] Verification dataset	*REF_PhIMD0007_Inference_Requirements_Verification_Tests* Verified Inference Model	*Document previously generated*

[1] If the model is optimized, the optimized model will be used as input; otherwise, the converted model will be utilized.

5 Inference Requirements Specifications

Regarding the inference requirements specifications, the *REF_PhIMD0001_Inference_Requirements_Specifications.docx* document stored in the "*PhIM Inference Management*" folder must be completed. It provides a set of instructions and recommendations for defining the requirements.

After defining the inference requirements specifications, *REF_PhIMD0007_Inference_Requirements_Verification_Tests.docx* must be completed. This document collects and defines the mechanisms or tests that must be performed to verify the inference requirements specifications. It must explicitly define at least one verification activity related to each data requirement. Additionally, this document must incorporate metrics, key performance indicators (KPIs), or other significant indicators to authenticate that the inference model maintains behavior consistent with the verified learning model, or at least within an acceptable range. In cases where deviations are anticipated and permissible, acceptance criteria must be clearly defined.

Upon completion of *REF_PhIMD0001_Inference_Requirements_Specifications.docx* and *REF_PhIMD0007_Inference_Requirements_Verification_Tests.docx*, an internal review of both documents should be conducted by completing *REF_PhIMD0002_Inference_Requirements_Specifications_IR.xlsx* and *REF_PhIMD0008_Inference_Requirements_Verification_Tests_IR.xlsx*, respectively).

6 Model Conversion

The model conversion is the process of transforming a trained model into a frozen model, which generally includes preserving the model's architecture and weights in a fixed format optimized for inference. In DL, a frozen model refers to a trained model where the weights and parameters have been fixed or "frozen" after the training process. This means that the model is no longer undergoing updates or learning from new data. At this step, the *REF_PhIMD0003_Model_Conversion_Log. docx* document must be fulfilled, which includes recommendations to facilitate its completion. As outlined in the document, the provided information should encompass, but is not limited to, the following aspects:

- Elimination of training-specific operations. Clarifying the modifications, removals, or replacements of operations with respect to the trained model, ensuring the model is tailored for inference.
- Loading and converting the trained model: Articulating the procedure for loading the saved trained model using relevant tools provided by the DL framework. Additionally, specifying the conversion process to transition to the frozen model format.

After the completion of this document, an internal review shall be conducted by filling out the *REF_PhIMD0004_Model_Conversion_Log_IR.xlsx* document.

7 Model Optimization

At this step, a process is undertaken to enhance performance and/or resource utilization while maintaining, enhancing, or even refining the behavior of the model within a defined range of values specified in the requirements specifications. This guideline proposes completing the *REF_PhIMD0005_Model_Optimization_Log.docx* document with the information related to the model optimization process. It includes recommendations to facilitate its completion. As outlined in the document, the provided information should encompass, but is not limited to, the following aspects:

- *Calibration fundamental operations*: Calibration is the process of choosing α and β for model weights and activations, defining the transformation function or the scale factor.
- *Post-training quantization specifications:* Post-training quantization reduces model memory and computational requirements by converting high-precision parameters to lower precision. Specifications include quantization precision (choosing bit-width) or quantization techniques (i.e., weight or integer quantization), among others.

- *Pruning specifications:* Pruning reduces the computational complexity of a trained model by removing less important connections or parameters. Specifications include criteria for pruning and pruning schedules or patterns.
- *Techniques to recover accuracy*: After applying compression techniques like quantization or pruning, accuracy may be compromised. Recovery techniques involve configurations such as partial quantization or quantization-aware training to find a balance between model efficiency and accuracy.

Upon completion of *REF_PhIMD0005_Model_Optimization_Log.docx*, an internal review shall be conducted by completing *REF_PhIMD0006_Model_Optimization_Log_IR.xlsx.*

8 Model Deployment

Once the inference model has been converted and optimized, the next step is its implementation on the target platform. This phase does not require the completion of any specific template or internal review. The focus is on deploying the optimized model to perform the inference verification of next step on the intended platform. Note that integration of the DL component with other system components is done in subsequent phases of the lifecycle, after inference verification is successfully passed.

9 Inference Verification

The objective of this step is to assess the inference model, compare the results of the inference model verification with the training model verification, and verify whether the deviation in the model's behavior meets the specifications outlined in the inference requirements. For this purpose, the tests previously defined in *REF_PhIMD0007_Inference_Requirements_Verification_Tests.docx* shall be executed and the results collected in the same document. In the event that the requirements collected in *REF_PhIMD0001_Inference_Requirements_Specifications.docx* are not meet, the inference model process shall be reiterated. If the inference model still does not meet the inference requirements specifications, further corrective actions or adjustments in the Data Management and the Learning Management may be required before proceeding with subsequent steps in the process.

Reminder:

- Update the state of REF_Ph0D0003_AI_Document_List.docx when a document is generated or modified, including the last version generated.

- The status of the tests (Not done/Pass/Fail) must be updated in the REF_Ph0D0009_AI_Log_of_Tests.docx.

- The tools and frameworks employed must be listed in REF_Ph0D0011_AI_Tools_Selection.docx.

- The traceability between DL and inference requirements must be updated in REF_Ph0D0013_AI_Traceability_Matrix.docx

- The tests results must be documented in REF_PhIMD0007_Inference_Requirements_Verification_Tests.docx

10 Acronyms and Abbreviations

Below is a list of acronyms and abbreviations employed in this document:

- CNN – Convolutional Neural Networks
- DL – Deep Learning
- HPC – High Performance Computing
- KPI – Key Performance Indicator
- ODD – Operational Design Domain
- RNN – Recurrent Neural Networks
- SGD – Stochastic Gradient Descent

11 Bibliography

[1] D. P. K. a. J. Ba, «ADAM: A Method for Stochastic Optimization,» *arXiv,* 2017.

[2] V. Bushave, «Understanding RMSprop — faster neural network learning,» 2018.

[3] P. I. Frazier, «A Tutorial on Bayesian Optimization,» *arXiv,* 2018.

[4] S. Ruder, «An overview of gradient descent optimization algorithms,» *arXiv,* 2016.