SAFEXPLAIN

Safe and Explainable
Critical Embedded Systems based on AI

PhDMT0003 Data Preparation Log

Version 2.0

Documentation Information

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

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 2
2 Objective 3
3 Scope 3
4 Data Preparation 3
5 Acronyms and Abbreviations 5
6 Bibliography 6

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
V0.3	04/12/2023	Modifications and improvements based on internal review
V0.2	30/08/2023	Modifications and improvements
V0.1	23/05/2023	First draft

*Note. The paragraphs/name of the project/Rev./Ref./history table in **blue** must be replaced with the information for the specific project. The paragraphs written in **red** are instructions that can be used as a guide, so they must be deleted.*

2 Objective

The aim of this document is to collect the steps performed in the data preparation process of the data management phase described in the Data Management guideline.

3 Scope

This template applies to the data preparation process of the Data Management phase performed through the Artificial Intelligence - Functional Safety Management (AI-FSM).

4 Data Preparation

The deliverable generated from this template must include all the information related to the Data Preparation step. This template provides the minimum information that should be collected in this step.

Table 1 collects all the information related to the description of the data preparation step.

Table 1: Information related to the Data Preparation step

Data Preparation			
Date	Date of the preparation: Format YYYY/MM/DD (Year/month/day)		
Responsible	The person or team who annotates, cleans, preprocess, or structures the data.		
Lifecycle Phase	Data Management		
Description (technique used)	• *Data cleaning:* Removing anomalies using an anomaly detector, or correcting erroneous values or standardizing values (e.g., cropping to remove irrelevant information from an image). • *Data processing:* Normalization (e.g., mi-max scaling, z-score normalization, robust scaling to reduce the sensibility to outliers…), scaling, feature selection, dimensionality reduction, data balance, fixing up formats through harmonizing units (e.g., using consistent units), filling in missing values (different strategies can apply in this case, either removing the corresponding row in the dataset or filling missing data) … • *Data annotation:* Manual annotation, Program-based annotation, etc.		
Reason for the Modification	Need to correct errors, improve data quality, adjust to new requirements, etc.		
Data ID of prepared data			
Previous IDs	Previous IDs:	**News IDs**	Proposal. Rename the previous identifier by adding the subindex 'PREP_' at the beginning of the name
Tools/Programs (optional)	Description of the tools and programs employed. Include the required information to replicate the preparation process from scratch. (I.e., Amazon Sage Maker Ground Truth)		
Details of the implementation (optional)	Details of the implementation (libraries, packages): • *Data annotation:* Annotate data using OpenCV. • *Data cleaning:* Removing anomalies using sklearn.svm.OneClassSVM. • *Data pre-processing:* Normalization of the data using sklearn.preprocessing.StandardScaler().		
Configuration of the environment	Package version, input parameters of the function used, etc. For example: train_test_split with parameters test_size=0.2 and random_state=0.		
Expected results	The set of expected results for the modification of the data applied.		
Observations	Additional information. I.e., specify that it has not been possible to collect the required amount of data to meet the data requirements and that for that reason it is necessary to generate new data.		

5 Acronyms and Abbreviations

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

- AI-FSM – Artificial Intelligence - Functional Safety Management

6 Bibliography

Add here the reference to used bibliography / references (if any).