International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

16 March 2023

PRE-PUBLISHED VERSION

DEL5.4

Training and test data specification

ITU-T

International Telecommunication Union

Summary

Deliverable 5.4 provides guidelines on the systematic way of preparing technical requirements specification for datasets used in training and testing of machine learning models and discusses the best practices of data quality assurance aimed at minimizing the data error risks during the training and test data preparation phase of machine learning process lifecycle.

Keywords

Artificial intelligence; health; data requirements; benchmarking platform; test data

Change Log

This document contains Version 1 of the Deliverable DEL5.4 on "*Training and test data specification*" approved on 16 March 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor: Pradeep Balachandran, Technical Consultant, India, E-mail: pbn.tvm@gmail.com

Luis Oala, WG-DAISAM & DotPhoton, Switzerland, E-mail: luis.oala@dotphoton.com

CONTENTS

Page

List of Tables

Page

ITU-T FG-AI4H Deliverable DEL5.4

Training and test data specification

Summary

Deliverable 5.4 provides guidelines on the systematic way of preparing technical requirements specification for datasets used in training and testing of machine learning models and discusses the best practices of data quality assurance aimed at minimizing the data error risks during the training and test data preparation phase of machine learning process lifecycle.

1 Scope

This document is intended to guide the target audience with a systematic way of preparing technical requirements specification for datasets used in training and testing of machine ML models.

This document explains the best practices of data quality assurance aimed at minimizing the data error risks during the training and test data preparation phase of machine learning process lifecycle.

The training and test data requirement specifications follow the data integrity, data security and data safety norms of the AI data governance lifecycle process.

2 References

[1] Timnit Gebru,Google Jamie Morgenstern, et.al. "DatasheetsforDatasets", 19 Mar 2020, arXiv:1803.09010v7

[2] Yun Xu, Royston Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning", *Journal of Analysis and Testing*, 29 October 2018,(https ://doi.org/10.1007/s4166 4-018-0068-2)

[3] Yuji Roh, Geon Heo, Steven Euijong Whang, "A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective",12 Aug 2019, arXiv:1811.03402v2

[4] Zahraa S. Abdallah,Lan Du,Geoffrey I. Webb, "Data Preparation", *C Sammut and G I Webb (Eds) Encyclopedia of Machine Learning and Data Mining*, Springer, 2017

[5] Z. Reitermanov´a, "Data Splitting", *WDS'10 Proceedings of Contributed Papers*, Part I, 31–36, 2010

[6] ISO 7498-2:1989, Information processing systems – Open Systems Interconnection – Basic Reference Model – Part 2: Security Architecture

[7] Oala, Luis, Jana Fehr, Luca Gilli, Pradeep Balachandran, Alixandro Werneck Leite, Saul Calderon-Ramirez, Danny Xie Li et al. "Ml4h auditing: From paper to practice." In Machine learning for health, pp. 280-317. PMLR, 2020.

[8] Parziale, Antonio, Monica Agrawal, Shengpu Tang, Kristen Severson, Luis Oala, Adarsh Subbaswamy, Sayantan Kumar et al. "Machine Learning for Health (ML4H) 2022." In Machine Learning for Health, pp. 1-11. PMLR, 2022.

[9] Roy, Subhrajit, Stephen Pfohl, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen et al. "Machine learning for health (ml4h) 2021." In Machine Learning for Health, pp. 1-12. PMLR, 2021.

[10] Oala, Luis, Andrew G. Murchison, Pradeep Balachandran, Shruti Choudhary, Jana Fehr, Alixandro Werneck Leite, Peter G. Goldschmidt et al. "Machine learning for health: algorithm auditing & quality control." Journal of medical systems 45 (2021): 1-8.

[11] Parziale, Antonio, Monica Agrawal, Shalmali Joshi, Irene Y. Chen, Shengpu Tang, Luis Oala, and Adarsh Subbaswamy. "Machine Learning for Health symposium 2022--Extended Abstract track." arXiv preprint arXiv:2211.15564 (2022).

[12] Falck, Fabian, Yuyin Zhou, Emma Rocheteau, Liyue Shen, Luis Oala, Girmaw Abebe, Subhrajit Roy, Stephen Pfohl, Emily Alsentzer, and Matthew McDermott. "A collection of the accepted abstracts for the Machine Learning for Health (ML4H) symposium 2021." arXiv e-prints (2021): arXiv-2112.

[13] Oala, Luis, Marco Aversa, Gabriel Nobis, Kurt Willis, Yoan Neuenschwander, Michèle Buck, Christian Matek et al. "Data Models for Dataset Drift Controls in Machine Learning With Images." arXiv preprint arXiv:2211.02578 (2022).

[14] Fehr, Jana, Giovanna Jaramillo-Gutierrez, Luis Oala, Matthias I. Gröschel, Manuel Bierwirth, Pradeep Balachandran, Alixandro Werneck-Leite, and Christoph Lippert. "Piloting a Survey-Based Assessment of Transparency and Trustworthiness with Three Medical AI Tools." In Healthcare, vol. 10, no. 10, p. 1923. MDPI, 2022.

[15] Calderon-Ramirez, Saul, Shengxiang Yang, Armaghan Moemeni, Simon Colreavy-Donnelly, David A. Elizondo, Luis Oala, Jorge Rodríguez-Capitán, Manuel Jiménez-Navarro, Ezequiel López-Rubio, and Miguel A. Molina-Cabello. "Improving uncertainty estimation with semi-supervised deep learning for covid-19 detection using chest x-ray images." Ieee Access 9 (2021): 85442-85454.

[16] Willis, Kurt, and Luis Oala. "Post-hoc domain adaptation via guided data homogenization." arXiv preprint arXiv:2104.03624 (2021).

[17] Ramirez, Saul Calderon, Luis Oala, Jordina Torrentes-Barrena, Shengxiang Yang, David Elizondo, Armaghan Moemeni, Simon Colreavy-Donnelly, Wojciech Samek, Miguel Molina-Cabello, and Ezequiel Lopez-Rubio. "Dataset similarity to assess semi-supervised learning under distribution mismatch between the labelled and unlabelled datasets." IEEE Transactions on Artificial Intelligence (2022).

[18] Oala, Luis, Cosmas Heiß, Jan Macdonald, Maximilian März, Gitta Kutyniok, and Wojciech Samek. "Detecting failure modes in image reconstructions with interval neural network uncertainty." International Journal of Computer Assisted Radiology and Surgery 16 (2021): 2089-2097.

3 Terms and definitions

This document defines the following terms:

3.2.1 Training Dataset: A subset of the input dataset that is used to train the ML model.

3.2.2 Test Dataset: A subset of the input dataset that is different from the training dataset (undisclosed) and is used to evaluate and benchmark the ML model performance.

4 Abbreviations

API	Application Programming Interface
App	Application
ARFF	Attribute-Relation File Format
CSV	Comma-Separated Values
DICOM	Digital Imaging and Communications in Medicine
ETL	Extract, Transform, and Load
ICD-10	International Classification of Diseases, Tenth Revision

JPEG	Joint Photographic Experts Group
LOINC	Logical Observation Identifiers Names and Codes
MOV	QuickTime File Format
MP3	MPEG-1 Audio Layer III
MP4	MPEG-4 Part 14
PACS	Picture Archiving and Communication System
PNG	Portable Network Graphics
RAID	Redundant Array of Independent Disks
RxNORM	Prescription (Rx) Normalized Names
SHA-256	Secure Hash Algorithm 256-bit
SNOMED	Systematized Nomenclature of Medicine
SNR	Signal-to-Noise Ratio
SQL	Structured Query Language
SSL	Secure Sockets Layer
Weka	Waikato Environment for Knowledge Analysis

5 Conventions

The following conventions apply in this document:

– "Shall": states a **mandatory** requirement.

– "Should": states a **recommended** requirement.

– "May": states an **optional** requirement.

6 Structure of this document

This document covers all the important steps involved in the preparation of training and test datasets for machine learning starting from clause 7 (with data acquisition requirements) to clause 15 (with test data quality test requirements). Each clause is provided with a corresponding table for stating the requirements specifications and their descriptions.

7 Data acquisition requirements

Table 1 – Data acquisition requirements

REQ. ID	Requirement specification	Description
1	Data specification SHALL state the data acquisition modality	E.g. sensed, self-reported
2	Data specification SHALL state the data acquisition device /sensor type / hardware	E.g. device name, device UID (if any), device model. device manufacturer, etc.
3	Data specification SHALL state the data acquisition device / app firmware	E.g. firmware name, firmware version, etc.
4	Data specification SHALL state the data acquisition device / app operating system (OS)	E.g. Android, iOS, other embedded OS with their version numbers

REQ. ID	Requirement specification	Description
5	Data specification SHALL have a documented procedure / protocol for data acquisition	E.g. data acquisition protocol should support data reproducibility with information on (who, when, where, how, etc.)

8 Data management requirements

Table 2 – Data management requirements

REQ. ID	Requirement specification	Description
6	Data specification SHALL define the data source types	E.g. real & synthetic data sources which includes: electronic health records (anonymised), medical images, vital signs signals, lab test results, photographs, non-medical data-socioeconomic, Environmental, etc), questionnaire responses, free text (discharge / summary, medical history / notes, etc.), PACS, web portal, mobile health app, medical device, etc.
7	Data specification SHALL define the data directory structure and file naming convention	E.g. – organization of parent directory and child directories – file naming convention based on version control appended with title of the file, date, and author name
8	Data specification SHALL have description of the data directory backup structure	
9	Data specification SHALL define the data variable naming convention	E.g. optimized, short and self-explanatory variable names
10	Data specification SHALL define the metadata	E.g. – data creation place – data creation time – data creation authors – data sampling rate – data time frame length – data point IDs – data update version – data migration protocol – other Data creation authors may include: medical personnel (physician/ clinician / nurse /pharmacist/ etc.), support personnel, patient (or proxy person), machine-generated

9 Data quality requirements

Table 3 – Data quality requirements

REQ. ID	Requirement specification	Description
11	Data specification SHALL define the data size	
12	Data specification SHALL define the input data type	E.g. Real valued, integer-valued, categorical value., ordinal value, strings. dates, times, complex data type, other
13	Data specification SHALL define the input data encoding/decoding format	E.g. – DICOM PS3.0 (latest versions) for diagnostic image (X-Ray, CT,MRI, PET, other pathological slides, etc) – JPEG / PNJ for static image – MP3 / OGG Vorbis for audio: – MP4 / MOV for video – SNOMED for clinical observations/terminology – LOINC for laboratory observations – WHO ICD-10 for disease classifications – RxNORM for medication code – Other
14	Data specification SHALL define the output data type	E.g. Binary/Class output (0 or 1) as in case of classification problems, probability output(0-1) as in case of classification problems, continuous valued output as in case of regression problems
15	Data specification SHALL define the data resolution / precision	E.g. Signal-to-Noise Ratio (SNR)
16	Data specification SHALL define the data value range	E.g. minimum and maxima values
17	Data specification SHALL define the data compression / decompression format, if any	E.g. lossy compression / Non-lossy compression techniques
18	Data specification SHALL define the encryption/decryption format, if any	E.g. homographic encryption
19	Data specification SHALL define the data integrity mechanisms used	E.g. integrity mechanisms- RAID, mirroring, checksum, digital signature, etc.
20	Data specification SHALL define the data bias factors, if any	
21	Data specification SHALL define the data privacy / ethical clearance and confidentiality protocol, if any	E.g. anonymization, pseudonymisation& De-identification methods used
22	Data specification SHALL define the data risk factors, if any	
23	Data specification SHALL define the data annotation & labelling protocol used	E.g. – Standards for health data vocabulary / labelling for training and test data

REQ. ID	Requirement specification	Description
		○ Standards for clinical terminology ○ Laboratory observations ○ Disease mapping ○ Procedure mapping ○ Messaging ○ Clinical data format – Procedure – to establish the reference or ground truth for the training data (whether based on objective measures, expert group consensus, etc) – Labelling accuracy calculation technique – Labelling error estimation technique
24	Data specification SHALL define the data safety & security protocol used	E.g. – Access control functions(authentication, authorization, monitoring, logging and auditing) – Audit logs for viewing, creation, modification, validation, copying, import, export, transmission, reception, etc. based ○ On block chain technology ○ Merkle trees, etc – Data repositories compliance with the ISO 7498-2:1989 security model and other allied standards for best practice recommendations on information security management – Implementing security standards based on digital certificate, SSL, SHA-256, etc
25	Data specification SHALL define the data interface protocol used	E.g. – Messaging coding Standards – APIs/Web services for data exchange, data loading/importing – Protocols and tools to collect and integrate diverse data

10 Data loading & pre-processing requirements

Table 4 – Data loading and pre-processing requirements

REQ. ID	Requirement specification	Description
26	Data specification SHALL define the data loading file conventions	E.g. CSV, ARFF (Weka), etc.
27	Data specification SHALL define the standard Extract, Transform, and Load (ETL) tools/ libraries used for data loading	E.g. – Pandas, NumPy, etc for CSV files – Cloud native tools o Alooma o Fivetran o Matillion o Snaplogic o Stitch Data, etc – Open source tools o Apache Airflow o Apache Kafka o Apache NiFi, etc. – Realtime tools o Alooma o Confluent o StreamSets o Striim, etc.
28	Data specification SHALL define the data export & import mechanisms.	E.g. writing and loading datasets to/from SQL database, SQL data warehouse, Hadoop, blob storage, table storage, web URLs, etc
29	Data specification SHALL define the data filtering technique used.	E.g. digital filters to remove the noise /interferences and improve the SNR, suppress or amplify desired frequency components/bands of interest, etc.
30	Data specification SHALL define the standardized data cleaning protocols for cleaning and correction for ranges, variations, outliers, missing values, etc.	E.g. – Verification for missing values and rectifying corrupt or missing values with statistical methods such as imputation- mean, median, mode, 1st or 3rd quartile values, etc. depending on the shape of the data distribution. – Verification for outliers due to data errors, sampling error, etc. and correcting them with flooring and capping of variable values. – Verification for typographical errors and correcting them with numerical coding of variable values. – Cross-verification of data sanity with standard data references.

11 Data visualization requirements

Table 5 – Data visualization requirements

REQ. ID	Requirement specification	Description
31	Data specification SHALL define the data descriptive statistical techniques used to summarize the distribution and relationships between variables	E.g. minimum value, maximum value, means, standard deviation. Pearson's correlation coefficient. skewness (for normal distributions), etc.
32	Data specification SHALL define for the data distribution plotting/ visualization modes and techniques used	E.g. charts, plots, and graphs including histograms. density plots. box plots, scatter plots, etc.

12 Data transformation requirements

Table 6 – Data transformation requirements

REQ. ID	Requirement specification	Description
33	Data specification SHALL define the data re-scaling technique used to normalize the data attributes with varying scales (e.g. data variability in terms of data variable property, data sensing hardware, data sensing software settings, etc.)	E.g. rescaling an input variable to the range between 0 and 1. This method is independent of any data distribution assumption
34	Data specification SHALL define the data re-scaling technique used to standardize the data attributes with normal distribution (differing means and standard deviations)	E.g. rescaling an input variable by configuring the mean of the distribution to the value '0' and the standard deviation to the value '1', '2', from the mean
35	Data specification SHALL define the data thresholding technique used	E.g. applying a binary threshold to the data, whereby data values above the threshold are marked 1' and data values equal to or below are marked as '0'
36	Data specification SHALL define other data transformation techniques used, if any	E.g. logarithm, square roots, exponents. power transforms, etc.
37	Data specification SHALL define other data manipulation techniques used, if any	E.g. merging multiple datasets using joins, merging columns /rows, modifying column names/headings, modifying column data types, etc.

13 Data feature selection requirements

Table 7 – Data feature selection requirements

REQ. ID	Requirement specification	Description
38	Data specification SHALL define the automatic data feature selection technique used	E.g. – Univariate selection – Feature Importance – Correlation matrix with heatmap – Principal component analysis, – Filter methods (Fisher score, Chi-squared score, Pearson's correlation coefficient, Spearman's correlation coefficient etc.) – Wrapper methods (Forward selection, backward selection, recursive feature elimination, etc) – Embedded methods (sparse multinomial logistic regression, automatic relevance determination regression, etc.)
39	Data specification SHALL define the data input features used	
40	Data specification SHALL define the class labels used(in case of classification Problem)	
41	Data specification SHALL define the data dimensions	
42	Data specification SHALL define the input variable names /labelling convention used	
43	Data specification SHALL define the output variable names /labelling convention used	

14 Train & test data configuration requirements

Table 8 – Train & test data configuration requirements

REQ. ID	Requirement specification	Description
44	Data specification SHALL define the data partitioning method used	E.g. – Sample and split method o Split data into a training and testing data set based on a custom percentage or ratio o filter training data based on a specific attribute in the data. – Cross validation method o K-fold validation – regular expression and relative expressions filtering based splitting

REQ. ID	Requirement specification	Description
45	Data specification SHALL define the 'percentage / ratio of training set' split (in case of sample and split method)	
46	Data specification SHALL define the 'percentage / ratio of test set' split (in case of sample and split method)	
47	Data specification SHALL define the 'split repetition count ' (in case of sample and split method)	
48	Data specification SHALL define the 'fold size' used (in case of K-fold validation)	
49	Data specification SHALL define the 'unit fold size' used (in case of K-fold validation)	

15 Test data quality test requirements

Table 9 – Test data quality test requirements

REQ. ID	Requirement specification	Description
50	Data specification SHALL define the test data quality test performed to minimize the noise and variance of the test data and to maximize the performance accuracy of ML algorithm	E.g. Test plan & procedure for – Training and testing on the same dataset – Split tests – Multiple split tests – Cross validation – Multiple cross validation – Statistical significance