|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-I-034 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | E-meeting, 7-8 May 2020 | |
| **DOCUMENT** | | | | |
| **Source:** | | Editors | | |
| **Title:** | | DEL5.4: Training and test data specification | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Luis Oala WG-DAISAM & Fraunhofer HHI Germany | | Email: [luis.oala@hhi.fraunhofer.de](mailto:luis.oala@hhi.fraunhofer.de) |
| **Contact:** | | Pradeep Balachandran Technical Consultant (eHealth) | | Email: [pbn.tvm@gmail.com](mailto:pbn.tvm@gmail.com) |

|  |  |
| --- | --- |
| **Abstract:** | This document contains the draft version 2.0 of the project deliverable FG-AI4H DEL5.4 on "Training and test data specification". This belongs to a set of four deliverables under the umbrella of the deliverable FG-AI4H-DEL05.1 "Data specification" |

*Call for Participation*

If you are interested in contributing to the *Training and Test Data Specification*, please contact the editors of this document (Pradeep Balachandran, pbn.tvm@gmail.com, or Luis Oala, [luis.oala@hhi.fraunhofer.de](mailto:luis.oala@hhi.fraunhofer.de)) and the Secretariat of the Focus Group ([tsbfgai4h@itu.int](mailto:tsbfgai4h@itu.int)) using a descriptive e-mail-subject, briefly introduce yourself, describe your relevant expertise, and explain your interest.

*Expertise profile of potential contributors*

AI data preparation, AI model training & testing

*Time plan (first draft, release 1)*

AI4H FG Meeting I, May 7-8, 2020

*Target audience*

AI data curators, AI data quality analysts, AI model developers, AI model testers

**CONTENTS**

| Page |
| --- |
| [1 Scope and objectives 3](#_Toc39490063)  [2 References 3](#_Toc39490064)  [3 Definitions 3](#_Toc39490065)  [4 Structure of this document 3](#_Toc39490066)  [5 Document conventions 3](#_Toc39490067)  [6 Data acquisition requirements 4](#_Toc39490068)  [7 Data management requirements 4](#_Toc39490069)  [8 Data quality requirements 5](#_Toc39490070)  [9 Data loading & pre-processing requirements 7](#_Toc39490071)  [10 Data visualization requirements 8](#_Toc39490072)  [11 Data transformation requirements 8](#_Toc39490073)  [12 Data feature selection requirements 9](#_Toc39490074)  [13 Train & test data configuration requirements 10](#_Toc39490075)  [14 Test data quality test requirements 10](#_Toc39490076) |

**List of Tables**

| Page |
| --- |
| [Table 1: Data acquisition requirements 4](#_Toc39490106)  [Table 2: Data management requirements 4](#_Toc39490107)  [Table 3: Data quality requirements 5](#_Toc39490108)  [Table 4: Data loading and pre-processing requirements 7](#_Toc39490109)  [Table 5: Data visualization requirements 8](#_Toc39490110)  [Table 6: Data transformation requirements 8](#_Toc39490111)  [Table 7: Data feature selection requirements 9](#_Toc39490112)  [Table 8: Train & test data configuration requirements 10](#_Toc39490113)  [Table 9: Test data quality test requirements 10](#_Toc39490114) |

FG-AI4H Deliverable 5.4

Training and test data specification

# Scope and objectives

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

# References

1. Timnit Gebru,Google Jamie Morgenstern, et.al. "Datasheets for Datasets", 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

# Definitions

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

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

# 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 section 4 – "data acquisition requirements" to section 12 – "test data quality test requirements". Each section is provided with a corresponding table for stating the requirements specifications and their descriptions

# Document conventions

This document shall conform to the following standard convention of specification language syntax to indicate its particular significance/‌compliance level.

| Term | Meaning |
| --- | --- |
| "SHALL" | states a **mandatory** requirement of this policy |
| "SHOULD" | states a **strongly suggested** requirement of this policy |
| "MAY" | states an **optional** requirement |

# 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 |
| 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.) |

# 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 |

# 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 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 * 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 ISO 7498-2 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 |

# 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 * Alooma * Fivetran * Matillion * Snaplogic * Stitch Data, etc * Open source tools * Apache Airflow * Apache Kafka * Apache NiFi, etc. * Realtime tools * Alooma * Confluent * StreamSets * 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. * 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 sanctity with standard data references |

# 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. |

# 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 root, 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. |

# 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 |  |

# 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 * Split data into a training and testing data set based on a custom percentage or ratio * filter training data based on a specific attribute in the data. * Cross validation method * K-fold validation * regular expression and relativeexpressions filtering based splitting |
| 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) |  |

# 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 |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_