ETSI TR 128 908 V18.0.0 (2024-05)

5G;
Study on Artificial Intelligence/Machine Learning (AI/ ML) management
(3GPP TR 28.908 version 18.0.0 Release 18)

Reference
DTR/TSGS-0528908vi00

Keywords
5G

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Legal Notice

This Technical Report (TR) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities. These shall be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between 3GPP and ETSI identities can be found under https://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

In the present document, modal verbs have the following meanings:

shall indicates a mandatory requirement to do something

shall not indicates an interdiction (prohibition) to do something

The constructions "shall" and "shall not" are confined to the context of normative provisions, and do not appear in Technical Reports.

The constructions "must" and "must not" are not used as substitutes for "shall" and "shall not". Their use is avoided insofar as possible, and they are not used in a normative context except in a direct citation from an external, referenced, non-3GPP document, or so as to maintain continuity of style when extending or modifying the provisions of such a referenced document.

should indicates a recommendation to do something

should not indicates a recommendation not to do something

may indicates permission to do something

need not indicates permission not to do something

The construction "may not" is ambiguous and is not used in normative elements. The unambiguous constructions "might not" or "shall not" are used instead, depending upon the meaning intended.

can indicates that something is possible

cannot indicates that something is impossible

The constructions "can" and "cannot" are not substitutes for "may" and "need not".

will indicates that something is certain or expected to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

will not indicates that something is certain or expected not to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

might indicates a likelihood that something will happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

might not indicates a likelihood that something will not happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

In addition:

is (or any other verb in the indicative mood) indicates a statement of fact

is not (or any other negative verb in the indicative mood) indicates a statement of fact

The constructions "is" and "is not" do not indicate requirements.

1 Scope

The present document studies the Artificial Intelligence / Machine Learning (AI/ML) management capabilities and services for 5GS where AI/ML is used, including management and orchestration (e.g. MDA, see 3GPP TS 28.104 [2]), 5GC (e.g. NWDAF, see 3GPP TS 23.288 [3], and NG-RAN (e.g. RAN intelligence defined in 3GPP TS 38.300 [16] and 3GPP TS 38.401 [19]).

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TS 28.104: "Management and orchestration; Management Data Analytics (MDA)".

[3] 3GPP TS 23.288: "Architecture enhancements for 5G System (5GS) to support network data analytics services".

[4] 3GPP TS 28.105: "Management and orchestration; Artificial Intelligence/Machine Learning (AI/ML) management".

[5] IBM Watson Studio: "Model Drift" [Online].

NOTE: Available at: https://www.ibm.com/cloud/watson-studio/drift.

[6] 3GPP TR 28.864: "Study on Enhancement of the management aspects related to NetWork Data Analytics Functions (NWDAF)".

NOTE: Available at https://www.3gpp.org/dynareport/28864.htm.

[7] 3GPP TS 28.310: "Management and orchestration; Energy efficiency of 5G".

[8] 3GPP TS 28.552: "Management and orchestration; 5G performance measurements".

[9] 3GPP TS 28.313: "Management and orchestration; Self-Organizing Networks (SON) for 5G networks".

[10] European Commission (21.04.2021): "Proposal for a Regulation laying down harmonized rules on artificial intelligence".

[11] High-level Expert Group on Artificial Intelligence setup by the European Commission (08.04.2019): "Ethical Guidelines for Trustworthy AI".

[12] ISO/IEC TR 24028:202: "Information technology -- Artificial intelligence -- Overview of trustworthiness in artificial intelligence".

[13] 3GPP TS 28.622: "Telecommunication management; Generic Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)".

[14] 3GPP TS 28.554: "Management and orchestration;5G end to end Key Performance Indicators (KPI)".

[15] 3GPP TR 37.817: "Study on enhancement for data collection for NR and ENDC".

[16] 3GPP TS 38.300: "NR; NR and NG-RAN Overall description; Stage-2".

[17] 3GPP TS 28.533: "Management and orchestration; Architecture framework".

[18] 3GPP TR 28.813: "Management and orchestration; Study on new aspects of Energy Efficiency (EE) for 5G".

[19] 3GPP TS 38.401: "NG-RAN; Architecture description".

[20] 3GPP TS 28.541: "Management and orchestration of 5G networks; Network Resource Model (NRM); Stage 2 and stage 3".

3 Definitions of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms given in 3GPP TR 21.905 [1], TS 28.105 [4] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

MAE Mean Absolute Error
MDCA Management Data Correlation Analytics
MSE Mean Squared Error

4 Concepts and overview

4.1 Concepts and terminologies

Biased data: Biased data in machine learning occurs when certain data samples of a training dataset are more heavily weighted and/or overrepresented in comparison to others. Biased data may lead to lower quality predictions and/or reduced accuracy of the trained ML model.

F1 score: (also known as F-measure, or balanced F-score) is a metric used to measure the training performance of classification ML models.

4.2 Overview

Artificial Intelligence/Machine Learning (AI/ML) techniques are being embraced by telecommunication service providers around the world to facilitate enabling the existing and the new challenging use cases that 5G offers. AI/ML capabilities are being increasingly adopted in mobile networks as a key enabler for wide range of features and functionalities that maximise efficiency and bring intelligence and automation in various domains of the 5GS. For example, these include the Management Data Analytics (MDA) in the management and orchestration [1], the Network Data Analytics Function (NWDAF) in the 5G core network domain [3], and NG-RAN (e.g. RAN intelligence) defined in 3GPP TS 38.300 [16] and 3GPP TS 38.401 [19].

The AI/ML inference functions in the 5GS use the ML model for inference and in order to enable and facilitate the AI/ML adoption, the ML model needs to be created, trained and then managed during its entire lifecycle.

To enable, facilitate and support AI/ML-capabilities in the 5GS, the following management capabilities are studied in the present document:

- Validation of ML model or entity.
- Testing of ML model or entity (before deployment).
- Deployment of ML model or entity (new or updated model/entity).
- Configuration of ML training and AI/ML inference.
- Performance evaluation of ML training and AI/ML inference.

NOTE: The ML model training capability is specified in 3GPP TS 28.105 [4].

4.3 AI/ML workflow for 5GS

4.3.1 AI/ML operational workflow

AI/ML techniques are widely used in 5GS (including 5GC, NG-RAN and management system), and the generic workflow of the operational steps in the lifecycle of an ML model or entity, is depicted in the figure 4.3.1-1.

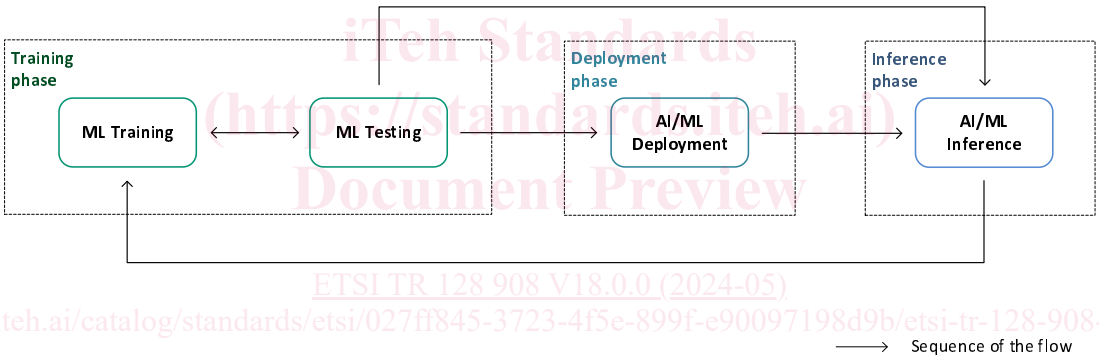

Figure 4.3.1-1: AI/ML operational workflow

The workflow involves 3 main phases; the training, deployment and inference phase, including the main operational tasks for each phase. These are briefly described below:

Training phase:

- **ML Training:** Learning by the Machine from the training data to generate the (new or updated) ML entity (see 3GPP TS 28.105 [4]) that could be used for inference. The ML Training may also include the validation of the generated ML entity to evaluate the performance variance of the ML entity when performing on the training data and validation data. If the validation result does not meet the expectation (e.g. the variance is not acceptable), the ML entity needs to be re-trained. This is the initial step of the workflow. The ML Training MnS is specified in 3GPP TS 28.105 [4].
- **ML Testing:** Testing of the validated ML entity with testing data to evaluate the performance of the trained ML entity for selection for inference. When the performance of the trained ML entity meets the expectations on both training data and validation data, the ML entity is finally tested to evaluate the performance on testing data. If the testing result meets the expectation, the ML entity may be counted as a candidate for use towards the intended use case or task, otherwise the ML entity may need to be further (re)trained. In some cases, the ML entity may need to be verified which is the special case of testing to check whether it works in the AI/ML inference function or the target node. In other cases, the verification step may be skipped, for instance in case the input and output data, data types and formats, have been unchanged from the last ML entity.

Deployment phase:

- **ML Deployment:** Deployment of the trained and tested ML entity to the target inference function which will use the subject ML entity for inference.

NOTE: The deployment phase may not be needed in some cases, for example when the training function and inference function are in the same entity.

Inference phase:

- **AI/ML Inference:** Performing inference using the ML entity by the inference function.

In telco-grade environments, it is worth noting that the selected learning method (see examples of learning methods captured in table 4.1-1 in 3GPP TS 28.105 [4]) can influence on how AI/ML operational workflow executes. In some cases (e.g. when using supervised learning methods), the inference phase cannot start until training phase gets ended. In other cases (e.g. when using reinforcement learning methods), the inference phase can start while training phase is still in progress.

4.3.2 AI/ML management capabilities

Each operational step in the workflow (as depicted in clause 4.3.1) may be related to one or more AI/ML management capabilities, including:

Management capabilities for training and testing phase

- **ML training data management:** This involves management capabilities for managing the data needed for training the ML entities. It may also include capabilities for processing of data as requested by a training function, by another management function or by the MLT MnS consumer.

- **ML training management:** allowing the MnS consumer to request and/or manage the model training/retraining. For example, activating/deactivating, training performance management and setting policy for the producer-initiated ML training (e.g. the conditions to trigger the ML (re-)training based on the AI/ML inference performance or AI/ML inference trustworthiness).

- **ML testing management:** allowing the MnS consumer to request the ML entity testing, and to receive the testing results for a trained ML model. It may also include capabilities for selecting the specific performance and trustworthiness metrics to be used or reported by the ML testing function.

- **ML validation:** ML training capability may also include validation to evaluate the performance and trustworthiness of the ML entity when performing on the validation data, and to identify the variance of the performance and trustworthiness on the training data and the validation data. If the variance is not acceptable, the entity would need to be tuned (re-trained) before being made available to the consumer and used for inference.

Management capabilities for deployment phase

- **AI/ML deployment control and monitoring:** This involves capabilities for loading the ML entity to the target inference function. It includes providing information to the consumer when new entities are available, enabling the consumer to request the loading of the ML entity or to set the policy for such deployment and to monitor the deployment process.

Management capabilities for inference phase

- **ML entity activation/deactivation:** allowing the MnS consumer to activate/deactivate the inference function and/or ML entity/entities, including instant activation, partial activation, schedule-based or policy-based activations.

- **AI/ML inference function control:** allowing the MnS consumer to control the inference function including the activation and deactivation of the function.

- **AI/ML inference performance management:** allowing the MnS consumer to monitor and evaluate the inference performance of an ML entity when used by an AI/ML inference function.

- **AI/ML trustworthiness management:** allowing the MnS consumer to monitor and evaluate the inference trustworthiness of an ML entity when used by an AI/ML inference function.