|  |  |
| --- | --- |
| 3GPP TR 22.850 V0.5.0 (2025-09) | |
| Technical Report | |
| 3rd Generation Partnership Project;  Technical Specification Group Services and System Aspects;  Study on 3GPP AI/ML Consistency Alignment  (Release 19) | |
|  | |
|  |  |
|  | |
| The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP. The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented. This Specification is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification. Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices. | |

|  |
| --- |
|  |
| ***3GPP***  Postal address  3GPP support office address  650 Route des Lucioles - Sophia Antipolis  Valbonne - FRANCE  Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16  Internet  https://www.3gpp.org |
| ***Copyright Notification***  No part may be reproduced except as authorized by written permission. The copyright and the foregoing restriction extend to reproduction in all media.  © 2025, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).  All rights reserved.  UMTS™ is a Trade Mark of ETSI registered for the benefit of its members  3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners  GSM® and the GSM logo are registered and owned by the GSM Association |

Contents

Foreword 7

1 Scope 9

2 References 9

3 Definitions of terms and abbreviations 12

3.1 Terms 12

3.2 Abbreviations 12

4 Background 12

5 AI/ML related activities in all Working Groups 13

5.1 General 13

5.2 AI/ML related activities in TSG SA & CT Working Groups 16

5.2.1 AI/ML related terminology 16

5.2.1.1 TSG SA WG2 16

5.2.1.2 TSG SA WG5 17

5.2.1.3 TSG SA WG6 18

5.2.2 AI/ML related activities 19

5.2.2.1 Rel-18 SA WG1 WID - AI/ML model transfer in 5GS (AIML\_MT) 19

5.2.2.1.1 Description 19

5.2.2.1.2 Activities summary 19

5.2.2.2 Rel-19 SA WG1 SID - AI/ML Model Transfer Phase 2 (FS\_AIML\_MT\_Ph2) 20

5.2.2.2.1 Description 20

5.2.2.2.2 Activities summary 20

5.2.2.3 Rel-19 SA WG1 WID - AI/ML Model Transfer Phase 2 (AIML\_MT\_Ph2) 20

5.2.2.3.1 Description 20

5.2.2.3.2 Activities summary 20

5.2.2.4 Rel-18 SA WG2 WID - Enablers for Network Automation for 5G - phase 3 (eNA\_Ph3) 21

5.2.2.4.1 Description 21

5.2.2.4.2 Activities summary 21

5.2.2.5 Rel-18 SA WG2 WID - System Support for AI/ML-based Services (AIMLsys) 22

5.2.2.5.1 Description 22

5.2.2.5.2 Activities summary 23

5.2.2.6 Rel-19 SA WG2 SID - Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML) (FS\_AIML\_CN) 24

5.2.2.6.1 Description 24

5.2.2.7 Rel-19 SA WG2 WID - Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML) (AIML\_CN) 24

5.2.2.7.1 Description 24

5.2.2.7.2 Activities summary 25

5.2.2.8 Rel-18 SA WG3 WID - Security aspects of enablers for Network Automation for 5G - phase 3 (eNA\_SEC\_PH3) 25

5.2.2.8.1 Description 25

5.2.2.8.2 Activities summary 25

5.2.2.9 Rel-19 SA WG3 SID - Security aspects of Core Network Enhanced Support for AIML (FS\_AIML\_CN\_SEC) 26

5.2.2.9.1 Description 26

5.2.2.10 Rel-19 SA WG4 SID - Artificial Intelligence (AI) and Machine Learning (ML) for Media (FS\_AI4Media) 26

5.2.2.10.1 Description 26

5.2.2.11 Rel-18 SA WG5 WID - AI/ML management (AIML\_MGT) 26

5.2.2.11.1 Description 26

5.2.2.11.2 Activities summary 27

5.2.2.12 Rel-19 SA WG5 SID - AI/ML management - phase 2 (FS\_AIML\_MGT\_Ph2) 27

5.2.2.12.1 Description 27

5.2.2.13 Rel-19 SA WG6 SID - Application layer support for AI/ML services (FS\_AIMLAPP) 28

5.2.2.13.1 Description 28

5.2.2.13.2 Activities summary 28

5.2.2.14 Rel-19 SA WG6 WID - Application enablement for AI/ML services (AIML\_App) 29

5.2.2.14.1 Description 29

5.2.2.14.2 Activities summary 29

5.2.2.15 Rel-19 CT WG4 WID - Protocol for AI Data Collection from UPF (FS\_PAIDC-UPF) 31

5.2.2.15.1 Description 31

5.2.2.15.2 Activities summary 31

5.2.2.16 Rel-19 SA WG3 WID - Security aspects of Core Network Enhanced Support for AIML (AIML\_CN\_SEC) 31

5.2.2.16.1 Description 31

5.2.2.16.2 Activities summary 31

5.2.2.17 Rel-19 SA WG5 WID - AI/ML management phase 2 (AIML\_MGT\_Ph2) 32

5.2.2.17.1 Description 32

5.2.2.17.2 Activities summary 33

5.2.2.18 Rel-19 SA WG6 WID - Application Data Analytics Enablement Service (TEI19\_ADAES) 33

5.2.2.18.1 Description 33

5.2.2.18.2 Activities summary 33

5.2.2.19 Rel-19 CT WG1/WG3 WID – CT aspects of application enablement for AI/ML services (AIML\_App) 33

5.2.2.19.1 Description 33

5.2.2.19.2 Activities summary 34

5.2.2.20 Rel-19 CT WG3 WID – Rel-19 Enhancements of Network Automation Enablers (eNetAE19) 34

5.2.2.20.1 Description 34

5.2.2.20.2 Activities summary 35

5.2.2.21 Rel-19 CT WG3/WG4 WID – CT aspects of Core Network Enhanced Support for Artificial Intelligence (AI) and Machine Learning (ML) (AIML\_CN) 35

5.2.2.21.1 Description 35

5.2.2.21.2 Activities summary 36

5.3 AI/ML related activities in TSG RAN Working Groups 36

5.3.1 AI/ML related terminology 36

5.3.1.1 TSG RAN WG1 36

5.3.1.2 TSG RAN WG3 38

5.3.2 AI/ML related activities 38

5.3.2.1 Rel-19 RAN WG1/RAN WG4 WID - Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface (NR\_AIML\_air) 38

5.3.2.1.1 Description 38

5.3.2.1.2 Activities summary 40

5.3.2.2 Rel-19 RAN WG2 SID - AIML for mobility in NR (FS\_NR\_AIML\_Mob) 40

5.3.2.2.1 Description 40

5.3.2.3 Rel-18 RAN WG3 WID - Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN (NR\_AIML\_NGRAN-Core) 40

5.3.2.3.1 Description 40

5.3.2.3.2 Activities summary 41

5.3.2.4 Rel-19 RAN WG3 SID - Enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN (FS\_NR\_AIML\_NGRAN\_enh) 41

5.3.2.4.1 Description 41

5.3.2.5 Rel-19 RAN WG1/RAN WG4 SID - Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface (FS\_NR\_AIML\_air\_Ph2) 41

5.3.2.5.1 Description 41

5.3.2.6 Rel-19 RAN WG3 WID - Enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN (FS\_NR\_AIML\_NGRAN\_enh-Core) 42

5.3.2.6.1 Description 42

5.3.2.6.2 Activities summary 43

6 Analysis on AI/ML across 3GPP 43

6.1 General 43

6.2 AI/ML related terminology 43

6.2.1 Analysis on AI/ML model related terminology consistency 43

6.2.1.1 Analysis on ML model 43

6.2.1.2 Analysis on ML model training 44

6.2.1.3 Analysis on ML model re-training 44

6.2.1.4 Analysis on ML model testing 45

6.2.1.5 Analysis on ML model inference 45

6.2.1.6 Analysis on ML model activation & ML model de-activation 46

6.2.1.7 Analysis on ML model lifecycle 46

6.2.1.8 Analysis on ML model lifecycle management 47

6.2.1.9 Analysis on usage of ML Model identifier in each Working Group 48

6.2.1.9.1 RAN WG 1 48

6.2.1.9.2 RAN WG 3 49

6.2.1.9.3 SA WG 2 49

6.2.1.9.4 SA WG5 50

6.2.1.9.5 SA WG6 51

6.2.1.10 Analysis on ML model pre-specialized training and ML model fine-tuning 52

6.2.1.11 Analysis on ML model distributed training 53

6.2.2 Analysis on Federated Learning 53

6.2.4 Analysis on Decision vs Prediction vs Output 54

6.2.5 Analysis on ML vs AI vs AI/ML 55

6.2.6 Analysis on Transfer Learning 55

6.3 AI/ML related features 55

6.3.1 Analysis on ML model training services 55

6.3.2 Analysis on analytics related services 57

6.3.3 Analysis on ML performance evaluation and monitoring 62

6.3.4 Analysis on data collection and management for AI/ML 63

7 Overall Evaluation 71

7.1 General evaluation on AI/ML related terminology 71

7.2 Detailed evaluation on AI/ML-related terminology 72

7.3 Evaluation summary on AI/ML-related terminology 72

7.4 Evaluation summary on AI/ML-related features 73

8 Conclusions 73

Annex A: ML Model 74

A.1 ML model life cycle management (LCM) 74

A.1.1 Observations and analyses: AI/ML LCM 74

A.2 ML model lifecycle management capabilities 74

A.2.1 Observations and analyses: ML model lifecycle management capabilities 75

A.3 AI/ML functionalities management scenarios 75

A.3.1 Observations and analyses: AI/ML functionalities management scenarios 75

Annex B: Change history 76

# 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

This study will investigate ongoing AI/ML work in TSG CT, TSG RAN and TSG SA Working Groups and identify instances of any potential misalignment and/or inconsistencies and provide information on any potential outcome to the respective WGs to resolve any issues with appropriate SA-level co-ordination as necessary.

The study is led by TSG SA in close collaboration and inputs from TSG CT and TSG RAN.

NOTE**:** The study item does not impact ongoing studies and normative work for AI/ML across all SA/RAN/CT WGs for Rel-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 TR 21.918: "Summary of Rel-18 Work Items".

[3] 3GPP TR 38.843: "Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface (Release 18)".

[4] 3GPP TR 23.700‑84: "Study on Core Network Enhanced Support for Artificial Intelligence (AI) / Machine Learning (ML)".

[5] 3GPP TR 22.874: "5G System (5GS); Study on traffic characteristics and performance requirements for AI/ML model transfer".

[6] 3GPP TS 22.261: "Service requirements for the 5G system".

[7] 3GPP TR 23.700‑82: "Study on application layer support for AI/ML services".

[8] 3GPP TS 23.288: "Architecture enhancements for 5G System (5GS) to support network data analytics services".

[9] 3GPP TS 28.105: "Management and orchestration; Artificial Intelligence/ Machine Learning (AI/ML) management".

[10] Void.

[11] 3GPP TS 38.300: "NR; NR and NG-RAN Overall description; Stage-2".

[12] 3GPP TR 26.927: "Study on Artificial Intelligence and Machine learning in 5G media services".

[13] 3GPP TS 38.401: "NG-RAN; Architecture description".

[14] 3GPP TS 38.420: "NG-RAN; Xn general aspects and principles".

[15] 3GPP TS 38.423: "NG-RAN; Xn Application Protocol (XnAP)".

[16] 3GPP TS 23.273: "5G System (5GS) Location Services (LCS); Stage 2".

[17] 3GPP TR 37.817: "Study on enhancement for data collection for NR and ENDC".

[18] 3GPP TR 28.908: "Study on Artificial Intelligence/Machine Learning (AI/ ML) management".

[19] 3GPP TR 28.858: "Study on Artificial Intelligence / Machine Learning (AI/ML) management; Phase 2".

[20] 3GPP TR 29.889: "Study on Protocol for AI Data Collection from UPF".

[21] 3GPP TR 22.876: "Study on AI/ML Model Transfer; Phase 2".

[22] 3GPP TS 23.501: "System architecture for the 5G System (5GS)".

[23] 3GPP TS 23.502: "Procedures for the 5G System (5GS)".

[24] 3GPP TS 23.503: "Policy and charging control framework for the 5G System (5GS); Stage 2".

[25] 3GPP TR 33.784: "Study on security aspects of core network enhanced support for Artificial Intelligence Machine Learning (AIML)".

[26] 3GPP TR 26.847: "Evaluation of Artificial Intelligence and Machine learning in 5G media services".

[27] 3GPP TS 28.552: "Management and orchestration; 5G performance measurements".

[28] 3GPP TS 32.425: "Telecommunication management; Performance Management (PM); Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN)".

[29] 3GPP TS 28.554: "Management and orchestration; 5G end to end Key Performance Indicators (KPI)".

[30] 3GPP TS 28.541: "Management and orchestration; 5G Network Resource Model (NRM); Stage 2 and stage 3".

[31] 3GPP TS 32.254: "Telecommunication management; Charging management; Exposure function Northbound Application Program Interfaces (APIs) charging".

[32] 3GPP TS 32.291: "Telecommunication management; Charging management; 5G system, charging service; Stage 3".

[33] 3GPP TS 23.436: "Functional architecture and information flows for Application Data Analytics Enablement Service".

[34] 3GPP TS 23.482: "Functional architecture and information flows for AIML Enablement Service".

[35] 3GPP TS 23.434: "Service Enabler Architecture Layer for Verticals (SEAL); Functional architecture and information flows".

[36] 3GPP TS 23.558: "Architecture for enabling Edge Applications".

[37] 3GPP TS 29.122: "T8 reference point for Northbound APIs".

[38] 3GPP TS 29.513: "5G System; Policy and Charging Control signalling flows and QoS parameter mapping; Stage 3".

[39] 3GPP TS 29.514: "5G System; Policy Authorization Service; Stage 3".

[40] 3GPP TS 29.517: "5G System; Application Function Event Exposure Service; Stage 3".

[41] 3GPP TS 29.591: "5G System; Network Exposure Function Southbound Services; Stage 3".

[42] 3GPP TS 29.519: "5G System; Usage of the Unified Data Repository Service for Policy Data, Application Data and Structured Data for Exposure; Stage 3".

[43] 3GPP TS 29.520: "5G System; Network Data Analytics Services; Stage 3".

[44] 3GPP TS 29.521: "5G System; Binding Support Management Service; Stage 3".

[45] 3GPP TS 29.522: "5G System; Network Exposure Function Northbound APIs; Stage 3".

[46] 3GPP TS 29.523: "5G System; Policy Control Event Exposure Service; Stage 3".

[47] 3GPP TS 29.525: "5G System; UE Policy Control Service; Stage 3".

[48] 3GPP TS 29.508: "5G System; Session Management Event Exposure Service; Stage 3".

[49] 3GPP TS 29.551: "5G System; Packet Flow Description Management Service; Stage 3".

[50] 3GPP TS 29.552: "5G System; Network Data Analytics signalling flows; Stage 3".

[51] 3GPP TS 29.574: "5G System; Data Collection Coordination Services; Stage 3".

[52] 3GPP TS 29.575: "5G System; Analytics Data Repository Services; Stage 3".

[53] 3GPP TS 29.576: "5G System; Messaging Framework Adaptor Services; Stage 3".

[54] 3GPP TS 29.503: "5G System; Unified Data Management Services; Stage 3".

[55] 3GPP TS 29.504: "5G System; Unified Data Repository Services; Stage 3".

[56] 3GPP TS 29.505: "5G System; Usage of the Unified Data Repository services for Subscription Data; Stage 3".

[57] 3GPP TS 29.510: "5G System; Network function repository services; Stage 3".

[58] 3GPP TS 29.564: "5G System; User Plane Function Services; Stage 3".

[59] 3GPP TS 29.518: "5G System; Access and Mobility Management Services; Stage 3".

[60] 3GPP TS 29.536: "5G System; Network Slice Admission Control Services; Stage 3".

[61] 3GPP TS 29.571: "5G System; Common Data Types for Service Based Interfaces; Stage 3".

[62] 3GPP TS 38.214: "NR; Physical layer procedures for data".

[63] 3GPP TS 38.215: "NR; Physical layer measurements".

[64] 3GPP TS 38.331: "NR; Radio Resource Control (RRC); Protocol specification".

[65] 3GPP TS 38.305: "NG Radio Access Network (NG-RAN); Stage 2 functional specification of User Equipment (UE) positioning in NG-RAN".

[66] 3GPP TS 37.355: "LTE Positioning Protocol (LPP)".

[67] 3GPP TS 38.455: "NG-RAN; NR Positioning Protocol A (NRPPa)".

[68] 3GPP TS 38.133: "NR; Requirements for support of radio resource management".

[69] 3GPP TR 38.743: "Study on enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN".

[70] 3GPP TR 38.744: "Study on Artificial Intelligence (AI)/Machine Learning (ML) for mobility in NR".

[71] 3GPP TS 28.104: "Management Data Analytics (MDA)".

[72] 3GPP TS 28.622: "Generic Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS)".

[73] 3GPP TR 28.871: "Study on Service Based Management Architecture enhancement phase 3".

[74] 3GPP TR 26.531: "Data Collection and Reporting; General Description and Architecture".

[75] 3GPP TS 24.559: "Application Data Analytics Enablement Services (ADAES); Stage 3".

[76] 3GPP TS 24.560: "Artificial Intelligence Machine Learning (AIML) Services - Service Enabler Architecture Layer for Verticals (SEAL); Protocol Specification; Stage 3".

[77] 3GPP TS 29.482: “Service Enabler Architecture Layer for Verticals (SEAL); Artificial Intelligence Machine Learning Enablement (AIMLE) Services; Stage 3”.

[78] 3GPP TS 29.549: “Service Enabler Architecture Layer for Verticals (SEAL); Application Programming Interface (API) specification; Stage 3”.

[79] 3GPP TS 29.558: “Enabling Edge Applications; Application Programming Interface (API) specification; Stage 3”.

[80] 3GPP TS 24.080: “Mobile radio interface layer 3 Supplementary services specification; Formats and coding”.

[81] 3GPP TS 29.570: “5G System; Service Communication Proxy Services Stage 3”.

[82] 3GPP TS 29.572: “5G System; Location Management Services; Stage 3”.

[83] 3GPP TS 37.340: “Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Overall Description; Stage 2”.

[84] 3GPP TS 37.480: “E1 general aspects and principles”.

[85] 3GPP TS 37.483: “E1 Application Protocol (E1AP)”.

[86] 3GPP TS 38.470: “NG-RAN; F1 general aspects and principles”.

[87] 3GPP TS 38.473: “NG-RAN; F1 Application Protocol (F1AP)”.

# 3 Definitions of terms and abbreviations

## 3.1 Terms

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

**example:** text used to clarify abstract rules by applying them literally.

## 3.2 Abbreviations

For the purposes of the present document, the abbreviations given in 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 TR 21.905 [1].

<ABBREVIATION> <Expansion>

# 4 Background

In Rel-18 and Rel-19, most working groups in TSG SA, TSG CT and TSG RAN have already performed SIs and/or have WIs relating to the AI/ML topic. These activities address different usage scenarios and associated specific use cases exploiting AI/ML for the operation of the 3GPP System ranging from radio interface operations (e.g. beam management, positioning), NG-RAN operations (e.g. energy saving, load balancing), 5GC operations (e.g. assistance to QoS and policy control, prevention and mitigation of signalling storms, etc.), to network management & orchestration, media services and application enablement aspects. In addition, some activities target the enhanced support of the 3GPP System for AI/ML-based applications and services which are themselves out of the scope of 3GPP.

With the complexity of the 3GPP systems and its operations and that of AI/ML, it is vital that the use of AI/ML in the operation of the 3GPP system (incl. related AI/ML model LCM) for any given use case be bound to specific principles, guidelines, design criteria and requirements to safeguard the operation of the 3GPP System. This includes the capability to, e.g. fallback to non-AI/ML operation (i.e. not relying on inference process) whenever necessary not to negatively affect the NW and E2E performance.

This requires, as a minimum, the introduction of a common set of definitions to prevent any inconsistencies in the definition and use of AI/ML LCM across 3GPP WGs, to identify any misalignments/inconsistencies and to communicate such inconsistencies to WGs for better alignment within 3GPP across different AI/ML related initiatives.

NOTE: AI/ML models and associated algorithms are certainly implementation specific and therefore out of scope of this study.

# 5 AI/ML related activities in all Working Groups

## 5.1 General

This clause will investigate and identify AI/ML related activities of all working groups of Rel-18 features and Rel-19 studies and work items in TSG CT, TSG RAN and TSG SA Working Groups, as given in the list in Table 5.1-1.

Table 5.1-1: List of Rel-18 work items and Rel-19 studies and work items for AI/ML activities in 3GPP working groups

| Unique ID | Release | Title | Acronym | Approved SID/WID | TRs/TSs |
| --- | --- | --- | --- | --- | --- |
| SA WG1 |  |  |  |  |  |
| 920030 | Rel-18 | Stage 1 of AMMT | AIML\_MT | SP-220440 | TS 22.261 [6] |
| 950008 | Rel-19 | Study on AI/ML Model Transfer Phase2 | FS\_AIML\_MT\_Ph2 | SP-220439 | TR 22.876 [21] |
| 1000030 | Rel-19 | AI/ML Model Transfer Phase 2 | AIML\_MT\_Ph2 | SP-230514 | TS 22.261 [6] |
| SA WG2 |  |  |  |  |  |
| 980019 | Rel-18 | System Support for AI/ML-based Services | AIMLsys | SP-231278 | TS 23.501 [22]  TS 23.502 [23]  TS 23.503 [24]  TS 23.288 [8] |
| 980020 | Rel-18 | Enablers for Network Automation for 5G - phase 3 | eNA\_Ph3 | SP-230110 | TS 23.501 [22]  TS 23.502 [23]  TS 23.503 [24]  TS 23.288 [8] |
| 1020068 | Rel-19 | Study on Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML) | FS\_AIML\_CN | SP-241936 | TR 23.700-84 [4] |
| 1040033 | Rel-19 | Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML) | AIML\_CN | SP-240991 | TS 23.288 [8]  TS 23.501 [22]  TS 23.502 [23]  TS 23.503 [24]  TS 23.273 [16] |
| SA WG3 |  |  |  |  |  |
| 990042 | Rel-18 | Security aspects of enablers for Network Automation for 5G - phase 3 | eNA\_Ph3\_SEC  Rel-18 | SP-230155 | TS 33.501 [22] |
| 1030035 | Rel-19 | Study on security aspects of Core Network Enhanced Support for AIML | FS\_AIML\_CN\_SEC | SP-240509 | TR 33.784 [25] |
| 1060059 | Rel-19 | Security aspects of Core Network Enhanced Support for AIML | AIML\_CN\_SEC | SP-241957 | TS 33.501 [22] |
| SA WG4 |  |  |  |  |  |
| 950011 | Rel-19  (started in Rel-18) | Feasibility Study on Artificial Intelligence (AI) and Machine Learning (ML) for Media | FS\_AI4Media | SP-220328 | TR 26.927 [12]  TR 26.847 [26] |
| SA WG5 |  |  |  |  |  |
| 990119 | Rel-18 | AI/ML management | AIML\_MGT | SP-230335 | TS 28.105 [9]  TS 28.552 [27]  TS 32.425 [28]  TS 28.554 [29]  TS 28.541 [30] |
| 1020023 | Rel-18 | NEF Charging enhancement to support AI/ML in 5GS | AIMLsysNEF\_CH | SP-231706 | TS 32.254 [31]  TS 32.291 [32] |
| 1020007 | Rel-19 | Study on AI/ML management - phase 2 | FS\_AIML\_MGT\_Ph2 | SP-241567 | TR 28.858 [19] |
| 1060011 | Rel-19 | AI/ML management - phase 2 | AIML\_MGT\_Ph2 | SP-241944 | TS 28.105 [9]  TS 28.552 [27]  TS 28.554 [29]  TS 28.541 [30] |
| SA WG6 |  |  |  |  |  |
| 970036 | Rel-18 | Application Data Analytics Enablement Service | ADAES | SP-230275 | TS 23.436 [33]  TS 23.434 [35]  TS 23.558 [36] |
| 1010005 | Rel-19 | Study on application layer support for AI/ML services | FS\_AIMLAPP | SP-231182 | TR 23.700-82 [7] |
| 1040075 | Rel-19 | Application enablement for AI/ML services | AIML\_App | SP-241008 | TS 23.482 [34]  TS 23.436 [33]  TS 23.434 [35]  TS 23.558 [36] |
| 1060068 | Rel-19 | Application Data Analytics Enablement Service | TEI19\_ADAES | SP-241695 | TS 23.436 [33] |
| CT WG1 |  |  |  |  |  |
| 1050024 | Rel-19 | CT1 aspects of application enablement for AIML services | AIML\_App | CP-243310 | TS 24.559 [75]  TS 24.560 [76] |
| CT WG3 |  |  |  |  |  |
| 990008 | Rel-18 | CT WG3 aspects of System Support for AI/ML-based Services | AIMLsys | CP-230329 | TS 29.122 [37]  TS 29.513 [38]  TS 29.514 [39]  TS 29.517 [40]  TS 29.591 [41]  TS 29.519 [42]  TS 29.520 [43]  TS 29.521 [44]  TS 29.522 [45] |
| 990010 | Rel-18 | CT WG3 aspects of Enablers for Network Automation for 5G - phase 3 | eNA\_Ph3 | CP-240079 | TS 29.508 [48]  TS 29.122 [37]  TS 29.513 [38]  TS 29.517 [40]  TS 29.519 [42]  TS 29.520 [43]  TS 29.522 [45]  TS 29.523 [46]  TS 29.525 [47]  TS 29.551 [49]  TS 29.552 [50]  TS 29.574 [51]  TS 29.575 [52]  TS 29.576 [53]  TS 29.591 [41] |
| 1050010 | Rel-19 | Rel-19 Enhancements of Network Automation Enablers | eNetAE19 | CP-243078 | TS 29.508 [48]  TS 29.517 [40]  TS 29.520 [43]  TS 29.522 [45]  TS 29.552 [50]  TS 29.574 [51]  TS 29.575 [52]  TS 29.576 [53]  TS 29.591 [41] |
| 1050014 | Rel-19 | CT3 aspects of Core Network Enhanced Support for Artificial Intelligence (AI) and Machine Learning (ML) | AIML\_CN | CP-242247 | TS 29.508 [48]  TS 29.513 [38]  TS 29.517 [40]  TS 29.520 [43]  TS 29.522 [45]  TS 29.552 [50]  TS 29.574 [51]  TS 29.575 [52]  TS 29.576 [53]  TS 29.591 [41] |
| 1050024 | Rel-19 | CT3 aspects of application enablement for AIML services | AIML\_App | CP-243310 | TS 29.482 [77]  TS 29.549 [78]  TS 29.558 [79] |
| 1070004 | Rel-19 | CT3 aspects of Rel-19 Application Data Analytics Enablement Service | TEI19\_ADAES | CP-250074 | TS 29.549 [78] |
| CT WG4 |  |  |  |  |  |
| 990008 | Rel-18 | CT WG4 aspects of AIML | Rel-18 CT WG4 aspects of AI/ML | CP-230329 | TS 29.503 [54]  TS 29.504 [55]  TS 29.505 [56]  TS 29.510 [57]  TS 29.564 [58] |
| 990010 | Rel-18 | CT WG4 aspects of Enablers for Network Automation for 5G - phase 3 | eNA\_Ph3 | CP-240079 | TS 29.503 [54]  TS 29.510 [57]  TS 29.518 [59]  TS 29.536 [60]  TS 29.564 [58]  TS 29.571 [61] |
| 1040005 | Rel-19 | Study on Protocol for AI Data Collection from UPF | FS\_PAIDC\_UPF | CP-241025 | TR 29.889 [20] |
| 1050014 | Rel-19 | CT4 aspects of Core Network Enhanced Support for Artificial Intelligence (AI) and Machine Learning (ML) | AIML\_CN | CP-242247 | TS 24.080 [80]  TS 29.503 [54]  TS 29.510 [57]  TS 29.518 [59]  TS 29.570 [81]  TS 29.572 [82] |
| RAN WG1 |  |  |  |  |  |
| 1020093 | Rel-19 | Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface | NR\_AIML\_air | RP-250792 | TR 38.843 [3]  TS 38.300 [11]  TS 38.214 [62]  TS 38.215 [63]  TS 38.331 [64]  TS 38.305 [65]  TS 37.355 [66]  TS 38.455 [67]  TS 38.133 [68] |
| 1060078 | Rel 19 | Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface Phase 2 | NR\_AIML\_air\_Ph2 | RP-250308 | TR 38.843 [3] |
| RAN WG2 |  |  |  |  |  |
| 1020084 | Rel-19 | Study on Artificial Intelligence (AI)/Machine Learning (ML) for mobility in NR | FS\_NR\_AIML\_Mob | RP-242393 | TR 38.744 [70] |
| RAN WG3 |  |  |  |  |  |
| 941110 | Rel-18 | Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN | NR\_AIML\_NGRAN-Core | RP-233441 | TS 38.300 [11]  TS 38.401 [13]  TS 38.420 [14]  TS 38.423 [15] |
| 1020083 | Rel-19 | Study on enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN | FS\_NR\_AIML\_NGRAN\_enh | RP-240323 | TR 38.743 [69] |
| 1051124 | Rel-19 | Enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN | NR\_AIML\_NGRAN\_enh-Core | RP-250326 | TS 38.300 [11]  TS 38.401 [13]  TS 38.420 [14]  TS 38.423 [15]  TS 37.340 [83]  TS 37.480 [84]  TS 37.483 [85]  TS 38.470 [86]  TS 38.473 [87] |
| RAN WG4 |  |  |  |  |  |
| 1020093 | Rel-19 | Perf. part: Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface | NR\_AIML\_air | RP-250792 | TS 38.133 [68]  TS 38.300 [11]  TS 38.214 [62]  TS 38.215 [63]  TS 38.331 [64]  TS 38.305 [65]  TS 37.355 [66]  TS 38.455 [67] |

Table 5.1-1 comprises 3GPP AI/ML related activities where the consumer of AI/ML models can be either the 3GPP System (including RAN, Core, OAM and Application Enablement) itself to optimize 3GPP-specified functionalities and capabilities (i.e. equivalent to the common term "AI for Network"), or the applications and services that the 3GPP System provides support for (i.e. equivalent to the common term "Network for AI").

Based on the above, the AI/ML related activities of 3GPP can be classified in the following two categories:

- 'AI for Network': Work to optimize the 3GPP System performance by employing AI/ML on 3GPP-specified functionalities and capabilities.

- 'Network for AI': Work to provide 3GPP System support for AI/ML-based applications and services.

Illustrative examples of 3GPP activities in Table 5.1-1 falling into the category 'AI for Network' include NR\_AIML\_air and AIML\_MGT while 'Network for AI' activities include AIMLsys and AIML\_App. In addition, certain items may address both 'AI for Network' and 'Network for AI' aspects, e.g. eNA\_Ph3, AIML\_CN.

## 5.2 AI/ML related activities in TSG SA & CT Working Groups

### 5.2.1 AI/ML related terminology

#### 5.2.1.1 TSG SA WG2

The following definitions are provided in clause 3 of TS 23.288 [8]:

- Analytics Accuracy Information: Represent a performance measure of an analytics ID provided by an NWDAF containing AnLF, which is composed of the number of correct predictions of the analytics ID out of all predictions and the corresponding number of samples.

- ML Model Accuracy Information: Represent a performance measure of a ML Model provided by an NWDAF containing MTLF, which is composed of the number of correct predictions by the ML Model out of all predictions and the corresponding number of samples.

- Analytics Feedback Information: Indicates that the consumer NF has taken action(s) influenced by the previously provided analytics, which may or may not affect the ground truth data.

The following definitions are provided in clause 5 of TS 23.288 [8]:

- Analytics logical function (AnLF): A logical function in NWDAF, which performs inference, derives analytics information (i.e. derives statistics and/or predictions based on Analytics Consumer request) and exposes analytics service i.e. Nnwdaf\_AnalyticsSubscription or Nnwdaf\_AnalyticsInfo.

- Model Training logical function (MTLF): A logical function in NWDAF, which trains Machine Learning (ML) models and exposes new training services (e.g. providing trained ML Model) as defined in clause 7.5 and clause 7.6.

The following definitions are provided in clause 3 of TR 23.700-84 [4]:

- Horizontal Federated Learning (HFL): a federated learning technique without exchanging/sharing local data set, wherein the local data set in different FL clients for local model training have the same feature space for different samples (e.g. UE IDs).

- Vertical Federated Learning (VFL): a federated learning technique without exchanging/sharing local data set, wherein the local data set in different VFL Participant for local model training have different feature spaces for the same samples (e.g. UE IDs).

- Label: A label is the training objective in supervised machine learning.

- VFL Server: An NWDAF or AF that integrates local training results, computes gradient information or loss information and send them to VFL client(s) for the local ML model update in VFL training process. It also coordinates the VFL training process by discovering and selecting VFL clients. In VFL inference process, The VFL server aggregates local inference results from VFL clients to generate the final VFL inference result and sends the final VFL inference result to the consumer. Only one VFL server may exist for each VFL process.

- VFL Client: An NWDAF or AF that holds the local dataset and performs local training and inference as asked by VFL Server. There can be multiple VFL Clients in VFL training and inference.

The following definitions are provided in clause 6.15 of TR 23.700-84 [4]:

- VFL active participant: A VFL function that owns part of an ML model for an analytic ID and knows the labels for the ML model. The active participant is the main function for training an ML model for an analytic ID.

- VFL passive participant: A VFL function that owns part of an ML model for an analytic ID but does not know the labels of the ML model but is able to collect local data for one or more features.

#### 5.2.1.2 TSG SA WG5

The following definitions are provided in clause 3 of TS 28.105 [9]:

- ML model: a manageable representation of an ML model algorithm.

NOTE 1: An ML model algorithm is a mathematical algorithm through which running a set of input data can generate a set of inference output.

NOTE 2: An ML model algorithm is proprietary and not in scope for standardization and therefore not treated in this specification.

NOTE 3: An ML model may include metadata. Metadata may include e.g. information related to the trained model and applicable runtime context.

- ML model training: a process performed by an ML training function to take training data, run it through an ML model algorithm, derive the associated loss and adjust the parameterization of that ML model iteratively based on the computed loss and generate the trained ML model.

- ML model initial training: a process of training an initial version of an ML model.

- ML model re-training: a process of training a previous version of an ML model and generate a new version.

NOTE 4: A new version of a trained ML model supports the same type of inference as the previous version of the ML model, i.e. the data type of inference input and data type of inference output remain unchanged between the two versions of the ML model, but parameter values might be different for the re-trained model.

- ML model pre-specialized training: the process of training an ML model on a dataset not specific to any type of inference.

NOTE 5: The pre-specialised trained model supports an inference scope that may be potentially adapted to support a list of inference types, such as MDA types in MDA, analytics types in NWDAF, type of AI/ML supported functions in NG-RAN, or vendor-specific extensions.

- ML model Fine-tuning: the process of training a pre-specialised trained ML model to narrow down its inference scope to a new single inference type, generating a new ML model.

NOTE 6: The inference scope refers to a list of inference types that the ML model may be potentially adapted to support.

NOTE 7: The type of inference represents the specific type of ML inference supported by the model, such as MDA types in MDA, Analytics types in NWDAF, type of AI/ML supported functions in NG-RAN, or vendor-specific extensions.

- Distributed training: a process of distributing the training workload across multiple ML training functions.

- Federated Learning: a distributed machine learning approach where the ML model is trained collaboratively by multiple ML training functions. This includes multiple FL clients, which perform training on local data, and one FL server, which aggregates model outcomes from the clients iteratively without exchanging data samples.

- Horizontal Federated Learning: a federated learning technique without exchanging/sharing local data set, wherein the local data set in different HFL clients for local model training have the same feature space for different samples.

- FL Client: a training function that trains an ML model on local data and shares only the model updates with the FL server, preserving data privacy.

- FL Server: a function that aggregates the ML model updates from FL Clients to produce a global ML model.

- Reinforcement Learning: a machine learning approach in which an RL agent interacts with an RL environment by observing states, taking actions and receiving rewards as feedback. The RL agent learns a decision-making policy by maximizing rewards over time through trial and error.

- ML model joint training: a process of training a group of ML models.

- ML training function: a logical function with ML model training capabilities.

- ML knowledge: the implicit information representing the experience gained by the training an ML Model

NOTE 8 Examples of experience include statistics (e.g. a distribution) or summaries (e.g. tables) indicating the ML model’s recommended output for a given set of input data.

- AI/ML inference function: a logical function that employs an ML model to conduct inference.

- ML model testing: a process of evaluating the performance of an ML model using testing data different from data used for model training and validation.

- ML model joint testing: a process of evaluating the performance of a group of ML models using testing data different from data used for model training and validation.

- ML testing function: a logical function with ML model testing capabilities.

- AI/ML inference: a process of running a set of input data through a trained ML model to produce set of output data, such as predictions.

NOTE 5: The inference represents the process to realize the AI capabilities by utilizing a trained ML model and other AI enablers if needed, hence the AI/ML prefix is used when referring to inference as compared to training and testing.

- AI/ML inference function: a logical function that employs trained ML model(s) to conduct inference.

- AI/ML inference emulation: running the inference process to evaluate the performance of an ML model in an emulation environment before deploying it into the target environment.

- ML model deployment: a process of making a trained ML model available for use in the target environment.

**-** ML model loading: a process of making a trained ML model available to an inference function.

**-** AI/ML activation: a process of enabling the inference capability of an AI/ML inference function.

**-** AI/ML deactivation: a process of disabling the inference capability of an AI/ML inference function.

#### 5.2.1.3 TSG SA WG6

The following definitions are provided in clause 3 of TR TS 23.482 [34]:

- ML model: According to TS 28.105 [9], mathematical algorithm that can be "trained" by data and human expert input as examples to replicate a decision an expert would make when provided that same information.

- ML model lifecycle: The lifecycle of an ML model aka ML model operational workflow consists of a sequence of ML operations for a given ML task / job (such job can be an analytics task or a VAL automation task). This definition is aligned with the 3GPP definition on ML model lifecycle according to TS 28.105 [9].

- ML model training: According to TS 28.105 [9], ML model training includes capabilities of an ML training function or service to take data, run it through an ML model, derive the associated loss and adjust the parameterization of that ML model based on the computed loss.

- ML model inference: According to TS 28.105 [9], ML model training includes capabilities of an ML model inference function that employs an ML model and/or AI decision entity to conduct inference.

- AI/ML intermediate model: For federated learning, members need to train models for multiple rounds, intermediate models indicate the model which do not meet the required training rounds and/or meet the requirements of the federated training.

- AIMLE service: An AIMLE service is an AIMLE capability which aims assisting in performing or enabling one or more AIML operations.

- AI/ML client: an application layer entity (also referred as ML client) which is an AI/ML endpoint and performs client-side operations (e.g. related to the ML model lifecycle). Such AI/ML client can be a VAL client or AIML enabler client and may be configured e.g. to provide ML model training and inference locally e.g. at the VAL UE side.

- AIMLE client set identifier**:** an identifier of the set of selected AIMLE clients.

- AI/ML server: an application layer entity which is an AI/ML endpoint and performs server-side operations (e.g. related to the ML model lifecycle). Such AI/ML server can be a VAL server or AIML enabler server.

- FL member: An FL member or participant is an entity which has a role in the FL process. An FL member can be an FL client performing ML model training, or an FL server performing aggregation/collaboration for the FL process.

- FL client: An FL member which locally trains the ML model as requested by the FL server. Such FL client functionality can be at the network (e.g. AIMLE server with FL client capability) or at the device side (e.g. AIMLE client with FL client capability).

- FL server: An FL member which generates global ML model by aggregating local model information from FL clients.

- Split AI/ML operation pipeline**:** A Split AI/ML operation pipeline is a workflow for ML model inference in which AI/ML endpoints are organized and collaborate to process ML models in sequential stages, where processing at each stage involves ML model inference on the output of the previous stage.

### 5.2.2 AI/ML related activities

#### 5.2.2.1 Rel-18 SA WG1 WID - AI/ML model transfer in 5GS (AIML\_MT)

##### 5.2.2.1.1 Description

The objective of this work item is to specify performance requirements (for end-to-end latency, experienced data rate, communication service availability) and service requirements (for AI/ML QoS management, AI/ML model /data distribution/transfer, network performance and resource utilization monitoring/prediction) for 5GS to support the following AI/ML operations for various applications (e.g. image/speech recognition, media editing/enhancements, robot control, automotive):

- AI/ML operation splitting between AI/ML endpoints.

- AI/ML model/data distribution and sharing over 5G system.

- Distributed/Federated Learning over 5G system.

##### 5.2.2.1.2 Activities summary

The three application AI/ML operations in 5.2.2.1.1 the 5G system can support were specified in clause 6.40.1 of TS 22.261 [6] as follows:

- AI/ML operation splitting between AI/ML endpoints: The AI/ML operation/model is split into multiple parts according to the current task and environment. The intention is to offload the computation-intensive, energy-intensive parts to network endpoints, whereas leave the privacy-sensitive and delay-sensitive parts at the end device. The device executes the operation/model up to a specific part/layer and then sends the intermediate data to the network endpoint. The network endpoint executes the remaining parts/layers and feeds the inference results back to the device.

- AI/ML model/data distribution and sharing: Multi-functional mobile terminals might need to switch the AI/ML model in response to task and environment variations. The condition of adaptive model selection is that the models to be selected are available for the mobile device. However, given the fact that the AI/ML models are becoming increasingly diverse and with the limited storage resource in a UE, it can be determined to not pre-load all candidate AI/ML models on-board. Online model distribution (i.e. new model downloading) is needed, in which an AI/ML model can be distributed from a Network endpoint to the devices when they need it to adapt to the changed AI/ML tasks and environments. For this purpose, the model performance at the UE needs to be monitored constantly.

- Distributed/Federated Learning over 5G system: The cloud server trains a global model by aggregating local models partially-trained by each end devices. Within each training iteration, a UE performs the training based on the model downloaded from the AI server using the local training data. Then the UE reports the interim training results to the cloud server via 5G UL channels. The server aggregates the interim training results from the UEs and updates the global model. The updated global model is then distributed back to the UEs and the UEs can perform the training for the next iteration.

It is worth emphasizing that the above descriptions refer to AI/ML operations over the application layer. The service requirements and performance requirements for AI/ML model transfer over the application layer in 5GS with direct network connection are specified in clause 6.40.2.1 and in clause 7.10.1 of TS 22.261 [6], respectively.

#### 5.2.2.2 Rel-19 SA WG1 SID - AI/ML Model Transfer Phase 2 (FS\_AIML\_MT\_Ph2)

##### 5.2.2.2.1 Description

The objective of this study is to explore new use cases and potential service and performance requirements to support efficient AI/ML operations using direct device connections. This includes:

- Distributed AI training and inference based on direct device connections, such as traffic KPIs, various QoS and functional requirements for sidelink transmission.

- Considerations for charging and security aspects.

##### 5.2.2.2.2 Activities summary

In this study, TR 22.876 [21] described study the use cases with potential functional and performance requirements to support efficient AI/ML operations over the application layer using direct device connection for various applications e.g. auto-driving, robot remote control, video recognition, etc. The agreed activities which were progressed in normative phase are described in clause 5.2.2.3.2.

#### 5.2.2.3 Rel-19 SA WG1 WID - AI/ML Model Transfer Phase 2 (AIML\_MT\_Ph2)

##### 5.2.2.3.1 Description

The objective of this work item is to specify KPI and functional requirements for 5GS to support the AIML data transfer by leveraging direct device connection under 5G network control. These objectives were derived based on outcome of Rel-19 study in SA WG1 that relates to how the 5GS supports the transmissions of AI/ML-based services over the application layer. The study addressed use cases and potential performance requirements for 5G system support of application layer Artificial Intelligence (AI)/Machine Learning (ML) model distribution and transfer (download, upload, updates, etc.) and identified traffic characteristics of AI/ML model distribution, transfer and training for various applications, e.g. video/speech recognition, robot control, automotive, other verticals.

##### 5.2.2.3.2 Activities summary

The service requirements and performance requirements for AI/ML model transfer over the application layer in 5GS with direct device connection are specified in clause 6.40.2.2 and in clause 7.10.2 of TS 22.261 [6], respectively.

#### 5.2.2.4 Rel-18 SA WG2 WID - Enablers for Network Automation for 5G - phase 3 (eNA\_Ph3)

##### 5.2.2.4.1 Description

The objective of this work item is to further enhance NWDAF, based on what has been specified in the previous releases to allow 5GS to support network automation. This work item focuses on architecture enhancement, new scenarios and the necessary inputs and outputs to the NWDAF based on the conclusions of the study in Rel-18. The work focuses on 10 key aspects as follows:

- Improve correctness of NWDAF analytics.

- NWDAF-assisted application detection.

- Data and analytics exchange in roaming case.

- Enhancements on Data collection and Storage.

- Enhancements on trained ML Model sharing.

- NWDAF-assisted URSP.

- Enhancements on QoS Sustainability analytics.

- Supporting Federated Learning in 5GC.

- Enhancement of NWDAF with finer granularity of location information.

- Interactions with MDAS/MDAF.

##### 5.2.2.4.2 Activities summary

5.2.2.4.2.1 AIML related LCM activities

**Data collection/storage/exposure**

Analysis of data collection activities as part of eNA, eNA\_Ph2 work:

- Data collection in TS 23.288 [8] refers to data collected by the NWDAF and DCCF. The data are collected for the purpose of analytics generation and training of ML models. NWDAF/DCCF/MFAF collects data from NFs/AFs, OAM and UE application.

- Data Collected and Analytics output information may be stored at ADRF.

As depicted in Figure 4.2.0-1 of TS 23.288 [8], the 5G System architecture allows NWDAF to collect data from any 5GC NF. The NWDAF is allowed to collect data from any 5GC NF directly and retrieve the management data from OAM by invoking OAM services as defined in clause 5.11 of TR 23.288 [8].

As defined in clause 4.2.0 of TS 23.288 [8], the NWDAF is allowed cdata from any 5GC NF using a DCCF or MFAF as defined in Figure 5A.3.2-1in TS 23.288 [8] and retrieve the management data from OAM by invoking OAM services as defined in clause 5.11 of TR 28.871 [73].As defined in clause 5A of TS 23.288 [8], Each Event Notification received from a Data Source NF is sent to the DCCF which propagates it to all Data Consumers / Notification Endpoints specified by the Data Consumers or determined by the DCCF. As defined in clause 5B of TS 23.288 [8], the ADRF offers services that enable a consumer to store, retrieve and delete data, analytics and ML Models. ML Model(s) may be stored in the ADRF by a consumer sending the ADRF a storage request containing the ML Model or ML Model address to be stored.

Analysis of Data Collection activities as part of eNA\_Ph3 work:

- Exposure of input data for analytics is allowed from VPLMN to HPLMN and vice versa.

- Data Processing enhancements at Data stored at ADRF.

**AI/ML Model Training**

Analysis AI/ML model training activities as part of eNA, eNA\_Ph2 work:

- MTLF trains an ML model for an Analytics output requested by an AnLF. Trained ML model is assigned an ML Model Identifier and may be stored in ADRF.

Analysis of AI/ML Model Training activities as part of eNA\_Ph3 work:

- Trained ML model sharing between different NWDAF vendors by defined ML Model Interoperability Indication/Information.

- Support of Model Training using Horizontal Federated Learning.

**AI/ML Model Inference**

Analysis of AI/ML Model inference activities as part of eNA\_Ph3 work:

- AnLF exposes analytics information to consumers (e.g. NFs). Analytics information may be stored at an ADRF.

- Analytics from multiple NWDAF can be aggregated.

- Analytics content can be transferred between NWDAFs, including roaming scenarios.

**Performance evaluation and accuracy monitoring**

Analysis of performance evaluation and accuracy monitoring activities as part of eNA\_Ph3 work:

- NWDAF (either AnLF or MTLF) supporting performance monitoring of a trained ML model.

5.2.2.4.2.2 AI/ML functional entities.

As part of eNA, eNA\_Ph2 and eNA\_Ph3 work the following functional entities have been defined:

- NWDAF (Network Data Analytics Function) is defined in TS 23.288 [8] with main function to generate analytics (statistics and/or predictions) for one or more network events. NWDAF is defined by two logical functions.

- AnLF (Analytics Logical Function): Derives and exposes analytics information (statistics or predictions).

- MTLF (Model Training Logical Function): Trains Machine Learning (ML) models and exposes new training services (e.g. providing trained ML Model).

- DCCF (Data Collection & Coordination Function) is defined in TS 23.288 [8] with main functionality for analytics collection from NWDAFs and data collection from multiple NF(s), AF and OAM.

- MFAF (Messaging Framework Adaptor NF) is defined in TS 23.288 [8] and is part of the DCCF architecture. MFAF offers 3GPP defined services that allow the 5GS to interact with a Messaging Framework.

- ADRF (Analytics Data Repository Function) is defined in TS 23.288 [8] with main functionality for storing and retrieving collected data and analytics.

#### 5.2.2.5 Rel-18 SA WG2 WID - System Support for AI/ML-based Services (AIMLsys)

##### 5.2.2.5.1 Description

This work item implements the conclusions of the Rel-18 study on the 5GS architectural and functional extensions to enable 5GS to assist the Application AI/ML operations. The normative text is defined based on the agreed conclusions on 6 key issues, ensuring consistency with other 5GS features. The agreed conclusions focus on the following aspects:

- Monitoring of network resource utilization to support the Application AI/ML operations.

- Exposure of 5GC information to authorized 3rd party for Application AI/ML operations.

- Enhancement of external parameter provisioning in 5GC to assist the Application AI/ML operations.

- Enhancement in 5GC to enable Application AI/ML traffic transport.

- Enhancement of QoS and Policy control to support Application AI/ML data transport over 5GS.

- 5GS assistance to federated learning operation.

##### 5.2.2.5.2 Activities summary

This work item specifies a list of principles that apply when the 5GS assists the AI/ML operation at the application layer as specified in clause 5.46 of TS 23.501 [22], namely:

- AF requesting 5GS assistance to AI/ML operations in the application layer shall be authorized by the 5GC using the existing mechanisms.

- Application AI/ML decisions and their internal operation logic reside at the AF and UE application client and is out of scope of 3GPP.

- Based on application logic, it is the application decision whether to request assistance from 5GC, e.g. for the purpose of selection of Member UEs that participate in certain AI/ML operation.

The activities of this work item are limited to providing assistance to AI/ML-based applications when the participating UEs are not roaming and the AI/ML operations in the application layer are conducted within a single slice. Policy and charging control as defined in TS 23.503 [24] are assumed to be used for traffic related to application AI/ML operations.

The overall objective of this item is to provide assistance by the 5GC to AI/ML operations in the application layer, which are described in clause 5.2.2.1 and specified in clause 6.40 of TS 22.261 [6]. A brief description of the specified capabilities in this work item can be found below, with further details provided in clause 5.46 of TS 23.501 [22] and clause 11.1 of TR 21.918 [2] and the references therein:

- Planned Data Transfer with QoS: this capability is used to enable the AF to negotiate a variable time window for the planned AI/ML operation e.g. application data transfer with specific QoS requirements and operational conditions via the support of the NEF.

- Enhanced external parameter provisioning: this capability enables an AF hosting an AI/ML based application to provision enhanced Expected UE Behaviour parameters and/or Application-Specific Expected UE Behaviour parameter(s) to the 5GC by including corresponding confidence and/or accuracy levels to the expected parameters, which UDM could check against a threshold.

- Member UE selection assistance functionality: this capability provided by NEF is used to assist the AF to select member UE(s) for AI/ML application operations (e.g. Federated Learning) according to the AF's request including a list of target member UEs and a set of filtering criteria.

- Multi-member AF session with required QoS: this capability enables the NEF to map a request for Multi-member AF session with required QoS to individual requests for AF session with required QoS per UE address and interact with each of the UE's serving PCFs on a per AF session basis.

- End-to-end data volume transfer time analytics: this analytics provides the consumer (e.g. AF, NEF) with analytics (i.e. statistics, predictions or both) referring to a time delay for completing the transmission of a specific data volume from UE to AF, or from AF to UE. The data volume may be the expected or observed data volume from UE to AF or from AF to UE.

- Enhanced NEF monitoring events: new NEF monitoring events are specified relevant to the operation of AI/ML based operations, namely session inactivity time, traffic volume exchanged between the UE and the AF and UL/DL consolidated data rate which is the aggregated data rate across all traffic flows corresponding to the list of UE addresses of the Multi-member AF session with required QoS.

5.2.2.5.2.1 AI/ML related LCM activities

No AI/ML related LCM activities or functional entities were specified as part of this work. Instead, the activities summarized in clause 5.2.2.5.2 specify 5GS support for AI/ML related LCM activities (e.g. AI/ML model training and inference) assumed to be conducted at the application layer.

#### 5.2.2.6 Rel-19 SA WG2 SID - Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML) (FS\_AIML\_CN)

##### 5.2.2.6.1 Description

The aim of this study is to investigate and identify potential architectural and system-level enhancements to support AI/ML enhancements. Specifically, the objectives include:

- AI/ML Cross-Domain Coordination Aspects: Investigate enhancements to support AI-enabled RAN based on the conclusions of the RAN study in TR 38.843 [3]. This task will discuss whether and how to support cross-domain (i.e. UE, RAN, 5GC, OAM and AF) collaborative AI/ML mechanisms for the aspects described below:

- Enhancements to LCS for AI/ML-Based Positioning: Examine whether and how to consider enhancements to LCS to support AI/ML-based positioning.

- Collaborative AI/ML Operations for Vertical Federated Learning (VFL): Determine potential enhancements needed to enable the 5G system to assist in collaborative AI/ML operations involving 5GC/NWDAF and/or AF for "Vertical Federated Learning (VFL)." This work will be based solely on and limited to the scope of justified use cases.

- Enhancements to Support NWDAF-Assisted Policy Control and Address Network Abnormal Behaviour:

- Investigate additional support needed to enhance 5GC NF operations (i.e. policy control and QoS) assisted by NWDAF. This task will first identify specific use cases to define the appropriate scope. It will analyse the impacts on NWDAF (e.g. the need to understand specific NF functionality) and the compatibility of new solutions with existing analytics to determine the necessity and benefits of new solutions.

- Study the prediction, detection, prevention and mitigation of network abnormal behaviours, such as signalling storms, with the assistance of NWDAF.

NOTE: The outcome of this study was used to support the AIML\_CN Rel-19 SA WG2 WID, see clause 5.2.2.7.

#### 5.2.2.7 Rel-19 SA WG2 WID - Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML) (AIML\_CN)

##### 5.2.2.7.1 Description

The objective of this work item is to specify the following enhancements to 5GS as per the conclusions reached within the Rel-19 study for the following aspects:

- Enhancements to LCS to Support Direct AI/ML-Based Positioning:

- LMF Enhancements: The LMF will be enhanced to perform location calculations based on an ML model. The triggers for data collection and model training within the LMF will be implementation-specific.

- MTLF and LMF Enhancements: Both the MTLF and the LMF will be enhanced to facilitate ML model training for AI/ML-based positioning.

- Procedure Development: Related procedures for data collection will be developed in coordination with RAN WGs.

- 5GC Support for Vertical Federated Learning:

- 5GC Enhancements: The 5GC will be enhanced to support vertical federated learning (VFL), a technique that does not involve exchanging or sharing local datasets or ML models, in the following scenarios:

- VFL among NWDAFs within a single PLMN.

- VFL between NWDAF(s) within a single PLMN and AF(s).

- NWDAF-Assisted Policy Control and QoS Enhancement:

- Assistance Information: Based on PCF requests, the NWDAF may provide assistance information to the PCF to aid in the determination and modification of QoS parameters.

- NWDAF Enhancements to Support Network Abnormal Behaviours Mitigation and Prevention:

- Signalling Storm Mitigation: The NWDAF will support signalling storm mitigation and prevention by providing analytics related to the detection and prediction of signalling storms caused by massive signalling from UEs and/or NFs.

##### 5.2.2.7.2 Activities summary

5.2.2.7.2.1 AIML related LCM activities

As part of the AIML\_CN work the following enhancements are supported:

**Data collection/exposure**

Data collection from Direct AIML positioning. Data is collected to train an ML model for LMF-based AIML positioning and to support inference.

**AI/ML model training**

- ML model training for LMF-side Direct AIML positioning. LMF or MTLF support model training for LMF-Side Direct AIML positioning.

- Collaborative ML model training using Vertical Federated Learning. Vertical Federated Learning is supported between NWDAFs or cross-domain between NWDAF and Application Functions

**AI/ML model inference**

- LMF is supports inference based on using trained ML Model for Direct AIML positioning provisioned by an MTLF or locally trained at LMF.

**Performance evaluation and accuracy monitoring**

- LMF or MTLF evaluates the performance of a trained ML Model by comparing the inference output against ground truth information.

#### 5.2.2.8 Rel-18 SA WG3 WID - Security aspects of enablers for Network Automation for 5G - phase 3 (eNA\_SEC\_PH3)

##### 5.2.2.8.1 Description

The main objective of this work is to produce normative specification based on the conclusions from Rel-18 study. More specifically, the following objectives are expected to be specified:

- Protection of data and analytics exchange in roaming case.

- Security for AI/ML model storage and sharing.

- Authorization of selection of participant NWDAF instances in the Federated Learning group.

##### 5.2.2.8.2 Activities summary

As part of this work, enhancements were specified for

* The protection of data and analytics exchange in roaming case including authorization and anonymization of data/analytics. Authorization at data and analytics level is enforced by the roaming entry NWDAF producer. The roaming entry NWDAF producer is responsible to control the amount of exposed data/analytics and to abstract or hide internal network aspects in the exposed data/analytics.

- The authorization for selecting participant NWDAF instances in the Federated Learning (FL) group, using token-based authorization.

- The procedure for secured and authorized AI/ML model sharing between different vendors and AI/ML model storing.

#### 5.2.2.9 Rel-19 SA WG3 SID - Security aspects of Core Network Enhanced Support for AIML (FS\_AIML\_CN\_SEC)

##### 5.2.2.9.1 Description

The objectives of this study are the following:

- Security Aspects on Enhancements to LCS: Study security aspects on enhancements to LCS to support AI/ML-based positioning, considering the conclusions in TR 38.843 [3] and TR 23.700-84 [4].

- Security Aspects of Cross-Domain Vertical Federated Learning (VFL):

- Authorization of VFL Group Members: Examine the authorization of members of the VFL group.

- Security Aspects of Enhancements on SA WG2 Architecture: Investigate security aspects of enhancements on SA WG2 architecture to support VFL.

NOTE: The outcome of this study was used to support the AIML\_CN\_SEC Rel-19 SA WG3 WID, see clause 5.2.2.16.

#### 5.2.2.10 Rel-19 SA WG4 SID - Artificial Intelligence (AI) and Machine Learning (ML) for Media (FS\_AI4Media)

##### 5.2.2.10.1 Description

The primary objective of this study item is to identify relevant interoperability requirements and implementation constraints of AI/ML in 5G media services. The specific objectives include:

- Use Cases for Media-Based AI/ML Scenarios: List and describe the use cases for media-based AI/ML scenarios, based on those defined in TR 22.874 [5].

- Media Service Architecture and Service Flows: Describe the media service architecture and relevant service flows for the scenarios. Identify the impacts on the architecture for each use case, including any potential gaps with existing 5G media service architectures. Also, describe the model operation configurations for each use case, including split AI/ML operations and identify where certain AI/ML operations occur.

- Data Formats and Protocols: Identify and document the available data formats and suitable protocols for the exchange of different data components of various AI/ML models, such as model data, metadata, media data and intermediate data necessary for such model operation configurations. Investigate the data traffic characteristics of these data components for delivery over the 5G system, including any needs and potentials for data rate reduction.

- Key Performance Indicators (KPIs): Identify and study key performance indicators for such scenarios, based on the initial considerations in TS 22.261 [6]. Emphasize the use cases, model operation configurations and data components identified in earlier objectives, focusing on objective performance metrics considering the identified KPIs.

- Normative Work and Collaboration: Identify potential areas for normative work as the next phase. Communicate and align with SA WG2 and other potential 3GPP working groups on relevant aspects related to the study.

#### 5.2.2.11 Rel-18 SA WG5 WID - AI/ML management (AIML\_MGT)

##### 5.2.2.11.1 Description

The objective of this work is to specify the AI/ML management capabilities, including use cases, requirements and solutions for each phase of the AI/ML operational workflow for managing the AI/ML capabilities in 5GS (i.e. management and orchestration, 5GC and NG-RAN), including:

- Management capabilities for ML training phase, which includes control of producer-initiated ML training, data management for ML training, performance evaluation for ML training, ML entity validation, ML context management, ML entity capability discovery and ML entity testing.

- Management capabilities for ML deployment phase, including management of ML entity loading.

- Management capabilities for AI/ML inference phase.

To describe the deployment scenarios of the AI/ML management capabilities, with consideration of alignment with other relevant 3GPP WGs (e.g. RAN WG3, SA WG2) and ETSI ISG ZSM.

##### 5.2.2.11.2 Activities summary

5.2.2.11.2.1 ML model life cycle management (LCM)

Rel-18 specification in TS 28.105 [9] addressed the AI/ML LCM management capabilities (including wide range of use cases, corresponding requirements (stage 1) and solutions (stage 2 NRMs & stage 3 OpenAPIs) for the ML model, including ML model training (which also includes validation), ML model testing, AI/ML inference emulation, ML model deployment and AI/ML inference steps of the lifecycle. The specification defined operational workflow as shown in Figure 5.2.2.11.2.1-1 below highlighting the main steps of an ML model lifecycle.



Figure 5.2.2.11.2.1-1: ML model lifecycle

5.2.2.11.2.2 ML model lifecycle management capabilities

Each step in the ML model lifecycle, defined in TS 28.105 [9] (see clause 6.1) i.e. the ML model training, ML model testing, AI/ML emulation, ML model deployment and AI/ML inference correspond to number of dedicated management capabilities. The specified capabilities are developed based on corresponding use cases and requirements.

5.2.2.11.2.3 AI/ML functionalities management scenarios (relation with managed AI/ML features)

The Rel-18 specification TS 28.105 [9] (see clause 4a.2) also documented AI/ML functionalities management scenarios in relation with managed AI/ML features which describe the possible locations of ML training function and AI/ML inference function involving the various 3GPP system domains.

#### 5.2.2.12 Rel-19 SA WG5 SID - AI/ML management - phase 2 (FS\_AIML\_MGT\_Ph2)

##### 5.2.2.12.1 Description

The objectives of this study item include:

- Continuation of AI/ML Studies: Continue the study on AI/ML emulation, AI/ML inference coordination and ML knowledge transfer that are left over from Rel-18.

- Management Aspects of AI/ML functionalities defined by other 3GPP WGs:

- AI/ML Model Transfer/delivery in RAN: Study the management aspects (LCM, CM and PM) of AI/ML model transfer in RAN.

- 5GS Support for AI/ML-Based Services: Investigate the management aspects of 5GS support for AI/ML-based services, as defined in SA WG2.

- Management Aspects of AI/ML Functionalities Defined by SA WG5:

- Management Data Analytics (MDA): Study the management aspects (LCM, CM and PM) of AI/ML functionalities defined by SA WG5, including MDA Phase 3.

- AI/ML Management and Operation Capabilities: Investigate the AI/ML management and operation capabilities to support different types of AI/ML technologies needed for AI/ML in 5GS, such as Federated Learning, Reinforcement Learning, Online and Offline Training, Distributed Learning and Generative AI.

- Sustainability Aspects of AI/ML:

- Energy consumption/efficiency impacts: Evaluate the energy consumption and efficiency impacts associated with AI/ML solutions for all operational phases (training, emulation, deployment, inference).

- Trustworthiness Aspects Related to AI/ML Functionalities in 5GS:

- Concept of Trustworthiness: Further study the concept of trustworthiness for AI/ML in the context of OAM.

- Data for Trustworthiness Indicators: Identify and analyse data (e.g. measurements, events) to support the calculation of trustworthiness indicators.

The outcome of this study was used to support the Rel-19 SA WG5 WID on AI/ML management phase 2 (AIML\_MGT\_Ph2).

#### 5.2.2.13 Rel-19 SA WG6 SID - Application layer support for AI/ML services (FS\_AIMLAPP)

##### 5.2.2.13.1 Description

The objective of this study is to enable support for AI/ML services at the application enablement layer. This includes the following:

- Analysis of Rel-18 and Rel-19 Requirements: Analyse the requirements in TS 22.261 [6] related to AI/ML model distribution, transfer and training. Identify key issues and develop corresponding architectural requirements at the application enablement layer, along with potential enhancements to the application layer architecture.

- Architectural and Functional Implications: Study the architectural and functional implications on existing SA WG6 application enablers (e.g. ADAES, other SEAL services, EDGEAPP) for supporting AI/ML lifecycle operations. This includes operations such as data collection, data preparation, training, inference and federated learning for ML models used in ADAE layer analytics.

- Potential Solutions and APIs: Identify potential solutions, including information flows and developer-friendly application enablement APIs, to satisfy the architectural requirements and enhancements identified in the previous points.

- Impact on Deployments and Business Models: Investigate the possible impacts of application layer support for AI/ML services on different deployments and business models.

##### 5.2.2.13.2 Activities summary

In this study, TR 23.700-82 [7] described the AI/ML enablement capabilities for supporting vertical use cases. The agreed AIMLE activities which were progressed in normative phase are described in clause 5.2.2.14.2.

#### 5.2.2.14 Rel-19 SA WG6 WID - Application enablement for AI/ML services (AIML\_App)

##### 5.2.2.14.1 Description

The objectives of this work include the following:

Develop Stage 2 normative technical specification for AIML enablement service as a new SEAL service, based on the key issues, architecture, solutions and conclusions captured in TR 23.700-82 [7]. The Stage 2 normative technical specification will include the following aspects:

- Architecture requirements, deployment models and application architecture for AIML service enablement over 3GPP networks.

- Procedures, information flows and APIs supporting concluded solutions related to AIML enablement capabilities for AI/ML, FL (e.g. Vertical FL among VAL UEs, Horizontal FL), Transfer Learning. Such capabilities include:

- Support AIML client management (e.g. registration, discovery) and selection.

- Support AIML service lifecycle management aspects (e.g. training, inference, data management).

- Support AIML operation split and ML model distribution operations.

- Support AIML operations in edge / distributed deployments.

- Procedures information flows and APIs supporting concluded solutions for application layer support capabilities related to new ADAE analytics services. Such new analytics services include:

- DN Energy Analytics.

- Analytics for supporting FL.

Identify potential enhancements to other enablement frameworks (e.g. SEAL, EGDEAPP and CAPIF) based on the specified solutions for the above objectives.

##### 5.2.2.14.2 Activities summary

5.2.2.14.2.1 AI/ML Functional Entities

In AIML\_App, the following logical entities have been introduced within SEAL framework:

- AIMLE server includes of a common set of services for exposure of AIML functionality, including federated and distributed learning (e.g. FL client registration management, FL client discovery and selection) and reference points. The AIMLE services are offered to the vertical application layer (VAL) and include:

- Support for application-layer ML model related aspects, including model retrieval, model training, model monitoring, model selection, model distribution, model update and model storage / discovery.

- Assistance in AI/ML task transfer and split AI/ML operations.

- Support HFL/VFL operations, including FL member registration, FL grouping and FL-related events notification, VFL feature alignment, HFL training, ML model training capability evaluation for FL (HFL/VFL).

- Support for AIMLE client registration, discovery, participation and selection.

- AIMLE client functional entity acts as the application client supporting AIMLE services.

- ML repository is a logical entity that serves both as a registry for AI/ML members or FL members and as a repository for application layer ML model related information. It can be accessed by the AIMLE server.

5.2.2.14.2.2 AI/ML related LCM activities

**Model Lifecycle enablement for AI/ML**

Some AIMLE capabilities are applicable to ML model lifecycle enablement which provides assistance for use cases where an ASP/VAL layer wants to find and use other application entities to perform some ML operations (e.g. ML model inference) and AIMLE server as a mediator to accomplish that.

An example including some capabilities related to lifecycle enablement is depicted in Annex C.4 of TS 23.482 [34]. The support capabilities are based on AIMLE capabilities identified in this specification. In particular, AIMLE is undertaking:

- ML model related support capabilities such as model retrieval, discovery and storage (as covered in procedures in clauses 8.2 and 8.11 of TS 23.482 [34] )

- ML operation related support capabilities such as VFL/ HFL and TL enablement, Split AI/ML Operation support, Data management assistance, AI/ML task transfer, FL assistance in member grouping, registration and event notification (as covered in procedures in clauses 8.4, 8.6, 8.12, 8.14, 8.15-8.18 of TS 23.482 [34]).

- AIMLE client related support capabilities, including AIMLE client registration, discovery, participation, monitoring, selection (as covered in procedures in clauses 8.7-8.10, 8.13 of TS 23.482 [34]).

**Data Collection/Storage/Exposure activities.**

Analysis of data collection activities as part of AIML\_App work.

Data Collection in TS 23.482 [34] refers to application data collection from the UE. EVEX mechanism can be reused for data collection as described in TS 26.531 [74].

ML model performance degradation can be detected in the AI/ML enablement by leveraging ADAES, e.g. based on information collected from analytics consumer. In AIML\_App, one possible use of AI/ML enablement is for supporting ML-enabled ADAES analytics services (as specified in TS 23.436 [33]). For Data Collection and Storage related to ADAES analytics:

- Application layer - Data Collection and Coordination Function (A-DCCF) coordinates the collection and distribution of data requested by the consumer (ADAE server). Data Collection Coordination is supported by a A-DCCF. ADAE server can send requests for data to the A-DCCF rather than directly to the Data Sources. A-DCCF may also perform data processing/abstraction and data preparation based on the VAL server requirements.

- Application layer - Analytics and Data Repository Function (A-ADRF) stores historical data and/or analytics, i.e. data and/or analytics related to past time period that has been obtained by the consumer (e.g. ADAE server). After the consumer obtains data and/or analytics, consumer may store historical data and/or analytics in an A-ADRF. Whether the consumer directly contacts the A-ADRF or goes via the A-DCCF is based on configuration.

**AI/ML-related information storage and discovery for AI/ML**

Analysis of ML model storage and exposure activities as part of AIML\_App work.

In AIML\_App, ML repository has been defined as:

1) a registry for AI/ML members or FL members (application layer entities participating in an AI/ML operation); and

2) as a repository for application layer ML model related information.

AIMLE server stores the ML model to the ML repository along with the ML model information (e.g. ML model ID). AIMLE server can also discover the ML models under certain filtering criteria (e.g. applicable to an ADAES analytics ID).

AIMLE server also registers and stores information on VAL servers, AIMLE servers or AIMLE clients which are expected to serve as AI/ML members or FL members in a model lifecycle operation (e.g. ML training, FL, TL). AIMLE clients or other VAL servers can discover the availability and capabilities of registered AI/ML members or FL members for a given ML model ID. Such discovery allows e.g. the VAL server identifying the candidate FL members to be considered for an FL process.

**Model training/delivery/ (de)-activation/inference emulation activities**

In AIML\_App, AIMLE server or the AIMLE client (at VAL UE side) can also be used for training an application layer ML model e.g. for given analytics service. Such ML model training can be used to support ADAES analytics services (as provided in TS 23.436 [33]). Based on the VAL request to provide ML-enabled analytics, ADAES may consume AIMLE services (e.g. for ML model training for a given analytics ID) to derive application layer data analytics.

The trained ML model can be delivered to VAL server or ADAES via the ML model training notification API.

SA WG6 has not defined any procedures for model (de)-activation and inference emulation.

**AI/ML model inference and delivery support for AI/ML**

Analysis of ML model inference activities as part of AIML\_App work.

SA WG6 has not defined dedicated procedures for supporting ML model inference; however, it provides assistance for registering and discovering AIMLE clients serving as ML model inference entities for a given analytics ID or model ID or split operation pipeline.

**Performance evaluation and accuracy monitoring activities**

Analysis of ML model performance evaluation and monitoring activities as part of AIML\_App work.

AIMLE server based on VAL request provides a capability for monitoring and detecting a degradation related to an ML operation / analytics operation and translating to an ML model performance degradation (expected or predicted) and performing a trigger action to alleviate this issue (new model training or re-training). Such trigger action may be either an adaptation of the AIMLE service, such as training of a new ML model for the AIMLE by the same or a different AIMLE client, or re-training of the ML model.

AIML\_App has provided the basic capability for performance monitoring activity, which is expected to be further worked in further release.

#### 5.2.2.15 Rel-19 CT WG4 WID - Protocol for AI Data Collection from UPF (FS\_PAIDC-UPF)

##### 5.2.2.15.1 Description

In Rel-18, the UPF offers services to the NEF, AF, SMF, NWDAF, DCCF, MFAF via the Nupf service based interface for data collecting in AI/ML related activities. In Rel-19, CT WG4 is studying "Protocol for AI Data Collection from UPF", which aims at studying UPF data Collection for AI/ML and whether alternative protocols, or enhancements to the existing SBI protocol, are needed to optimize the AI/ML data collection while ensuring secure, scalable and reliable data transfers across the core network identifies.

##### 5.2.2.15.2 Activities summary

Editor’s note: Reference to TR 21.919 can be added when the work item summary is made available.

#### 5.2.2.16 Rel-19 SA WG3 WID - Security aspects of Core Network Enhanced Support for AIML (AIML\_CN\_SEC)

##### 5.2.2.16.1 Description

The following objectives are expected to be specified as a result of this work item:

- Security aspects on enhancements to LCS to support AIML.

- Security aspects on VFL process.

##### 5.2.2.16.2 Activities summary

Editor’s note: Reference to TR 21.919 can be added when the work item summary is made available.

#### 5.2.2.17 Rel-19 SA WG5 WID - AI/ML management phase 2 (AIML\_MGT\_Ph2)

##### 5.2.2.17.1 Description

The objectives of AI/ML management phase 2 work item are to specify the management capabilities to support AI/ML functions defined by 3GPP, including:

- NG-RAN AIML-based Coverage and Capacity Optimization, and NG-RAN AIML-based Network Slicing defined by RAN3,

- Model delivery/transfer as defined by RAN1/2,

- ML model training and AI/ML inference functions for 5GC as defined by SA2, and

- MDA (Management Data Analytics) as defined by SA5.

To achieve these objectives, the following work tasks are defined:

**WT-1**: Specify the AI/ML management capabilities including use cases, requirements and solutions for the relevant AI/ML lifecycle operational steps based on TR 28.858 [19], including:

**WT-1.1**: Management capabilities for ML model training:

- ML-Knowledge-based Transfer Learning,

- ML pre-training and fine-tuning,

- ML model training for multiple contexts,

- ML training data statistics,

- ML model confidence,

- Management of Reinforcement Learning,

- ML model Distributed training,

- Management of Federated Learning,

- ML authentication.

**WT-1.2**: Management capabilities for AI/ML inference emulation:

- ML inference emulation,

- ML inference emulation environment selection.

**WT-1.3**: Management capabilities for ML model deployment:

- Enhancements to ML model loading,

- ML model transfer/delivery.

**WT-1.4**: Management capabilities for AI/ML inference:

- Coordination between the ML capabilities,

- ML remedial action management,

- Managing ML models in use in a live network,

- ML explainability.

##### 5.2.2.17.2 Activities summary

Editor’s note: Reference to TR 21.919 can be added when the work item summary is made available.

#### 5.2.2.18 Rel-19 SA WG6 WID - Application Data Analytics Enablement Service (TEI19\_ADAES)

##### 5.2.2.18.1 Description

The objectives of this work item include the following:

- Clarify metrics related to the analytics inputs/outputs introduced in TS 23.436 [33].

- Enhance the IEs for the request/response of data collection from Data Producer and/or A-ADRF (or via A-DCCF) in TS 23.436 [33].

- Complete the definition of IEs for the data/analytics storage subscription request to A-ADRF in TS 23.436 [33].

##### 5.2.2.18.2 Activities summary

Editor’s note: Reference to TR 21.919 can be added when the work item summary is made available.

#### 5.2.2.19 Rel-19 CT WG1/WG3 WID – CT aspects of application enablement for AI/ML services (AIML\_App)

##### 5.2.2.19.1 Description

The objective of this work item includes providing the stage 3 solutions and protocol support for application enablement for AIML services (AIML\_App) based upon the normative technical specification for the functionalities defined in stage 2 requirements under the AIML App WID in the SA WG6 working group.

Stage 3 work shall be started only after the applicable normative stage 2 work is available.

Stage 3 protocols and solutions will be specified in CT WG1 and CT WG3 respectively, to support the following stage 2 functionalities.

CT WG1, the expected work includes:

a) Definition of new APIs between the AIML server and AIML client provided by AIML layer to support the application enablement for AIML services based on normative stage-2 work developed in 3GPP SA WG6, to support:

1) ML client configuration provisioning;

2) AIMLE client selection;

3) AIMLE client registration;

4) AIML service lifecycle management;

5) AIML operational splitting and provisioning management;

6) vertical federated learning (VFL) and horizontal federated learning (HFL);

7) AIML data management;

8) AIML edge services; and

9) AIML model distribution.

b) Enhancement the APIs between the ADAE server and ADAE client provided by ADAE layer to support the application enablement for AIML services based on normative stage-2 work developed in 3GPP SA WG6, to support:

1) ML-enabled ADAE analytics; and

2) Application layer AIML Member Capability Analytics.

CT WG3, the expected work includes:

a) Definition of new APIs in network provided by AIML layer to support the application enablement for AIML services based on normative stage-2 work developed in 3GPP SA WG6, to support:

1) AIMLE server and client registration;

2) AIMLE client information retrieval;

3) AIML model lifecycle management;

4) AIML model distribution;

5) AIMLE client discovery; and

6) AIMLE client selection and reselection;

7) federated learning (FL) member registration and grouping;

8) vertical federated learning (VFL) and horizontal federated learning (HFL);

9) AIML operational splitting and provisioning management;

10) AIML policy provisioning and management;

11) AIML service lifecycle management;

12) AIML services for edge computing; and

13) AIML information transfer.

b) Enhancement of the APIs provided by ADAE layer to support the application enablement for AIML services based on normative stage-2 work developed in 3GPP SA WG6, to support:

1) ML-enabled ADAE analytics;

2) AI-enabled DN Energy Analytics; and

3) Application layer AIML Member Capability Analytics.

##### 5.2.2.19.2 Activities summary

Editor’s note: Reference to TR 21.919 can be added when the work item summary is made available.

#### 5.2.2.20 Rel-19 CT WG3 WID – Rel-19 Enhancements of Network Automation Enablers (eNetAE19)

##### 5.2.2.20.1 Description

The objective of this work item is to specify the technical improvements and enhancements to the network data analytics related services in Release 19 stage 3 level, mainly (but not exhaustively) including:

1 Potential completion of the support of Processing Instructions in MFAF and ADRF APIs.

2 Completion of the analytics transfer procedure.

3 Completion of the analytics aggregation.

4 Clarifications for input data information in ML model training procedure.

5 Enhancements of Collective Behaviour of UEs in AF, e.g., to include the average moving speed of the UE.

6 Enhancements of Nnwdaf\_MLModelTraining\_Notify service operation to provide the Global ML Model Accuracy information.

7 Other technical enhancements and corrections for the services related to Network Automation Enablers (i.e., the pure stage 3 protocol and interface enhancements are not included), which are not covered by the other dedicated WIs.

##### 5.2.2.20.2 Activities summary

Editor’s note: Reference to TR 21.919 can be added when the work item summary is made available.

#### 5.2.2.21 Rel-19 CT WG3/WG4 WID – CT aspects of Core Network Enhanced Support for Artificial Intelligence (AI) and Machine Learning (ML) (AIML\_CN)

##### 5.2.2.21.1 Description

The objective of this work is to specify the CT aspects of Core Network Enhanced Support for Artificial Intelligence (AI)/Machine Learning (ML) in order to implement the stage 2 normative work. The stage 3 work shall be started after the applicable normative stage 2 requirements are available, the detail impacts are subject to change when stage 2 normative CRs are agreed.

The following aspects of work are expected to be covered:

CT WG3:

- For Direct AI/ML based Positioning:

- update on NWDAF to support Direct AI/ML based Positioning, e.g., AI positioning model training, AI positioning model delivery, performance monitoring;

- update on NWDAF to support data collection for AI positioning model training;

- For Vertical Federated Learning (VFL):

- update on NWDAF and AF to support sample alignment for VFL;

- update on NWDAF and AF to support VFL training;

- update on NWDAF and AF to support VFL inference;

- update on NWDAF and AF to support performance monitoring for VFL;

- update on NEF to support VFL in case of an untrusted AF;

- For NWDAF-assisted policy control and QoS enhancement:

- update on PCF and NWDAF to support the analytics for QoS and policy assistance information;

- update on NFs and AF to support input data collection;

- update on NWDAF to support performance monitoring for NWDAF-assisted policy control;

- For signalling storm mitigation and prevention:

- update on NWDAF to support analytics for signalling storm Mitigation and Prevention caused by NFs;

- update on NWDAF to support analytics for Signalling storm Mitigation and Prevention caused by massive signalling of UEs;

- update on NFs to support input data collection;

- Potential impacts on ADRF, DCCF, and MFAF APIs for supporting the data sources or analytics.

CT WG4:

- For Direct AI/ML based Positioning:

- update on LMF to support Direct AI/ML based positioning;

- update on UDM to allow LMF as a consumer for retrieving user consent;

- update on NFs to support training data collection for AI positioning model;

- update to NWDAF discovery via NRF for training ML model for Direct AI/ML based Positioning;

- For Vertical Federated Learning:

- update on NRF to support VFL entity registration and discovery;

- For signalling storm mitigation and prevention:

- update on AMF and NRF to support input data collection;

- update to SCP entity to support input data collection.

##### 5.2.2.21.2 Activities summary

Editor’s note: Reference to TR 21.919 can be added when the work item summary is made available.

## 5.3 AI/ML related activities in TSG RAN Working Groups

### 5.3.1 AI/ML related terminology

#### 5.3.1.1 TSG RAN WG1

The following definitions are provided in clause 3 of TR 38.843 [3]:

- AI/ML-enabled Feature: refers to a Feature where AI/ML may be used.

- AI/ML Model: A data driven algorithm that applies AI/ML techniques to generate a set of outputs based on a set of inputs.

- AI/ML model delivery: A generic term referring to delivery of an AI/ML model from one entity to another entity in any manner.

NOTE 1: An entity could mean a network node/function (e.g. gNB, LMF, etc.), UE, proprietary server, etc.

- AI/ML model ID: A logical AI/ML model is identified by a Model ID. The Model ID, if needed, can be used in a Functionality (defined in functionality-based LCM) for LCM operations.

- AI/ML model Inference: A process of using a trained AI/ML model to produce a set of outputs based on a set of inputs.

- AI/ML model testing: A subprocess of training, to evaluate the performance of a final AI/ML model using a dataset different from one used for model training and validation. Differently from AI/ML model validation, testing does not assume subsequent tuning of the model.

- AI/ML model training: A process to train an AI/ML Model [by learning the input/output relationship] in a data driven manner and obtain the trained AI/ML Model for inference.

- AI/ML model transfer: Delivery of an AI/ML model over the air interface in a manner that is not transparent to 3GPP signalling, either parameters of a model structure known at the receiving end or a new model with parameters. Delivery may contain a full model or a partial model.

- AI/ML model validation: A subprocess of training, to evaluate the quality of an AI/ML model using a dataset different from one used for model training, that helps selecting model parameters that generalize beyond the dataset used for model training.

- Data collection: A process of collecting data by the network nodes, management entity, or UE for the purpose of AI/ML model training, data analytics and inference.

- Federated learning / federated training: A machine learning technique that trains an AI/ML model across multiple decentralized edge nodes (e.g. UEs, gNBs) each performing local model training using local data samples. The technique requires multiple interactions of the model, but no exchange of local data samples.

- Functionality identification: A process/method of identifying an AI/ML functionality for the common understanding between the NW and the UE.

NOTE 2: Information regarding the AI/ML functionality may be shared during functionality identification. Where AI/ML functionality resides depends on the specific use cases and sub use cases.

- Management instruction: Information needed to ensure proper inference operation. This information may include selection/(de)activation/switching of AI/ML models or AI/ML functionalities, fallback to non-AI/ML operation, etc.

- Model activation: enable an AI/ML model for a specific AI/ML-enabled feature.

- Model deactivation: disable an AI/ML model for a specific AI/ML-enabled feature.

- Model download: Model transfer from the network to UE.

- Model identification: A process/method of identifying an AI/ML model for the common understanding between the NW and the UE.

NOTE 3: The process/method of model identification may or may not be applicable.

NOTE 4: Information regarding the AI/ML model may be shared during model identification.

- Model monitoring: A procedure that monitors the inference performance of the AI/ML model.

- Model parameter update: Process of updating the model parameters of a model.

- Model selection: The process of selecting an AI/ML model for activation among multiple models for the same AI/ML enabled feature.

NOTE 5: Model selection may or may not be carried out simultaneously with model activation.

- Model switching: Deactivating a currently active AI/ML model and activating a different AI/ML model for a specific AI/ML-enabled feature.

- Model update: Process of updating the model parameters and/or model structure of a model.

- Model upload: Model transfer from UE to the network.

- Network-side (AI/ML) model: An AI/ML Model whose inference is performed entirely at the network.

- Offline field data: The data collected from field and used for offline training of the AI/ML model.

- Offline training: An AI/ML training process where the model is trained based on collected dataset and where the trained model is later used or delivered for inference.

NOTE 6: This definition only serves as a guidance. There may be cases that may not exactly conform to this definition but could still be categorized as offline training by commonly accepted conventions.

- Online field data: The data collected from field and used for online training of the AI/ML model.

- Online training: An AI/ML training process where the model being used for inference) is (typically continuously) trained in (near) real-time with the arrival of new training samples.

- Reinforcement Learning (RL): A process of training an AI/ML model from input (a.k.a. state) and a feedback signal (a.k.a. reward) resulting from the model's output (a.k.a. action) in an environment the model is interacting with.

- Semi-supervised learning: A process of training a model with a mix of labelled data and unlabelled data.

- Supervised learning: A process of training a model from input and its corresponding labels.

- Test encoder/decoder for TE: AI/ML model for UE encoder/gNB decoder implemented by TE.

- Two-sided (AI/ML) model: A paired AI/ML Model(s) over which joint inference is performed, where joint inference comprises AI/ML Inference whose inference is performed jointly across the UE and the network, i.e. the first part of inference is firstly performed by UE and then the remaining part is performed by gNB, or vice versa.

- UE-side (AI/ML) model: An AI/ML Model whose inference is performed entirely at the UE.

- Unsupervised learning: A process of training a model without labelled data.

- Proprietary-format models: ML models of vendor-/device-specific proprietary format, from 3GPP perspective. They are not mutually recognizable across vendors and hide model design information from other vendors when shared.

NOTE 7: An example is a device-specific binary executable format.

- Open-format models: ML models of specified format that are mutually recognizable across vendors and allow interoperability, from the 3GPP perspective. They are mutually recognizable between vendors and do not hide model design information from other vendors when shared.

#### 5.3.1.2 TSG RAN WG3

The following definitions are provided in clause 16.20 of TS 38.300 [11]:

- AI/ML Model Training follows the definition of the "ML model training" as specified in clause 3.1 of TS 28.105 [9].

- AI/ML Model Inference follows the definition of the "AI/ML inference" as defined in clause 3.1 of TS 28.105 [9].

### 5.3.2 AI/ML related activities

#### 5.3.2.1 Rel-19 RAN WG1/RAN WG4 WID - Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface (NR\_AIML\_air)

##### 5.3.2.1.1 Description

The objective of this work is to provide specification support for the following aspects:

- AI/ML general framework for one-sided AI/ML models within the realm of what has been studied in the FS\_NR\_AIML\_air project (RAN WG2):

- Signalling and protocol aspects of Life Cycle Management (LCM) enabling functionality and model (if justified) selection, activation, deactivation, switching, fallback:

- Identification related signalling is part of the above objective.

- Necessary signalling/mechanism(s) for LCM to facilitate model training, inference, performance monitoring, data collection (except for the purpose of CN/OAM/OTT collection of UE-sided model training data) for both UE-sided and NW-sided models.

- Signalling mechanism of applicable functionalities/models.

- Beam management - DL Tx beam prediction for both UE-sided model and NW-sided model, encompassing (RAN WG1/RAN WG2):

- Spatial-domain DL Tx beam prediction for Set A of beams based on measurement results of Set B of beams ("BM-Case1").

- Temporal DL Tx beam prediction for Set A of beams based on the historic measurement results of Set B of beams ("BM-Case2").

- Specify necessary signalling/mechanism(s) to facilitate LCM operations specific to the Beam Management use cases, if any.

- Enabling method(s) to ensure consistency between training and inference regarding NW-side additional conditions (if identified) for inference at UE.

- Positioning accuracy enhancements, encompassing (RAN WG1/RAN WG2/RAN WG3):

- Direct AI/ML positioning:

- Case 1: UE-based positioning with UE-side model, direct AI/ML positioning.

- Case 3b: NG-RAN node assisted positioning with LMF-side model, direct AI/ML positioning.

- AI/ML assisted positioning:

- Case 3a: NG-RAN node assisted positioning with gNB-side model, AI/ML assisted positioning.

- Specify necessary measurements, signalling/mechanism(s) to facilitate LCM operations specific to the Positioning accuracy enhancements use cases, if any.

- Investigate and specify the necessary signalling of necessary measurement enhancements (if any).

- Enabling method(s) to ensure consistency between training and inference regarding NW-side additional conditions (if identified) for inference at UE for relevant positioning sub use cases.

* CSI feedback enhancement, encompassing [RAN1/RAN2]:

- CSI prediction (UE-sided model):

- Functionality-based LCM leveraged from other use cases, when necessary and applicable,

- Study, and if necessary, specify consistency of training/inference,

- Core requirements for the above three use cases for AI/ML LCM procedures and UE features (RAN WG4):

- Specify necessary RAN WG4 core requirements for the above three use cases.

- Specify necessary RAN WG4 core requirements for LCM procedures including performance monitoring.

- For Beam Management and Positioning Accuracy enhancement use cases, specify performance requirements and test cases for AI/ML LCM procedures (including performance monitoring) and UE features enabled by UE-sided models:

- Specify necessary performance requirements and tests (including metrics) for the above-mentioned use cases.

- Specify necessary test cases and performance requirements for LCM procedure, including performance monitoring.

Note: the following aspects may be considered:

- Relation to legacy requirements

- Performance monitoring and LCM aspects considering use-case specifics

- Generalization aspects

- Static/non-static scenarios/conditions and propagation conditions for testing (e.g. CDL, field data, etc.)

- UE processing capability and limitations

- Post-deployment validation due to model change/drift

- RAN5 aspects related to testability and interoperability to be addressed on a request basis

##### 5.3.2.1.2 Activities summary

Editor's note: Reference to TR 21.919 can be added when the work item summary is made available.

#### 5.3.2.2 Rel-19 RAN WG2 SID - AIML for mobility in NR (FS\_NR\_AIML\_Mob)

##### 5.3.2.2.1 Description

The study will focus on mobility enhancement in RRC\_CONNECTED mode over air interface by following existing mobility framework, i.e. handover decision is always made in network side. Mobility use cases focus on standalone NR PCell change. UE-side and network-side AI/ML model can be both considered, respectively. The investigation is to evaluate potential benefits and gains of AI/ML aided mobility for network triggered L3-based handover, considering the following aspects:

- AI/ML based RRM measurement and event prediction:

- Cell-level measurement prediction including intra and inter-frequency (UE sided and NW sided model) (RAN WG2):

- Inter-cell Beam-level measurement prediction for L3 Mobility (UE sided and NW sided model) (RAN WG2).

- HO failure/RLF prediction (UE sided model) (RAN WG2).

- Measurement events prediction (UE sided model) (RAN WG2).

- Study the need/benefits of any other UE assistance information for the network side model (RAN WG2).

- The evaluation of the AI/ML aided mobility benefits should consider HO performance KPIs (e.g. Ping-pong HO, HOF/RLF, Time of stay, Handover interruption, prediction accuracy and measurement reduction) etc.) and complexity trade-offs (RAN WG2).

- Potential AI mobility specific enhancement should be based on the Rel19 AI/ML-air interface WID general framework (e.g. LCM, performance monitoring etc) (RAN WG2).

- Potential specification impacts of AI/ML aided mobility (RAN WG2).

- Evaluate testability, interoperability and impacts on RRM requirements and performance (RAN WG4).

NOTE: There was no normative work performed in Rel-19.

#### 5.3.2.3 Rel-18 RAN WG3 WID - Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN (NR\_AIML\_NGRAN-Core)

##### 5.3.2.3.1 Description

The objective of this work is to specify data collection enhancements and signalling support within existing NG-RAN interfaces and architecture (including non-split architecture and split architecture) for AI/ML-based Network Energy Saving, Load Balancing and Mobility Optimization.

Support of AI/ML for NG-RAN, as a RAN internal function, is used to facilitate Artificial Intelligence (AI) and Machine Learning (ML) techniques in NG-RAN. The objective of AI/ML for NG-RAN is to improve network performance and user experience, through analysing the data collected and autonomously processed by the NG-RAN, which can yield further insights, e.g. for Network Energy Saving, Load Balancing, Mobility Optimization.

Support of AI/ML in NG-RAN requires inputs from neighbour NG-RAN nodes (e.g. predicted information, feedback information, measurements) and/or UEs (e.g. measurement results).

Signalling procedures used for the exchange of information to support AI/ML in NG-RAN are use case and data type agnostic, which means that the intended usage of the data exchanged via these procedures (e.g. input, output, feedback) is not indicated. The collection and reporting of information are configured through the Data Collection Reporting Initiation procedure, while the actual reporting is performed through the Data Collection Reporting procedure.

Support of AI/ML in NG-RAN does not apply to ng-eNB.

For the deployment of AI/ML in NG-RAN, the following scenarios may be supported:

- AI/ML Model Training is located in the OAM and AI/ML Model Inference is located in the NG-RAN node.

- AI/ML Model Training and AI/ML Model Inference are both located in the NG-RAN node.

##### 5.3.2.3.2 Activities summary

Summary is available in clause 11.2 of TR 21.918 [2].

#### 5.3.2.4 Rel-19 RAN WG3 SID - Enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN (FS\_NR\_AIML\_NGRAN\_enh)

##### 5.3.2.4.1 Description

The objective of this study is to further investigate new AI/ML based use cases and identify enhancements to support AI/ML functionality and further discussions on the Rel-18 leftovers. The detailed objectives of the study are listed as follows:

- Study two new AI/ML based use cases, i.e. Network Slicing and CCO, with existing NG-RAN interfaces and architecture (including non-split architecture and split architecture).

- Rel-18 leftovers as candidates for normative work, based on the Rel-18 principles, as follows:

- Mobility optimization for NR-DC.

- Split architecture support for Rel-18 use cases based on the conclusions from Rel-18 WI.

- Energy Saving enhancements, e.g. Energy Cost Prediction.

- Continuous MDT collection targeting the same UE across RRC states.

- Multi-hop UE trajectory across gNBs.

NOTE: The outcome of this study was used to support the NR\_AIML\_NGRAN\_enh-Core Rel-19 RAN WG3 WID, see clause 5.3.2.6.

#### 5.3.2.5 Rel-19 RAN WG1/RAN WG4 SID - Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface (FS\_NR\_AIML\_air\_Ph2)

##### 5.3.2.5.1 Description

The objective of this work is to provide further study on some outstanding issues identified during the prior study. specifically for the following aspects:

- CSI feedback enhancement [RAN1]:

- For CSI temporal prediction: further update TR 38.843 [3] with additional evaluations.

- For CSI compression (two-sided model), further study ways to:

- Improve trade-off between performance and complexity/overhead:

- e.g. considering extending the spatial/frequency compression to spatial/temporal/frequency compression, cell/site specific models, CSI compression plus prediction (compared to Rel-18 non-AI/ML based approach), etc.

- Alleviate/resolve issues related to inter-vendor training collaboration,

while addressing necessary specification impact analysis, as well as, other aspects requiring further study/conclusion as captured in the conclusions clause of TR 38.843 [3]:

- Necessity and details of model Identification concept and procedure in the context of LCM for two-sided models [RAN2/RAN1].

- CN/OAM/OTT collection of UE-sided model training data [RAN2/RAN1]:

- For the FS\_NR\_AIML\_Air study use cases, identify the corresponding contents of UE data collection.

- Analyse the UE data collection mechanisms identified during the FS\_NR\_AIML\_Air (clause 7.2.1.3.2 of TR 38.843 [3]) study along with the implications and limitations of each of the methods.

- Model transfer/delivery [RAN2/RAN1]:

- Determine whether there is a need to consider standardised solutions for transferring/delivering AI/ML model(s) considering at least the solutions identified during the FS\_NR\_AIML\_Air study.

- Testability and interoperability [RAN4]:

o Further analyse the various testing options for two-sided models, in collaboration with RAN1, and including at least:

- Relation to legacy requirements.

- Performance monitoring and LCM aspects considering use-case specifics.

- Generalization aspects.

- Static/non-static scenarios/conditions and propagation conditions for testing (e.g. CDL, field data, etc.).

- UE processing capability and limitations.

- Post-deployment validation due to model change/drift.

- RAN5 aspects related to testability and interoperability to be addressed on a request basis.

#### 5.3.2.6 Rel-19 RAN WG3 WID - Enhancements for Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN (FS\_NR\_AIML\_NGRAN\_enh-Core)

##### 5.3.2.6.1 Description

The aim of this work item is to specify new AI/ML-based use cases and introduce further enhancements to finalize the Rel-18 leftovers based on the conclusions captured in TR 38.743 [69] (FS\_NR\_AIML\_NGRAN\_enh).

The objective of this work is to provide specification support for the following aspects:

- Specify data collection enhancements and signalling support within existing NG-RAN interfaces and architecture (including non-split architecture and split architecture) for AI/ML-based Slicing and AI/ML based CCO. [RAN3]

- Support of the Leftovers in Rel-18 AI/ML for NG-RAN [RAN3]:

- Mobility Optimization for NR-DC

- Split architecture support for Rel-18 use cases

- Continuous MDT collection targeting the same UE across RRC states

NOTE: Coordination with RAN WG2, SA WG5 when needed if any.

##### 5.3.2.6.2 Activities summary

Editor's note: Reference to TR 21.919 can be added when the work item summary is made available.

# 6 Analysis on AI/ML across 3GPP

## 6.1 General

This clause will identify any potential misalignments and inconsistencies for AI/ML across 3GPP, based on clause 5.

NOTE: Any RAN related aspects are subject to early coordination and feedback from TSG RAN.

## 6.2 AI/ML related terminology

### 6.2.1 Analysis on AI/ML model related terminology consistency

This clause identifies any potential misalignments and inconsistencies for AI/ML terminology across 3GPP, based on clause 5.

#### 6.2.1.1 Analysis on ML model

The term 'ML model' has been defined differently by SA WG5, SA WG6 and RAN WG1, as illustrated in Table 6.2.1.1-1.

Table 6.2.1.1-1: Definition of ML model as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model |
| SA WG5 TS 28.105 [9] | A manageable representation of an ML model algorithm.  (NOTE 1, NOTE 2, NOTE 3) |
| SA WG6 TS 23.482 [34] | According to TS 28.105 [9], mathematical algorithm that can be "trained" by data and human expert input as examples to replicate a decision an expert would make when provided that same information. |
| RAN WG1 TR 38.843 [3] | A data driven algorithm that applies AI/ML techniques to generate a set of outputs based on a set of inputs. |
| NOTE 1: An ML model algorithm is a mathematical algorithm through which running a set of input data can generate a set of inference output.  NOTE 2: An ML model algorithm is proprietary and not in scope for standardization and therefore not treated in this specification.  NOTE 3: An ML model may include metadata. Metadata may include e.g. information related to the trained model and applicable runtime context. | |

The following unified definition for 'ML model' is proposed:

**ML model:** A mathematical algorithm that applies ML techniques to generate a set of outputs based on a set of inputs. It may include metadata which consists of, e.g. information related to the model and applicable runtime context.

NOTE:An ML model can be managed, stored and transferred as artifacts, which may be containers, images, or proprietary file formats.

#### 6.2.1.2 Analysis on ML model training

The term 'ML model training' has been defined differently by SA WG5, SA WG6 and RAN WG1, as illustrated in Table 6.2.1.2-1. RAN WG3 follows the definition of SA WG5.

Table 6.2.1.2-1: Definition of ML model training as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model training |
| SA WG5 TS 28.105 [9] | A process performed by an ML training function to take training data, run it through an ML model algorithm, derive the associated loss and adjust the parameterization of that ML model iteratively based on the computed loss and generate the trained ML model. |
| SA WG6 TS 23.482 [34] | According to TS 28.105 [9], ML model training includes capabilities of an ML training function or service to take data, run it through an ML model, derive the associated loss and adjust the parameterization of that ML model based on the computed loss. |
| RAN WG1 TR 38.843 [3] | A process to train an AI/ML Model [by learning the input/output relationship] in a data driven manner and obtain the trained AI/ML Model for inference. |
| RAN WG3 TS 38.300 [11] | AI/ML Model Training follows the definition of the "ML model training" as specified in clause 3.1 of TS 28.105 [9]. |
| RAN WG3 TS 38.401 | AI/ML Model Training follows the definition of the "ML model training" as specified in clause 3.1 of TS 28.105 [9]. |

The following unified definition for 'ML model training' is proposed:

**ML model training:** A process to train an ML Model by learning the input/output relationship in a data driven manner and obtain the trained ML Model for e.g. inference.

#### 6.2.1.3 Analysis on ML model re-training

The term 'ML model re-training' has been defined differently by SA WG5 and RAN WG1, as illustrated in Table 6.2.1.3-1. RAN WG1 introduces two new terms, i.e. ML model parameter update and ML model update, which is nothing but ML model re-training.

Table 6.2.1.3-1: Definition of ML model re-training / ML model parameter update as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model re-training / ML model parameter update / ML model update |
| SA WG5 TS 28.105 [9] | *ML model re-training:* A process of training a previous version of an ML model and generate a new version. |
| RAN WG1 TR 38.843 [3] | *ML model parameter update:* A process of updating the model parameters of a model.  *Model update:* A process of updating the model parameters and/or model structure of a model |
| SA WG6 TS 23.482 [34] | *ML model update:* A process of training a new version of a ML model and updating its parameters. |

The term ML model re-training is proposed as unified term rather than using different terms such as 'ML model parameter update' or 'ML model update' which align conceptually but vary in technical detail and scope depending on the WG definition.

The following unified definition for 'ML model re-training' and 'ML model update' is proposed:

**ML model re-training:** A process of training a previous version of an ML model and generate a new version.

**ML model update:** A process that improves the performance or behaviour of an ML model through actions such as re-training, parameter adjustment, structural modification, or deployment of a new version.

In addition, it should be clarified that the term "model update" as used by RAN WG1 [3] may include changes to both model parameters and model structure. In contrast, TS 28.105 [9] defines ML model re-training as a process that does not alter the model structure, focusing instead on updating the model’s parameters, for example, when model performance degrades or when a new set of training data becomes available. TS 28.105 also describes ML model update as a broader concept that may involve various internal actions to improve the inference capabilities of an AI/ML inference function. These actions may include re-training, download new configurations, or trigger related processes. The term is used in contexts where the consumer requests improved capabilities, but the exact internal update mechanism may not be exposed. This illustrates that ML model update in TS 28.105 encompasses re-training but may also refer to other means of updating model behaviour or performance.

#### 6.2.1.4 Analysis on ML model testing

The term 'ML model testing' has been defined differently by SA WG5 and RAN WG1, as illustrated in Table 6.2.1.4-1.

Table 6.2.1.4-1: Definition of ML model testing as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model testing |
| SA WG5 TS 28.105 [9] | A process of evaluating the performance of an ML model using testing data different from data used for model training and validation. |
| RAN WG1 TR 38.843 [3] | A subprocess of training, to evaluate the performance of a final AI/ML model using a dataset different from one used for model training and validation. Differently from AI/ML model validation, testing does not assume subsequent tuning of the model. |

The following definition for 'ML model testing' in TS 28.105 [9] is proposed as a unified definition:

**ML model testing:** A process of evaluating the performance of an ML model using test data different from data used for model training and validation.

**Note:** Regarding the definition adopted in TR 38.843 [3], it should be clarified that RAN WG1 considers ML model testing as a subprocess of training, and notes that testing does not assume subsequent tuning of the model. In contrast, TS 28.105 [9] defines ML model testing as a distinct step from training, focused solely on evaluating the performance of an ML model using test data that is different from the data used for training and validation. Furthermore, according to TS 28.105 [9], if the performance of the ML model does not meet the target set during testing, the model may be retrained. This highlights that while testing itself does not include tuning, it can lead to retraining actions based on performance outcomes. Therefore, the definition in TS 28.105 reflects a clearer separation between training and testing phases, with explicit linkage to lifecycle management decisions such as retraining.

#### 6.2.1.5 Analysis on ML model inference

The term 'ML model inference' has been defined differently by SA WG5, SA WG6 and RAN WG1, as illustrated in Table 6.2.1.5-1. RAN WG3 follows the definition of SA WG5.

Table 6.2.1.5-1: Definition of ML model inference as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model inference |
| SA WG5 TS 28.105 [9] | A process of running a set of input data through a trained ML model to produce set of output data, such as predictions. |
| SA WG6 TS 23.482 [34] | According to TS 28.105 [9], ML model training includes capabilities of an ML model inference function that employs an ML model and/or AI decision entity to conduct inference. |
| RAN WG1 TR 38.843 [3] | A process of using a trained AI/ML model to produce a set of outputs based on a set of inputs. |
| RAN WG3 TS 38.300 [11] | AI/ML Model Inference follows the definition of the "AI/ML inference" as defined in clause 3.1 of TS 28.105 [9]. |
| RAN WG3 TS 38.401 | AI/ML Model Inference follows the definition of the "AI/ML inference" as defined in clause 3.1 of TS 28.105 [9]. |

The following unified definition for 'ML model inference' is proposed:

**ML model inference:** A process of running a set of inputs through a trained ML model to produce a set of outputs.

#### 6.2.1.6 Analysis on ML model activation & ML model de-activation

The term 'ML model activation' and 'ML model deactivation' have been defined by RAN WG1, as illustrated in Table 6.2.1.6-1. SA WG5 mentions the terms ML activation and ML deactivation several times in TS 28.105 [9] but does not provide a definition. It defines instead AI/ML activation/deactivation of the scope of an inference function.

Table 6.2.1.6-1: Definition of ML model activation & ML model de-activation as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model activation & ML model de-activation |
| SA WG5 TS 28.105 [9] | ***AI/ML activation****: a process of enabling the inference capability of an AI/ML inference function.*  ***AI/ML deactivation****: a process of disabling the inference capability of an AI/ML inference function.* |
| RAN WG1 TR 38.843 [3] | *ML Model activation:* enable an AI/ML model for a specific AI/ML-enabled feature.  *ML Model deactivation:* disable an AI/ML model for a specific AI/ML-enabled feature. |

The following unified definition for 'ML model activation' and 'ML model deactivation' is proposed:

**ML model activation:** A process to enable an ML model for a specific AI/ML-enabled feature.

**ML model deactivation:** A process to disable an ML model for a specific AI/ML-enabled feature.

Editor's note: Further analysis is required to determine whether ML model activation and deactivation should be associated specifically with inference capability (according to TS 28.105 [9]) or with the ML model more broadly (according to TR 38.843 [3]).

#### 6.2.1.7 Analysis on ML model lifecycle

The term 'ML model lifecycle' has been defined by SA WG6, as illustrated in Table 6.2.1.7-1. However, SA WG2 TS 23.288 [8], SA WG2 TR 23.700-84 [4], SA WG4 TR 26.927 [12], SA WG5 TS 28.105 [9], SA WG6 TR 23.700-82 [7], RAN WG1 TR 38.843 [3] and RAN WG3 also mentions one or more phases of ML model life cycle without providing a clear definition of ML model lifecycle.

Table 6.2.1.7-1: Definition of ML model lifecycle as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model lifecycle |
| SA WG6 TS 23.482 [34] | The lifecycle of an ML model aka ML model operational workflow consists of a sequence of ML operations for a given ML task / job (such job can be an analytics task or a VAL automation task). This definition is aligned with the 3GPP definition on ML model lifecycle according to TS 28.105 [9]. |
| SA WG5 TS 28.105 [9] | **ML model training:** includes initial training and re-training, as well as validation of the ML model using training and validation data. If the validation results do not meet expectations (e.g. unacceptable variance), re-training is required.  **ML model testing:** evaluates the performance of a trained ML model using testing data. If the results do not meet expectations, re-training is required before proceeding.  **AI/ML inference emulation (optional):** allows testing the inference performance of an ML model in an emulation environment before deploying it to the target network or system. If the emulation performance does not meet the target requirements, the model may require further re-training.  **ML model deployment:** involves the process of loading a trained ML model to make it available for use at the target AI/ML inference function. Deployment may not be needed if the training and inference functions are co-located.  **AI/ML inference:** performing inference using a trained ML model at the AI/ML inference function. The inference process may trigger model re-training or updates based on performance monitoring and evaluation. |

The following unified definition for 'ML model lifecycle' is proposed:

**ML model lifecycle:** The end-to-end process typically consisting of data processing, model training, model testing, model deployment, model inference, model monitoring and model maintenance.

NOTE 1: Data processing includes collecting and preparing the data for model training and model inference.

NOTE 2: Model training includes training and validating the model before model deployment.

NOTE 3: Model testing includes testing the model before model deployment.

NOTE 4: Model deployment includes making a trained ML model available for use in the target environment.

NOTE 5: Model monitoring includes observing the performance of the model during the model maintenance process.

NOTE 6: Model maintenance includes updating the model, retraining the model and (de-)activating the model.

#### 6.2.1.8 Analysis on ML model lifecycle management

SA WG5 describes the ML model lifecycle in clause 4a.0 of TS 28.105 [9] and ML model lifecycle management capabilities for ML model training, ML model testing, ML inference emulation as optional, ML model deployment and AI/ML inference in clause 6.1 of TS 28.105 [9]. The terms 'ML model-based lifecycle management', 'ML-enabled functionality' and 'Functionality-based lifecycle management' have been defined by RAN1, as illustrated in Table 6.2.1.8-1.

Table 6.2.1.8-1: Definitions of ML model-based lifecycle management, ML-enabled functionality and Functionality-based lifecycle management as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | ML model lifecycle management / Functionality-based lifecycle management |
| RAN WG1 TR 38.843 [3] | **ML model-based lifecycle management:** Operates based on identified logical models, where a model may be associated with specific configurations/conditions associated with UE capability of an AI/ML-enabled Feature / Feature Group and additional conditions (e.g. scenarios, sites and datasets) as determined/identified between UE-side and NW-side. The models are identified at the Network and Network/UE may activate/deactivate/select/switch individual AI/ML models via model ID.  **(ML-enabled) Functionality:** An AI/ML-enabled Feature/Feature Group enabled by configuration(s), where configuration(s) is(are) supported based on conditions indicated by UE capability.  **Functionality-based lifecycle management:** Signaling procedure where network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signaling (e.g. RRC, MAC-CE, DCI); operates based on, at least, one configuration of AI/ML-enabled Feature/FG or specific configurations of an AI/ML-enabled Feature / Feature Group. |
| SA WG5 TS 28.105 [9] | **ML model training management:** enables requesting, consuming and controlling ML model training and re-training processes. It includes training performance management and policy setting for producer-initiated training.  **ML model testing management:** allows requesting and receiving ML model testing results, selecting performance metrics and triggering model re-training based on test performance.  **ML model loading management:** supports triggering, controlling and monitoring the ML model loading process as part of model deployment.  **AI/ML inference management:** allows managing inference functions and/or ML model(s), including activation/deactivation, output parameter configuration, performance monitoring and triggering model updates if necessary. |

The following unified definition for 'ML model lifecycle management' is proposed:

**ML model lifecycle management:** The management capabilities allowing a producer or consumer to manage different phases of the ML model lifecycle as defined in clause 6.2.1.7.

The following definition for 'Functionality-based lifecycle management' is proposed for adoption by all 3GPP RAN Working Groups:

**Functionality-based lifecycle management:** Signalling procedure where network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signalling (e.g. RRC, MAC-CE, DCI); operates based on, at least, one configuration of AI/ML-enabled Feature / Feature Group or specific configurations of an AI/ML-enabled Feature/FG.

NOTE 1: In the context of RAN1, RAN2 and RAN4, functionality-based lifecycle management does not consider training, testing and maintenance phases and consider them as implementation-specific.

NOTE 2: Applicability of Functionality-based lifecycle management definition to/in TSG SA WGs is optional.

Editor's note: The following analyses on the key differences between ML Model LC and LCM is to be revised and possibly relocated to different clause in this TR.

**Key differences between ML Model lifecycle (LC) and ML Model lifecycle management (LCM)**

TS 28.105 [9] defines both ML model lifecycle (LC) and ML model lifecycle management (LCM) within the scope of AI/ML management in 3GPP networks. The key differences between the two are:

**ML model lifecycle (LC)** describes the essential steps (phases) an ML model undergoes, from training to inference. It consists of:

- **ML model training** (e.g. initial training & re-training).

- **ML model testing.**

- **ML emulation.**

- **ML model deployment** (including ML model loading).

- **AI/ML inference.**

**ML model lifecycle management (LCM)** focuses on the management capabilities that control and optimize each phase of the ML model lifecycle. LCM enables functionalities such as:

- **Training management** (e.g. triggering re-training, setting policies).

- **Testing management** (e.g. evaluating performance, determining retraining needs).

- **Deployment management** (including ML model loading).

- **Inference management** (e.g. monitoring inference results, managing AI/ML inference functions).

**LCM**, as specified in TS 28.105 [9], encompasses the full lifecycle management of both ML models and AI/ML inference functions. This means that LCM does not only manage the ML model while inference remains a separate process; rather, it ensures a unified management approach that includes both:

- **ML model lifecycle management** covers the entire lifecycle of the ML model itself, including its training, validation, deployment and inference.

- **AI/ML inference function lifecycle management** is also part of LCM, ensuring that inference operations are properly activated, configured, monitored and optimized.

For example, LCM in TS 28.105 [9] enables not just the deployment of an ML model but also the continuous management of its inference functions, such as their activation, configuration and real-time monitoring. This differs from a narrower view of lifecycle (LC), which only considers inference as a step where the ML model is applied, without addressing its ongoing management.

#### 6.2.1.9 Analysis on usage of ML Model identifier in each Working Group

##### 6.2.1.9.1 RAN WG 1

As part of the RAN1 lead work " Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface AI/ML" work in RAN WGs in TR 38.843 [3], RAN is studying two flavours of LCM; Functionality-based and ML model-based (details in clause 4.2.1 of TR 38.843 [3] and clause 6.2.1.8.

- Study on the usage of ML Model identifier is still ongoing and some interim agreements within TR 38.843 [3] are:

- For Functionality-based LCM: Model ID, if needed, can be used in a Functionality (defined in functionality-based LCM) for LCM operations.

NOTE: Functionality-based LCM is most suitable for UE-side ML models

- For Model-ID-based LCM of UE-side models and/or UE-part of two-sided models, model-ID-based LCM operates based on identified models, where a model may be associated with specific configurations/conditions associated with UE capability of an AI/ML-enabled Feature/FG and additional conditions (e.g. scenarios, sites and datasets) as determined/identified between UE-side and NW-side.

- For two-side ML models, in order to select a UE-side ML model (CSI generation model) that is compatible with the NW-side ML model (CSI reconstruction model) pairing information (model pairing) between the UE and gNB can be established based on ML Model identifier(s).

**Analysis on usage of ML Model identifier:**

- Study is ongoing and no concrete conclusions so far.

- For two-sided models, model pairing between UE-side ML model and NW-side ML model is based on ML Model identifiers.

- How an ML Model identifier is assigned to a trained ML model has not been discussed.

- How an ML model identifier is related to different functions has not been discussed.

##### 6.2.1.9.2 RAN WG 3

As part of the RAN WG3 work in TS 38.300 [11]. The following scenarios are supported:

- AI/ML Model Training is located in the OAM and AI/ML Model Inference is located in the NG-RAN node.

- AI/ML Model Training and AI/ML model inference are both located in the NG-RAN node.

**Analysis on usage of ML Model identifier:**

- For the case where AI/ML model is trained at OAM, ML model ID is used as defined in TS 28.105 [9].

##### 6.2.1.9.3 SA WG 2

As part of the work defined in TS 23.288 [8].

- An ML model is trained by the NWDAF MTLF.



Figure 6.2.1.9.3-1: ML model training/identification in AIML related work in SA WG2

- The training may be triggered by request(s) from one more ML model consumer(s) (i.e. NWDAF AnLF). The NWDAF AnLF indicates the purpose of the trained ML model by including an Analytics identifier (and other parameters) as described in TS 23.288 [8].

- The NWDAF MTLF trains an ML model and assigns an ML Model identifier. The trained ML model and assigned ML Model identifier is provisioned to the NWDAF AnLF.

- The AnLF associates the trained ML model and its corresponding ML model identifier to a specific analytics request (identified by an Analytics ID).

- A trained ML model may be stored at a repository (i.e. ADRF) for use by other analytics consumers. The trained ML model is identified at the ADRF based on the ML Model Identifier. No additional metadata are stored at the ADRF to identify the capabilities (e.g. supported Analytics) of the trained ML model.

**Analysis on usage of ML Model identifier:**

- The ML Model identifier identifies the provisioned ML model.

- Only the ML model consumer (AnLF) is aware of the capabilities of the trained ML model by associating the trained ML model and its corresponding ML model identifier to a specific analytics request, identified by an Analytics ID, during an ML model training request.

- When a trained ML model is stored in a repository, the ML Model identifier by its own cannot be used to identify the capabilities of the ML model

##### 6.2.1.9.4 SA WG5

As part of the work defined in TS 28.105 [9]:

- An ML model is trained by the ML training MnS producer.



Figure 6.2.1.9.4-1: ML model training/identification in AIML related work in SA WG5

- The training may be triggered by request(s) from one or more ML training MnS consumer(s). The consumer may be for example a network function, a management function, or an operator. The MnS consumer specifies in the ML training request the inference type which indicates the function or purpose of the ML model, e.g. CoverageProblemAnalysis, that is the MDA type for the coverage problem analysis, see TS 28.104 [71] or NgRanInferenceType which indicates the type of inference that the ML model for NG-RAN supports, see TS 28.105 [9].

- The ML training MnS producer assigns an ML Model identifier to the trained ML model that is provisioned to the MnS consumer. The ML Model identifier identifies the provisioned ML model.

- A trained ML model may be stored at a repository for use by other MnS consumers. The trained ML model is identified at the repository based on the ML Model Identifier. No additional metadata are stored in the repository to identify the capabilities of the trained ML model.

**Analysis on usage of ML Model identifier:**

- The ML Model identifier identifies the provisioned ML model.

- The ML model identifier is used to uniquely identify an ML model instance managed within the 5G system.

-- ML model consumer (MnS Consumer) is aware of the capabilities of the trained ML model by associating the trained ML model and its corresponding ML model identifier to a specific inference type during the ML training request.

- When a trained ML model is stored in a repository, the ML Model identifier by its own cannot be used to identify the capabilities of the ML model.

To support various management operations such as training, inference, and deployment, the following layered identification structure is defined:

- ML model identifier uniquely identifies an ML model. It is assigned when an ML model is created (e.g. after initial training).

- ML Model version specifies a particular version of an ML model to differentiate between versions resulting from re-training or updates.

- ML Model ref is a reference construct that encapsulates both the ML model identifier and ML model version and can optionally include a URI pointing to the model in an internal or external registry/repository.

The use of ML model ref is particularly relevant when:

- Referring to a specific ML model version in ML training request, ML training report, or ML inference job.

- Supporting re-training workflows, where the MnS consumer explicitly refers to an existing model to be updated.

- Supporting external ML model repositories, by using URIs in ML model ref.

- Simplifying job definitions by using a single reference field instead of separate ID and version fields.

NOTE: In initial training scenarios, the ML model ref is typically not used, as the ML model does not yet exist. Instead, the AIML Inference Name is provided in the ML training request to indicate the intended inference behaviour or use case of the new ML model. The MnS producer then generates the ML model identifier and optionally assigns a model version upon successful completion of training.

This structured identification framework enables:

- Lifecycle management and tracking of ML models.

- Flexible orchestration of ML training and ML inference operations.

- Potential extension toward integration with federated or external model repositories.

##### 6.2.1.9.5 SA WG6

As part of the work defined in TS 23.482 [34]

- The ML model ID uniquely identifies the application-layer ML model.



Figure 6.2.1.9.5-1: ML model training/identification in AIML related work in SA WG6

- A VAL server may offload training of an ML model to an AIMLE server.

- An AIML server may train an ML model based on request from consumer(s) (VAL server). Two options are supported:

- A VAL server may request to offload training of an application layer ML model to an AIML server where the model training request includes the ML Model identifier.

- A VAL server may request to train a model for an analytics supported by ADAES where the model training request includes the analytics identifier.

- The ML model information, including ML Model identifier and model capabilities can be stored in a model repository.

**Analysis on usage of ML Model identifier:**

- In scenarios where AIMLE trains an application-layer ML model an ML Model identifier can implicitly identify the capabilities of the ML model.

- In scenarios where AIMLE trains an ML model for ADAES services, ML Model identifier identifies the provisioned ML model. If the trained ML model is stored in an ML repository, then the information stored in the repository may include the capabilities of the ML model identified by an ML Model identifier (e.g. supported Analytics ID).

#### 6.2.1.10 Analysis on ML model pre-specialized training and ML model fine-tuning

The terms 'ML model pre-specialized training' has been defined by SA WG5 as illustrated in Table 6.2.1.10-1. No other 3GPP WG has yet adopted these terms in their activities.

**Table 6.2.1.10-1: Definition of ML model pre-specialized training and ML model fine-tuning**

|  |  |
| --- | --- |
| **TSG (TS/TR)** | **ML model pre-specialized training and ML model fine-tuning** |
| SA WG5 TS 28.105 [9] | **ML model pre-specialized training**: the process of training an ML model on a dataset not specific to any type of inference.  **ML model Fine-tuning**: the process of training a pre-specialised trained ML model to narrow its inference scope to a new single inference type, generating a new ML model.  NOTE 1: The pre-specialised trained model supports an inference scope that may be potentially adapted to support a list of inference types, such as MDA types in MDA, analytics types in NWDAF, type of AI/ML supported functions in NG-RAN, or vendor-specific extensions.  NOTE 2: The inference scope refers to a list of inference types that the ML model may be potentially adapted to support.  NOTE 3: The type of inference represents the specific type of ML inference supported by the model, such as MDA types in MDA, Analytics types in NWDAF, type of AI/ML supported functions in NG-RAN, or vendor-specific extensions. |

The following unified definition for 'ML modelpre-specialized training” and “ML model fine-tuning” is proposed:

**ML model pre-specialized training**: the process of training an ML model on a dataset not specific to any type of inference.

**ML model Fine-tuning**: the process of training a pre-specialised trained ML model to narrow its inference scope to a new single inference type, generating a new ML model.

NOTE 1: The pre-specialised trained model supports an inference scope that may be potentially adapted to support a list of inference types, such as MDA types in MDA, analytics types in NWDAF, type of AI/ML supported functions in NG-RAN, or vendor-specific extensions.

NOTE 2: The inference scope refers to a list of inference types that the ML model may be potentially adapted to support.

NOTE 3: The type of inference represents the specific type of ML inference supported by the model, such as MDA types in MDA, Analytics types in NWDAF, type of AI/ML supported functions in NG-RAN, or vendor-specific extensions.

The terms “ML model pre-specialized training” and “ML model fine-tuning” as defined by SA WG5 introduce a layered training paradigm that differs significantly from traditional concepts such as initial training and re-training. Initial training typically refers to the first-time development of an ML model using a task-specific dataset, while re-training involves updating an existing model with new data to refine or correct its behavior. In contrast, pre-specialized training is task-agnostic and aims to produce a broadly capable model with a wide inference scope. Fine-tuning then adapts this general model to a specific inference type, effectively narrowing its scope and generating a new, specialized model. This two-step approach supports modularity and reuse across multiple domains, such as MDA, NWDAF, and NG-RAN, and enables more efficient deployment of AI/ML capabilities in 3GPP systems. The analysis highlighted above underscores the importance of harmonizing these definitions across WGs to avoid semantic fragmentation and ensure consistent implementation.

#### 6.2.1.11 Analysis on ML model distributed training

The terms 'ML model distributed training' has been defined by SA WG5 as illustrated in Table 6.2.1.11-1. No other 3GPP WG has yet adopted these terms in their activities.

**Table 6.2.1.11-1: Definition of ML model distributed training**

|  |  |
| --- | --- |
| **TSG (TS/TR)** | **ML model distributed training** |
| SA WG5 TS 28.105 [9] | **Distributed training:** a process of distributing the training workload across multiple ML training functions. |

The following unified definition for 'ML modeldistributed training” is proposed:

**Distributed training:** a process of distributing the training workload across multiple ML training functions**.**

**Clarification on terminology: Distributed learning, Distributed training, and Federated learning**  
Within 3GPP documents, the terms *distributed learning*, *distributed training*, and *federated learning* are sometimes used, which can create ambiguity if not clearly distinguished (see also clause 6.2.2):

* **Distributed training** (as applied in SA5 NRM and AI/ML management) refers to splitting a single training job across multiple training functions or nodes to accelerate the process and/or optimise resource utilisation. Depending on the training requirements, the training data may either be partitioned and distributed among nodes (data-parallel approach) or kept intact while nodes focus on different parts of the model (model-parallel approach).
* **Distributed learning** is defined in sub clause 6.2.2,
* **Federated learning** is a specific form of distributed learning designed to protect data privacy.

### 6.2.2 Analysis on Federated Learning

The term 'Horizontal Federated Learning' and 'Vertical Federated Learning' have been defined in SA WG2 and RAN WG1 as well as SA WG5 defines 'Federated Learning', as illustrated in Table 6.2.2-1.

Table 6.2.2-1: Definition of Federated Learning as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | Federated Learning |
| SA WG2 TR 23.700-84 [4] | *Horizontal Federated Learning*: A federated learning technique without exchanging/sharing local data set, wherein the local data set in different FL clients for local model training have the same feature space for different samples (e.g. UE IDs). |
| SA WG2 TS 23.288 [8] | *Vertical Federated Learning*: A federated learning technique without exchanging/sharing local data set, wherein the local data set in different VFL Participant for local model training have different feature spaces for the same samples (e.g. UE IDs). |
| RAN WG1 TR 38.843 [3] | *Federated Learning*: A machine learning technique that trains an AI/ML model across multiple decentralized edge nodes (e.g. UEs, gNBs) each performing local model training using local data samples. The technique requires multiple interactions of the model, but no exchange of local data samples. |
| 3GPP SA5 TS 28.105 [9] | *Federated Learning*: a distributed machine learning approach where the ML model is trained collaboratively by multiple ML training functions. This includes multiple FL clients, which perform training on local data, and one FL server, which aggregates model outcomes from the clients iteratively without exchanging data samples. |

The definition of Federated Learning provided by RAN WG1 appears to only apply to Horizontal Federated Learning, as the phrase "each performing local model training using local data samples" implies that the data samples at individual nodes are distinct. The key difference between Horizontal Federated Learning and Vertical Federated Learning lies in the characteristics of the local datasets:

- Horizontal Federated Learning: Local datasets have the same features but different samples.

- Vertical Federated Learning: local data sets have different features but share same samples.

The definition of Federated Learning provided by SA WG5 highlights the collaborative training process among multiple FL participants, including an FL server and FL clients, without specifying the characteristics of the client datasets. This broader definition facilitates a more comprehensive understanding of both Horizontal Federated Learning (HFL) and Vertical Federated Learning (VFL) that are already defined in the specifications and also offers greater flexibility across more general scenarios, see definitions in clause 5.2.1.2.

The terms "distributed learning" and "federated learning" are often used together as "distributed/federated learning" in SA WG1 TS 22.261 [6]. "Distributed learning" typically refers to a broader set of learning techniques including "federated learning". Although the two terms are related, they are not identical and should be used appropriately based on the context.

The following unified definition for 'Federated Learning' is proposed:

**Federated Learning:** A distributed machine learning approach where the ML model(s) are collaboratively trained by multiple participants, including one acting as an FL server and multiple acting as FL clients, iteratively without exchanging data samples.

The following unified definition for 'Horizontal Federated Learning' is proposed:

**Horizontal Federated Learning:** A federated learning technique without exchanging/sharing local data set, wherein the local data set in different clients for local model training have the same feature space for different samples.

The following unified definition for 'Vertical Federated Learning' is proposed:

**Vertical Federated Learning:** A federated learning technique without exchanging/sharing local data set, wherein the local data set in different clients for local model training have different feature spaces for the same samples.

### 6.2.4 Analysis on Decision vs Prediction vs Output

RAN WG1 and RAN WG3 only uses "prediction" in all corresponding ML related TRs/TSs. SA WG2 uses "output" in all corresponding TRs/TSs where output may include both statistics and predictions. SA WG5 uses "decision" in all corresponding TRs/TSs with few occurrences of "prediction".

The term "output" is proposed as unified term since output may include decision or prediction or statistic or recommendation.

### 6.2.5 Analysis on ML vs AI vs AI/ML

RAN WG1, RAN WG2, RAN WG3 and SA WG1 only uses "AI/ML" in all corresponding ML related TRs/TSs. SA WG2 uses a mix of "ML" and "AI/ML" in all corresponding ML related TRs/TSs. SA WG3, SA WG4 and SA WG6 uses a mix of "AI/ML", "AI" and "ML" in all corresponding ML related TRs/TSs. SA WG5 uses "ML" for training/testing/emulation and "AI/ML" for inference in all corresponding ML related TRs/TSs.

The term "AI/ML" is to be used a unified definition encompassing "AI/ML", "AI" and "ML" in all corresponding ML related TRs/TSs.

### 6.2.6 Analysis on Transfer Learning

The term '*ML Knowledge-based Transfer Learning*' has been defined by SA WG5 in TR 28.858 [19], as illustrated in Table 6.2.6-1. However, the term “Transfer learning” has not been defined in the study or normative phase for SA WG5 Release 19. SA WG6 mentions the term Transfer Learning' in TS 23.482 [34], but definition of the term is not given. SA WG1 also uses the term 'Transfer Learning' in TS 22.261 [6] and TR 22.876 [21] without providing definitions of the term.

Table 6.2.6-1: Definition of Transfer Learning as defined across 3GPP WGs

|  |  |
| --- | --- |
| TSG (TS/TR) | Transfer Learning |
| 3GPP SA5 TR 28.858 [19] | *ML Knowledge-based Transfer Learning*: a technique where the knowledge gained from training of one or more ML models is applied or adapted to improve or develop another ML model. |

The following unified definition for 'Transfer Learning' is proposed (based on SA WG5 definition):

**Transfer Learning:** A machine learning technique where the knowledge acquired from training one or more ML models is leveraged to enhance the performance or accelerate the training of another ML model.

## 6.3 AI/ML related features

### 6.3.1 Analysis on ML model training services

The analysis focuses on the specifications from SA WG2, SA WG5 and SA WG6, considering these are the working groups defining services and operations related to ML model training in 3GPP Release 18. SA WG1, SA WG3, SA WG4, RAN WG1, RAN WG2 and RAN WG3 have not defined any services or operations related to ML model training.

Table 6.3.1-1 provides a detailed overview of the specific services defined by each working group.

The key findings from the analysis are as follows:

- SA WG2: Emphasizes a structured approach to ML model training services by defining a clear consumer-producer relationship. This enables specific entities to consume and produce these services, ensuring a well-defined and controlled environment for service utilization.

- SA WG5: Offers a more flexible approach by defining generic ML model training services. This allows for greater adaptability in implementation and usage, without the constraints of a specific consumer-producer relationship.

- SA WG6: Mirrors the approach of SA WG2, prioritizing a clear consumer-producer relationship for its defined services. This aligns with the structured approach advocated by SA WG2. Additional ML training features for applications are supported in the application enablement layer, such as: federated learning (horizontal, vertical), ML model training capability evaluation for selection of FL members, AI/ML task transfer to a suitable AIML enablement member, support for transfer learning enablement, support for FL member grouping, etc.

While SA WG2 and SA WG6 restrict the potential producers and consumers, SA WG5 emphasizes flexibility and adaptability. The choice of approach will depend on the specific needs and requirements of the individual service provider and consumer.

Table 6.3.1-1: ML model training related services and operations as specified across 3GPP WGs

| ML Model Training | | | |
| --- | --- | --- | --- |
| TSG (TS/TR) | Service/API Type | Service/API/IOC Name | Description [Consumer, Producer] |
|  |  | Nnwdaf\_MLModelProvision\_Subscribe | The consumer subscribes to NWDAF ML model provision with specific parameters to receive a notification when an ML Model matching the subscription parameters becomes available.  *Consumer:* NWDAF AnLF, LMF  *Producer:* NWDAF MTLF |
|  | ML Model Provisioning Services | Nnwdaf\_MLModelProvision\_Unsubscribe | The consumer unsubscribes to NWDAF ML model provision.  *Consumer:* NWDAF AnLF, LMF  *Producer:* NWDAF MTLF |
|  |  | Nnwdaf\_MLModelProvision\_Notify | The NWDAF notifies the ML model information to the consumer which has subscribed to the NWDAF ML model provision service.  *Consumer:* NWDAF AnLF, LMF  *Producer:* NWDAF MTLF |
| SA WG2 TS 23.288 [8] | ML Model Information Services | Nnwdaf\_MLModelInfo\_Request | The consumer requests and gets NWDAF ML Model Information.  *Consumer:* NWDAF AnLF, LMF  *Producer:* NWDAF MTLF |
|  |  | Nnwdaf\_MLModelTraining\_Subscribe | The consumer subscribes to NWDAF ML model training with specific parameters.  *Consumer:* NWDAF MTLF  *Producer:* NWDAF MTLF |
|  | ML Model Training Services | Nnwdaf\_MLModelTraining\_Unsubscribe | The consumer terminates NWDAF ML model training.  *Consumer:* NWDAF MTLF  *Producer:* NWDAF MTLF |
|  |  | Nnwdaf\_MLModelTraining\_Notify | The NWDAF notifies about the trained ML model to the consumer which has subscribed to the NWDAF ML model training service.  *Consumer:* NWDAF MTLF  *Producer:* NWDAF MTLF |
|  | ML Model Training Information Services | Nnwdaf\_MLModelTrainingInfo\_Request | The consumer requests for the information about NWDAF ML model training with specific parameters.  *Consumer:* NWDAF MTLF  *Producer:* NWDAF MTLF |
|  |  | Nnwdaf\_VFLTraining\_Subscribe | The consumer subscribes to VFL ML Model training information.  *Consumer:* NWDAF, AF, NEF  *Producer:* NWDAF |
|  | VFL Training Services | Nnwdaf\_VFLTraining\_Unsubscribe | The consumer terminates NWDAF VFL ML Model training.  *Consumer:* NWDAF, AF, NEF  *Producer:* NWDAF |
|  |  | Nnwdaf\_VFLTraining\_Notify | The NWDAF notifies the consumer of client intermediate training result of the local ML model.  *Consumer:* NWDAF, AF, NEF  *Producer:* NWDAF |
|  |  | Nnwdaf\_VFLTraining\_Request | The consumer requests NWDAF VFL client to check if it can support requirements for VFL.  *Consumer:* NWDAF, AF, NEF  *Producer:* NWDAF |
|  |  | MLTrainingRequest | It represents the ML model training request to train an ML model which is triggered by the ML training MnS consumer towards the ML training MnS producer.  *Consumer:* Any authorized network function, any authorized management function, operator  *Producer:* Any function that is capable of training an ML model |
| SA WG5 TS 28.105 [9] | ML Training Management Services | MLTrainingReport | It represents the ML model training report provided by the ML training MnS producer to the ML training MnS consumer who has requested for ML model training.  *Consumer:* Any authorized network function, any authorized management function, operator  *Producer:* Any function that is capable of training an ML model |
|  |  | MLTrainingProcess | It represents the ML model training process. When a ML model training process starts, an instance of the MLTrainingProcess is created by the MnS Producer and notification is sent to MnS consumer who has subscribed to it.  *Consumer:* Any authorized network function, any authorized management function, operator  *Producer:* Any function that is capable of training an ML model |
| SA WG6 TS 23.482 [34] | ML Model Training APIs | Aimles\_MLModelTraining Request | The consumer sends an ML model training request to the producer, requesting to assist in its ML model training. This request consists of ML model information or ML model requirement information, etc.  *Consumer:* VAL server  *Producer:* AIMLE Server |
|  |  | Aimles\_MLModelTraining Response | If the consumer is authorized, the producer identifies and selects the appropriate ML model for training based on the ML model requirement information. The producer returns a success response indicating the selected ML model for training; otherwise, a failure response indicating the reason for failure.  *Consumer:* VAL server  *Producer:* AIMLE Server |

### 6.3.2 Analysis on analytics related services

This clause focuses on the specifications from SA WG2, SA WG5 and SA WG6, considering these are the working groups defining services and operations related to ML model inference in 3GPP Release 18. SA WG1, SA WG3, SA WG4, RAN WG1, RAN WG2 and RAN WG3 have not defined any services or operations related to ML model inference.

Table 6.3.2-1 provides a detailed overview of the specific services defined by each working group.

The key findings from the analysis are as follows:

- SA WG2: Describes ML model inference by defining analytics services through a clear consumer-producer relationship. It defines several analytics types in TS 23.288 [8], each one supported by the NWDAF/RE-NWDAF AnLF and requested/subscribed by the NWDAF/RE-NWDAF AnLF consumer using the defined analytics services.

- SA WG5: Describes ML model inference by defining generic analytics services without specific consumer-producer relationship in TS 28.105 [9] and TS 28.104 [71]. It defines several analytics types in TS 28.104 [71], each one supported by an MnS producer and requested by the MnS consumer using the defined analytics services.

- SA WG6: Describes ML model inference by defining individual analytics services for each analytics type. It defines several analytics types in TS 23.436 [33].

While SA WG2 and SA WG6 restrict the potential producers and consumers, SA WG5 emphasizes flexibility and adaptability. Moreover, SA WG2 and SA WG5 defines several analytics types that can be supported by an entity and requested by another entity using the defined ML model inference services. In addition, in SA WG6, individual services are defined for each analytics type.

Table 6.3.2-1: Analytics and inference related services and operations as specified across 3GPP WGs

| ML Model inference | | | |
| --- | --- | --- | --- |
| TSG (TS/TR) | Service/API Type | Service/API/IOC Name | Description [Consumer, Producer] |
|  | Network Data Analytics Subscription Services | Nnwdaf\_AnalyticsSubscription\_Subscribe | The consumer subscribes for network data analytics and optionally its corresponding analytics accuracy information with specific parameters.  *Consumer:* PCF, NSSF, AMF, SMF, NEF, AF, OAM, CEF, NWDAF, DCCF, LMF  *Producer:* NWDAF AnLF |
|  |  | Nnwdaf\_AnalyticsSubscription\_Unsubscribe | The consumer unsubscribes for network data analytics.  *Consumer:* PCF, NSSF, AMF, SMF, NEF, AF, OAM, CEF, NWDAF, DCCF, LMF  *Producer:* NWDAF AnLF |
|  |  | Nnwdaf\_AnalyticsSubscription\_Notify | The NWDAF notifies the analytics and optionally Analytics Accuracy Information to the consumer which has subscribed to the NWDAF analytics subscription service.  *Consumer:* PCF, NSSF, AMF, SMF, NEF, AF, OAM, CEF, NWDAF, DCCF, LMF  *Producer:* NWDAF AnLF |
|  |  | Nnwdaf\_AnalyticsSubscription\_Transfer | The consumer NWDAF requests NWDAF for transferring analytics subscriptions from the consumer NWDAF.  *Consumer:* NWDAF AnLF  *Producer:* NWDAF AnLF |
| SA WG2 TS 23.288 [8] | Network Data Analytics Information Services | Nnwdaf\_AnalyticsInfo\_Request | The consumer requests NWDAF operator specific analytics and optionally Analytics Accuracy Information with specific parameters.  *Consumer:* PCF, NSSF, AMF, SMF, NEF, AF, OAM, CEF, NWDAF, DCCF, LMF  *Producer:* NWDAF AnLF |
|  |  | Nnwdaf\_AnalyticsInfo\_ContextTransfer | The consumer NWDAF requests NWDAF to transfer context information related to analytics subscriptions.  *Consumer:* NWDAF AnLF  *Producer:* NWDAF AnLF |
|  |  | Nnwdaf\_RoamingAnalytics\_Subscribe | The consumer subscribes for network data analytics related to roaming UEs.  *Consumer:* H-RE-NWDAF, V-RE-NWDAF  *Producer:* H-RE-NWDAF, V-RE-NWDAF |
|  | Network Data Roaming Analytics Services | Nnwdaf\_RoamingAnalytics\_Unsubscribe | The consumer unsubscribes for network data analytics related to roaming UEs.  *Consumer:* H-RE-NWDAF, V-RE-NWDAF  *Producer:* H-RE-NWDAF, V-RE-NWDAF |
|  |  | Nnwdaf\_RoamingAnalytics\_Notify | The NWDAF notifies the analytics related to roaming UE(s) to the consumer which has subscribed to the NWDAF roaming analytics subscription service.  *Consumer:* H-RE-NWDAF, V-RE-NWDAF  *Producer:* H-RE-NWDAF, V-RE-NWDAF |
|  |  | Nnwdaf\_RoamingAnalytics\_Request | The consumer requests NWDAF operator specific related to roaming UEs.  *Consumer:* H-RE-NWDAF, V-RE-NWDAF  *Producer:* H-RE-NWDAF, V-RE-NWDAF |
|  |  | Nnwdaf\_VFLInference\_Subscribe | The consumer subscribes to VFL inference.  Consumer: NWDAF, AF, NEF  Producer: NWDAF |
|  | Nnwdaf\_VFLInference | Nnwdaf\_VFLInference\_Unsubscribe | The consumer unsubscribes to VFL inference.  Consumer: NWDAF, AF, NEF  Producer: NWDAF |
|  |  | Nnwdaf\_VFLInference\_Notify | The consumer notifies VFL inference result.  Consumer: NWDAF, AF, NEF  Producer: NWDAF |
|  |  | Nnwdaf\_VFLInference\_Request | The consumer requests the NWDAF to perform a one-time VFL inference.  Consumer: NWDAF, AF, NEF  Producer: NWDAF |
| SA WG5 TS 28.104 [71] | Management Data Analytics Services | MDARequest | It represents the management data analytics output request which is created by an MDA MnS consumer towards the MDA MnS producer.  *Consumer:* Any authorized network function, any authorized management function, operator  *Producer:* Any function that is capable of producing management data analytics |
|  |  | MDAReport | It represents the management data analytics report containing the outputs for one or more MDA types delivered to the MDA consumer who has requested for management data analytics.  *Consumer:* Any authorized network function, any authorized management function, operator  *Producer:* Any function that is capable of producing management data analytics |
|  | SS\_ADAE\_VAL\_performance\_analytics | VAL\_performance\_analytics\_subscribe | The consumer subscribes for VAL performance analytics.  *Consumer:* VAL server  *Producer:* ADAE server |
|  |  | VAL\_performance\_analytics\_notify | The consumer is notified by ADAES on the VAL performance analytics.  *Consumer:* VAL server  *Producer:* ADAE server |
|  | SS\_ADAE\_slice\_performance\_analytics | slice\_performance\_analytics\_subscribe | The consumer subscribes for slice specific performance analytics.  *Consumer:* VAL server  *Producer:* ADAE server |
|  |  | slice\_performance\_analytics\_notify | The consumer is notified by ADAES on the slice specific performance analytics.  *Consumer:* VAL server  *Producer:* ADAE server |
|  | SS\_ADAE\_UE-to-UE\_performance\_analytics | UE-to-UE performance\_analytics\_subscribe | The consumer subscribes for UE-to-UE performance analytics. |
|  |  | UE-to-UE performance\_analytics\_notify | The consumer is notified by ADAES on the slice specific performance analytics.  *Consumer:* VAL server  *Producer:* ADAE server |
| SA WG6 TS 23.436 [33] | SS\_ADAE\_server-to-server\_performance\_analytics | server-to-server\_performance\_analytics\_subscribe | The consumer subscribes to the ADAE server for Server-to-server performance analytics.  *Consumer:* VAL server, EES  *Producer:* ADAE server |
|  |  | server-to-server\_performance\_analytics\_notify | The consumer is notified by the ADAE server on the Server-to-server performance analytics.  *Consumer:* VAL server, EES  *Producer:* ADAE server |
|  | SS\_ADAE\_location\_accuracy\_analytics | Location\_accuracy\_analytics\_subscribe | The consumer subscribes for location accuracy analytics.  *Consumer:* VAL server  *Producer:* ADAE server |
|  |  | Location\_accuracy\_analytics\_notify | The consumer is notified by ADAES on the location accuracy analytics.  *Consumer:* VAL server  *Producer:* ADAE server |
|  | SS\_ADAE\_service\_API\_analytics | Service\_API\_analytics\_subscribe | The consumer subscribes for service API analytics.  *Consumer:* VAL server, Subscriber, API invoker  *Producer:* ADAE server |
|  |  | Service\_API\_analytics\_notify | The consumer is notified by ADAES on the location accuracy analytics.  *Consumer:* VAL server, Subscriber, API invoker  *Producer:* ADAE server |
|  | SS\_ADAE\_slice\_usage\_pattern\_analytics | slice\_usage\_pattern\_analytics\_subscribe | The consumer subscribes for slice usage pattern analytics.  *Consumer:* VAL server, SEAL server  *Producer:* ADAE server |
|  |  | slice\_usage\_pattern\_analytics\_notify | The consumer is notified by ADAES on the slice usage pattern analytics.  *Consumer:* VAL server, SEAL server  *Producer:* ADAE server |
|  | SS\_ADAE\_edge\_analytics | edge\_analytics\_subscribe | The consumer subscribes for edge load analytics.  *Consumer:* VAL server, ECS, EES  *Producer:* ADAE server |
|  |  | edge\_analytics\_notify | The consumer is notified by ADAES on the edge load analytics.  *Consumer:* VAL server, ECS, EES  *Producer:* ADAE server |
|  |  | edge\_analytics\_get | The consumer requests edge analytics data.  *Consumer:* VAL server, ECS, EES  *Producer:* ADAE server |
|  | SS\_ADAES\_slice\_usage\_stats | slice\_usage\_stats\_get | The consumer requests and receives slice usage statistics from ADAE server.  *Consumer:* VAL server  *Producer:* ADAE server |
|  | SS\_ADAES\_edge\_preparation\_analytics | edge\_preparation\_analytics\_subscribe | The consumer subscribes for edge computing preparation analytics.  *Consumer:* VAL server, ECS, EES  *Producer:* ADAE server |
|  |  | edge\_preparation\_analytics\_notify | The consumer is notified by the ADAE server on the edge computing preparation analytics.  *Consumer:* VAL server, ECS, EES  *Producer:* ADAE server |
|  |  | edge\_preparation\_analytics\_get | The consumer requests edge computing preparation analytics  *Consumer:* VAL server, ECS, EES  *Producer:* ADAE server |
|  | SS\_ADAE\_collision\_detection\_analytics | collision\_detection\_analytics\_subscribe | The consumer subscribes for collision detection analytics.  *Consumer:* VAL server, LM server, UAE server, UAS application specific server  *Producer:* ADAE server |
|  |  | collision\_detection\_analytics\_notify | The consumer is notified by the ADAE server on collision detection analytics.  *Consumer:* VAL server, LM server, UAE server, UAS application specific server  *Producer:* ADAE server |
|  |  | collision\_detection\_analytics\_get | The consumer requests collision detection analytics.  *Consumer:* VAL server, LM server, UAE server, UAS application specific server  *Producer:* ADAE server |
|  | SS\_ADAE\_location-related\_UE\_group\_analytics | location-related\_UE\_group\_analytics\_subscribe | The consumer subscribes for location-related UE group analytics.  *Consumer:* LM server  *Producer:* ADAE server |
|  |  | location-related\_UE\_group\_analytics\_notify | The consumer is notified by the ADAE server on location-related UE group analytics.  *Consumer:* LM server  *Producer:* ADAE server |
|  |  | location-related\_UE\_group\_analytics\_get | The consumer requests location-related UE group analytics.  *Consumer:* LM server  *Producer:* ADAE server |
|  | SS\_ ADAE\_AIML\_member\_capability\_analytics | AIML\_member\_capability\_analytics\_subscribe | The consumer subscribes for application layer AIML member capability analytics.  *Consumer:* VAL server, AIMLE server  *Producer:* ADAE server |
|  |  | AIML\_member\_capability\_analytics\_notify | The consumer is notified by the ADAE server on application layer AIML member capability analytics.  *Consumer:* VAL server, AIMLE server  *Producer:* ADAE server |
|  |  | AIML\_member\_capability\_analytics\_get | The consumer requests application layer AIML member capability analytics.  *Consumer:* VAL server, AIMLE server  *Producer:* ADAE server |
|  | SS\_ADAE\_UE\_RAT\_connectivity\_analytics API | UE\_RAT\_connectivity\_analytics\_subscribe | The consumer requests UE RAT connectivity analytics.  *Consumer:*VAL server  *Producer:*ADAE server |
|  |  | UE\_RAT\_connectivity\_analytics\_notify | The consumer is notified by the ADAE server on UE RAT connectivity analytics.  *Consumer:*VAL server  *Producer:*ADAE server |

### 6.3.3 Analysis on ML performance evaluation and monitoring

The analysis focuses on the specifications from SA WG2, SA WG5 and SA WG6, considering these are the working groups defining services and operations related to ML performance evaluation in 3GPP Release 18 / Release 19. SA WG1, SA WG3, SA WG4, RAN WG1, RAN WG2 and RAN WG3 have not defined any services or operations related to ML performance evaluation.

Table 6.3.3-1 provides a detailed overview of the specific services defined by each working group.

The key findings from the analysis are as follows:

- SA WG2: Dedicated ML model monitoring services are defined with a specific consumer-producer relationship. Additionally, analytics subscription and ML model training subscription services may indicate the performance requirements which the producer has to satisfy when providing the analytics or training the ML model. The focus in SA WG2 has been on the accuracy aspects of ML model performance.

- SA WG5: SA WG5 defines the framework and mechanisms for performance assurance including performance metrics (ModelPerformance clause 7.4.1 of 28.105 [9]) on which the performance of an ML Model can be ascertained. ML training, ML testing and ML inference services indicate the performance requirements which the producer has to satisfy for the consumer when training the ML model or testing the ML model or providing the inferences. The achieved performance of ML Model is communicated to the consumer via MLTrainingReportMLTestingReport and AIMLInferenceReport.

- SA WG6: Dedicated ML model monitoring services are defined with specific consumer-producer relationship. Additionally, ML model training APIs indicate the performance requirements that the producer has to satisfy when providing the ML model. No specific ML model performance metrics are standardized in SA WG6.

Table 6.3.3-1: ML model performance monitoring services and operations as specified in 3GPP WGs

|  |  |  |  |
| --- | --- | --- | --- |
| ML Model Training | | | |
| TSG (TS/TR) | Service/API Type | Service/API/IOC Name | Description [Consumer, Producer] |
|  |  | Nnwdaf\_MLModelMonitor\_Subscribe | The consumer subscribes to NWDAF for the monitored ML Model accuracy information and Analytics Feedback Information for the analytics generated by the NWDAF with specific parameters.  *Consumer*: NWDAF  *Producer*: NWDAF |
|  |  | Nnwdaf\_MLModelMonitor\_Unsubscribe | The consumer unsubscribes to the NWDAF for the monitored ML Model accuracy information and Analytics Feedback Information for the analytics generated by the NWDAF.  *Consumer*: NWDAF  *Producer*: NWDAF |
| SA WG2 TS 23.288 [8] | ML Model Monitoring Services | Nnwdaf\_MLModelMonitor\_Notify | NWDAF notifies the monitored ML Model accuracy information and Analytics Feedback Information for the analytics generated by the NWDAF to the consumer who has subscribed to the specific NWDAF service.  *Consumer*: NWDAF  *Producer*: NWDAF |
|  |  | Nnwdaf\_MLModelMonitor\_Register | The consumer registers the use and monitoring capability for an ML Model at an NWDAF containing MTLF.  *Consumer*: NWDAF  *Producer*: NWDAF |
|  |  | Nnwdaf\_MLModelMonitor\_Deregister | The consumer deregisters, from an NWDAF containing MTLF, a previous MLModelMonitor registration, e.g. when the consumer is no longer using or monitoring the accuracy of the analytics generated using the ML Model.  *Consumer*: NWDAF  *Producer*: NWDAF |
| SA WG6 TS 23.482 [34] | ML Model Performance Monitoring APIs | MLModelPerfMonitor\_Subscribe | The consumer subscribes for ML model performance monitoring.  *Consumer:* VAL server  *Producer:* AIMLE Server |
|  |  | MLModelPerfMonitor\_Notify | The consumer is notified by ML repository on the ML model performance monitoring.  *Consumer:* VAL server  *Producer:* AIMLE Server |

### 6.3.4 Analysis on data collection and management for AI/ML

The analysis focuses on the specifications from SA WG2, SA WG6 and RAN WG3, considering these are the working groups defining services and operations related to data collection for AI/ML in 3GPP Release 18. SA WG1, SA WG3, SA WG4, SA WG5, RAN WG1 and RAN WG2 have not defined any services or operations related to data collection for AI/ML in Release 18. SA WG5 specifies data collection and performance measurement services that can be leveraged for AI/ML purposes (see TS 28.622 [72]). Table 6.3.4-1 provides a detailed overview of the specific services defined by each working group.

The key findings from the analysis are as follows:

- SA WG2: Defines multiple network functions capable of producing data collection services and defines a function for data storage related services. For example, the DCCF coordinates and manages the collection of data from various network functions for purposes such as computation of analytics and Analytics/ML Model Accuracy monitoring. Leverages event exposure framework and defines event exposure services for network functions that can be consumed by NWDAF (see clause 6.2.2.1 of TS 23.288 [8]) and DCCF (see clause 6.2.6.3 of TS 23.288 [8]). Defines data collection for AI/ML services through a clear consumer-producer relationship.

- SA WG6: Defines network functions similar to those in SA WG2. Data collection services defined for A-DCCF as well as data storage services defined for A-ADRF are generic and applicable to any ADAE services; however, while A-DCCF APIs are defined in a generic manner, the A-ADRF APIs (for data storage and fetching) are defined in a per use case specific manner.

- RAN WG3: Defines data collection messages exchanged between two gNBs over the Xn interface, in a P2P manner. It is to be noted that procedures used for AI/ML support in the NG-RAN shall be "data type agnostic", which means that the intended use of the data (e.g. input, output, feedback) shall not be indicated.

- RAN WG2: Defines data collection configuration procedures for offline training of network side models over existing RRC messages between UE and gNB. RAN WG2 also introduces the logging of data within UE and the retrieval of the logged data by gNB via UE Information procedure.

While SA WG2 and SA WG6 both define data collection services, their approaches to data storage and retrieval are different. SA WG2 defines generic data storage and retrieval services that can be supported by an entity (ADRF) and requested by another entity but SA WG6 defines both generic and individual services (related to each analytics type) for storing and retrieving data. RAN WG3 operates independently and is unrelated to services defined in SA WGs and therefore can coexist.

Table 6.3.4-1: Data Collection for AI/ML related services and operations as specified across 3GPP WGs

|  |  |  |  |
| --- | --- | --- | --- |
| Data Collection for AI/ML | | | |
| TSG (TS/TR) | Service/API/Message Type | Service/API/IOC/Message Name | Description [Consumer, Producer] |
|  |  | Namf\_EventExposure\_Subscribe | The NWDAF uses this service operation to subscribe to or modify event reporting for one UE, a group of UE(s) or any UE.  *Producer*: AMF |
|  |  | Namf\_EventExposure\_Unsubscribe | The NWDAF uses this service operation to unsubscribe for a specific event for one UE, group of UE(s), any UE.  *Producer*: AMF |
|  |  | Namf\_EventExposure\_Notify | Provides the previously subscribed event information to the NWDAF which has subscribed to that event before.  *Producer*: AMF |
|  |  | Nsmf\_EventExposure\_Subscribe | This service operation is used by an NWDAF to subscribe or modify a subscription for event notifications on a specified PDU Session or for all PDU Sessions of one UE, group of UE(s) or any UE.  *Producer*: SMF |
| SA WG2 TS 23.288 [8] |  | Nsmf\_EventExposure\_UnSubscribe | This service operation is used by an NWDAF to unsubscribe event notifications.  *Producer*: SMF |
|  | Event Exposure services | Nsmf\_EventExposure\_Notify | Report UE PDU Session related event(s) to the NWDAF which has subscribed to the event report service.  *Producer*: SMF |
|  |  | Npcf\_EventExposure\_Subscribe | The NWDAF uses this service operation to subscribe to or modify event reporting for a group of UE(s) or any UE accessing a combination of (DNN, S-NSSAI).  *Producer*: PCF |
|  |  | Npcf\_EventExposure\_Unsubscribe | The NWDAF uses this service operation to unsubscribe for a specific event for a group of UE(s) or any UE accessing a combination of (DNN, S-NSSAI).  *Producer*: PCF |
|  |  | Npcf\_EventExposure\_Notify | This service operation reports the event to the NWDAF that has previously subscribed either using Npcf\_EventExposure\_Subscribe service operation or provided as part of the Data Set Application Data and Data Subset Service Parameters stored in UDR.  *Producer*: PCF |
|  |  | Nudm\_EventExposure\_Subscribe | The NWDAF subscribes to receive an event.  *Producer*: UDM |
|  |  | Nudm\_EventExposure\_Unsubscribe | The NWDAF deletes the subscription of an event if already defined in UDM.  *Producer*: UDM |
|  |  | Nudm\_EventExposure\_Notify | UDM reports the event to the NWDAF that has previously subscribed.  *Producer*: UDM |
|  |  | Nudm\_EventExposure\_ModifySubscription | The NWDAF requests to modify an existing subscription to event notifications.  *Producer*: UDM |
|  |  | Nnef\_EventExposure\_Subscribe | The NWDAF subscribes to receive an event, or if the event is already defined in NEF, then the subscription is updated.  *Producer*: NEF |
|  |  | Nnef\_EventExposure\_Unsubscribe | The NWDAF deletes an event if already defined in NEF.  *Producer*: NEF |
|  |  | Nnef\_EventExposure\_Notify | NEF reports the event to the NWDAF that has previously subscribed.  *Producer*: NEF |
|  |  | Naf\_EventExposure\_Subscribe | The NWDAF subscribes the event to collect AF data for UE(s), group of UEs, or any UE, or updates the subscription which is already defined in AF.  *Producer*: AF |
|  |  | Naf\_EventExposure\_Unsubscribe | The NWDAF unsubscribes for a specific event.  *Producer*: AF |
|  |  | Naf\_EventExposure\_Notify | The AF provides the previously subscribed event information to the NWDAF which has subscribed to that event before.  *Producer*: AF |
|  |  | Nnsacf\_SliceEventExposure\_Subscribe | This service operation is used by the NWDAF to subscribe or modify a subscription with the NSACF for event based notifications of the current number of UEs registered for a network slice or the current number of PDU Sessions established on a network slice.  *Producer*: NSACF |
|  |  | Nnsacf\_SliceEventExposure\_Unsubscribe | This service operation is used by the NWDAF to unsubscribe from the event notification.  *Producer*: NSACF |
|  |  | Nnsacf\_SliceEventExposure\_Notify | This service operation is used by the NSACF to report the current number of UEs registered with a network slice or the current number of PDU Sessions established on a network slice in numbers or in percentage from the maximum allowed numbers, based on threshold or at expiry of periodic timer.  *Producer*: NSACF |
|  |  | Nupf\_EventExposure\_Subscribe | This service operation reports the event and information to the NWDAF that has subscribed implicitly.  *Producer*: UPF |
|  |  | Nupf\_EventExposure\_Unsubscribe | This service operation is used by an NWDAF to subscribe or modify a subscription to UPF event exposure notifications e.g. for the purpose of UPF data collection on a specified PDU Session or for all PDU Sessions of one UE or any UE.  *Producer*: UPF |
|  |  | Nupf\_EventExposure\_Notify | The NF consumer uses this service operation to unsubscribe for a specific event.  *Consumer*: Any NF  *Producer*: UPF |
|  |  | Nscp\_EventExposure\_Notify | The NWDAF uses this service operation to unsubscribe for a specific event.  *Producer*: SCP |
|  |  | Nscp\_EventExposure\_Subscribe | This service operation is used by an NWDAF to subscribe or modify a subscription to SCP event exposure notifications.  *Producer*: SCP |
|  |  | Nscp\_EventExposure\_Unsubscribe | The NWDAF uses this service operation to unsubscribe from an existing subscription.  *Producer*: SCP |
|  |  | Nnwdaf\_DataManagement\_Subscribe | The consumer subscribes to data exposed by an NWDAF. It can be historical data or runtime data. The subscription includes service operation specific parameters that identify the data to be provided and may include formatting and processing instructions that specify how the data is to be delivered to the consumer.  *Consumer*: NWDAF, DCCF  *Producer*: NWDAF |
|  | NWDAF Data Management services | Nnwdaf\_DataManagement\_Unsubscribe | The consumer unsubscribes to the data exposed by an NWDAF.  *Consumer*: NWDAF, DCCF  *Producer*: NWDAF |
|  |  | Nnwdaf\_DataManagement\_Notify | The NWDAF notifies the consumer of the requested data or notifies of the availability of previously subscribed data when delivery is via an NWDAF. The NWDAF may also notify the consumer when Data or Analytics is to be deleted.  *Consumer*: NWDAF, DCCF, MFAF, ADRF  *Producer*: NWDAF |
|  |  | Nnwdaf\_DataManagement\_Fetch | The consumer retrieves from the NWDAF subscribed data, as indicated by Fetch Instructions from Nnwdaf\_DataManagement\_Notify.  *Consumer*: NWDAF, DCCF, MFAF, ADRF  *Producer*: NWDAF |
|  |  | Nnwdaf\_RoamingData\_Subscribe | The consumer subscribes for input data related to roaming UE(s) for NWDAF analytics. The subscription includes service operation specific parameters that identify the data to be provided and may include formatting and processing instructions that specify how the data is to be delivered to the consumer.  *Consumer*: H-RE-NWDAF, V-RE-NWDAF  *Producer*: H-RE-NWDAF, V-RE-NWDAF |
|  | NWDAF Roaming Data services | Nnwdaf\_RoamingData\_Unsubscribe | The consumer unsubscribes to input data related to roaming UE(s).  *Consumer*: H-RE-NWDAF, V-RE-NWDAF  *Producer*: H-RE-NWDAF, V-RE-NWDAF |
|  |  | Nnwdaf\_RoamingData\_Notify | NWDAF notifies the consumer about input data related to roaming UE(s) that the consumer has subscribed to.  *Consumer*: H-RE-NWDAF, V-RE-NWDAF  *Producer*: H-RE-NWDAF, V-RE-NWDAF |
|  |  | Ndccf\_DataManagement\_Subscribe | The consumer subscribes to receive data or analytics from the DCCF. The subscription includes service operation specific parameters that identify the data or analytics to be provided and may include formatting and processing instructions that specify how the data is to be delivered to the consumer. The consumer may also request that data be stored in an ADRF or an NWDAF hosting ADRF functionality.  *Consumer*: NWDAF, PCF, NSSF, AMF, SMF, NEF, AF, ADRF  *Producer*: DCCF |
|  |  | Ndccf\_DataManagement\_Unsubscribe | The consumer unsubscribes to DCCF for data or analytics.  *Consumer*: NWDAF, PCF, NSSF, AMF, SMF, NEF, AF, ADRF  *Producer*: DCCF |
|  | DCCF Data Management Services | Ndccf\_DataManagement\_Notify | DCCF notifies the consumer instance of the requested data or analytics according to the request or notifies of the availability of previously subscribed Data or Analytics when data delivery is via the DCCF. The DCCF may also notify the consumer instance when Data or Analytics is to be deleted.  *Consumer*: NWDAF, PCF, NSSF, AMF, SMF, NEF, AF, ADRF  *Producer*: DCCF |
|  |  | Ndccf\_DataManagement\_Fetch | The consumer retrieves from the DCCF, data or analytics as indicated by Ndccf\_DataManagement\_Notify Fetch Instructions.  *Consumer*: NWDAF, PCF, NSSF, AMF, SMF, NEF, AF, ADRF  *Producer*: DCCF |
|  |  | Ndccf\_DataManagement\_Transfer | The Source DCCF transfers UE data subscription context to the target DCCF.  *Consumer*: DCCF  *Producer*: DCCF |
|  |  | Nmfaf\_3daDataManagement\_Configure | The consumer configures or reconfigures the MFAF to map data or analytics received by the MFAF to out-bound notification endpoints and to format and process the out-bound data or analytics.  *Consumer*: DCCF, NWDAF  *Producer*: MFAF |
|  | MFAF Data Management Services | Nmfaf\_3daDataManagement\_Deconfigure | The consumer configures the MFAF to stop mapping data or analytics received by the MFAF to one or more out-bound notification endpoints.  *Consumer*: DCCF, NWDAF  *Producer*: MFAF |
|  |  | Nmfaf\_3caDataManagement\_Notify | MFAF provides data or analytics or notification of availability of data or analytics to notification endpoints.  *Consumer*: NWDAF, PCF, NSSF, AMF, SMF, NEF, AF, ADRF  *Producer*: MFAF |
|  |  | Nmfaf\_3caDataManagement\_Fetch | The consumer retrieves from the MFAF, data or analytics as indicated by Nmfaf\_3caDataManagement\_Notify Fetch Instructions.  *Consumer*: NWDAF, PCF, NSSF, AMF, SMF, NEF, AF, ADRF  *Producer*: MFAF |
|  |  | Nadrf\_DataManagement\_StorageRequest | The consumer NF uses this service operation to request the ADRF to store data or analytics. Data or analytics are provided to the ADRF in the request message.  *Consumer*: DCCF, NWDAF, MFAF  *Producer*: ADRF |
|  |  | Nadrf\_DataManagement\_StorageSubscriptionRequest | The consumer (NWDAF or DCCF) uses this service operation to request the ADRF to initiate a subscription for data or analytics. Data or analytics provided in notifications as a result of the subsequent subscription by the ADRF are stored in the ADRF.  *Consumer*: NWDAF, DCCF  *Producer*: ADRF |
|  | ADRF Data Management Services | Nadrf\_DataManagement\_StorageSubscriptionRemoval | The consumer NF uses this service operation to request that the ADRF no longer subscribes to data or analytics it is collecting and storing.  *Consumer*: NWDAF, DCCF  *Producer*: ADRF |
|  |  | Nadrf\_DataManagement\_RetrievalRequest | The consumer NF uses this service operation to retrieve stored data or analytics from the ADRF. The Nadrf\_DataManagement\_RetrievalRequest response either contains the data or analytics or provides instructions for fetching the data or analytics.  *Consumer*: NWDAF, DCCF  *Producer*: ADRF |
|  |  | Nadrf\_DataManagement\_RetrievalSubscribe | The consumer NF uses this service operation to retrieve stored data or analytics from the ADRF and to receive future notifications containing the corresponding data or analytics received by ADRF.  *Consumer*: NWDAF, DCCF  *Producer*: ADRF |
|  |  | Nadrf\_DataManagement\_RetrievalUnsubscribe | The consumer NF uses this service operation to request that the ADRF no longer sends data or analytics to a notification endpoint.  *Consumer*: NWDAF, DCCF  *Producer*: ADRF |
|  |  | Nadrf\_DataManagement\_RetrievalNotify | This service operation provides consumers with either data or analytics from an ADRF, or instructions to fetch the data or analytics from an ADRF. The notifications are provided to consumers that have subscribed using the Nadrf\_DataManagement\_RetrievalSubscribe service operation.  *Consumer*: NWDAF, DCCF  *Producer*: ADRF |
|  |  | Nadrf\_DataManagement\_Delete | This service operation instructs the ADRF to delete stored data.  *Consumer*: NWDAF, DCCF  *Producer*: ADRF |
|  |  | SS\_AADRF\_Data\_Collection Subscribe | The consumer subscribes for offline data from A-ADRF.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_Data\_Collection Notify | The consumer is receiving the offline data from A-ADRF as notification, based on subscription.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_ AADRF\_Historical\_ServiceAPI\_Logs Get | The consumer requests API logs from A-ADRF.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_NetworkSlice\_Data Get | The consumer requests network slice data from A-ADRF.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_Location\_Accuracy\_Data Get | The consumer is receiving offline location analytics/data from A-ADRF.  *Consumer*: ADAES  *Producer*: A-ADRF |
| SA WG6 TS 23.436 [33] | A-ADRF Data Collection APIs | SS\_AADRF\_EdgeData\_Collection Subscribe | The consumer subscribes for offline edge data from A-ADRF.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_EdgeData\_Collection Notify | The consumer is receiving the offline edge data from A-ADRF as notification, based on subscription.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_Edge\_Preparation\_Data Get | The consumer is receiving offline edge computing preparation data from the A-ADRF.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_Data\_Storage Request Subscription | The consumer requests A-ADRF to subscribe for data or analytics from ADAE server or A-DCCF for store. This service operation provides parameters needed by the A-ADRF to initiate the subscription (to an ADAE server or A-DCCF).  *Consumer*: ADAE server, A-DCCF  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_Data\_Storage Store Data | The consumer requests A-ADRF to store data or analytics from ADAE server or A-DCCF. Data or analytics are provided to the A-ADRF in the request message.  *Consumer*: ADAE server  *Producer*: A-ADRF |
|  |  | SS\_ADRF\_ ServerToServer\_Analytics Get | The consumer is receiving offline server-to-server analytics/data from A-ADRF.  *Consumer*: ADAES  *Producer*: A-ADRF |
|  |  | SS\_AADRF\_UE RAT connectivity analytics Get | The consumer is receiving offline UE RAT connectivity analytics/data from A-ADRF.  *Consumer*: ADAE server  *Producer*: A-ADRF |
|  |  | SS\_ADCCF\_Data\_Collection Subscribe | The consumer subscribes to receive data or analytics from A-DCCF. The subscription includes service operation specific parameters that identify the data or analytics to be provided.  *Consumer*: ADAE server  *Producer*: A-DCCF |
|  | A-DCCF Data Collection APIs | SS\_ADCCF\_Data\_Collection Notify | The A-DCCF notifies the consumer of the requested data or analytics according to the request or notifies of the availability of previously subscribed data or analytics when data delivery is via the A-DCCF. The A-DCCF may also notify the consumer when data or analytics is to be deleted.  *Consumer*: ADAE server  *Producer*: A-DCCF |
|  |  | SS\_ADCCF\_Data\_Collection Get | The consumer retrieves data or analytics from the A-DCCF.  *Consumer*: ADAE server  *Producer*: A-DCCF |
|  |  | DATA COLLECTION REQUEST | NG-RAN node 1 initiates the procedure by sending the DATA COLLECTION REQUEST message to NG-RAN node 2 to start information reporting or to stop information reporting. Upon receipt, NG-RAN node 2:  shall initiate the requested information reporting according to the parameters given in the request in case the Registration Request for Data Collection IE is set to "start"; or  shall stop all measurements and predictions and terminate the reporting in case the Registration Request for Data Collection IE is set to "stop".  Report Characteristics for Data Collection IE in the DATA COLLECTION REQUEST message indicates the type of objects NG-RAN node 2 performs measurements or predictions on. |
| RAN WG3 TS 38.423 [15] | Data Collection procedures | DATA COLLECTION RESPONSE | If NG-RAN node 2 is capable of providing all of the requested information, it shall initiate the information reporting as requested by NG-RAN node 1 and respond with the DATA COLLECTION RESPONSE message.  If NG-RAN node 2 is capable of providing some but not all of the requested information, it shall initiate the information reporting for the admitted requested information and include the Node Measurement Initiation Result List IE or the Cell Measurement Initiation Result List IE or both in the DATA COLLECTION RESPONSE message. |
|  |  | DATA COLLECTION FAILURE | If none of the requested information can be initiated, NG-RAN node 2 shall send the DATA COLLECTION FAILURE message with an appropriate cause value. |
|  |  | DATA COLLECTION UPDATE | NG-RAN node 2 shall include in the DATA COLLECTION UPDATE message one or more of the following IEs based on the request: SSB Area Radio Resource Status List IE, Predicted Radio Resource Status, Predicted Number of Active UEs, Predicted RRC Connections, Average UE Throughput DL, Average UE Throughput UL, Average Packet Delay, Average Packet Loss, Energy Cost and Measured UE Trajectory. These IEs are specified in Rel. 18 to support three AI/ML for NG-RAN use cases, i.e. Energy Saving, Load Balancing and Mobility Optimization. |

# 7 Overall Evaluation

Editor's note: This clause will provide a general evaluation of potential terminology inconsistency #X and potential feature misalignment #X.

## 7.1 General evaluation on AI/ML related terminology

The study reviewed AI/ML-related activities across TSG SA, TSG RAN, and TSG CT Working Groups, to identify potential misalignment in AI/ML terminology and AI/ML feature descriptions across Working Groups.

The study confirms that AI/ML-related terminology across 3GPP WGs is broadly aligned at the conceptual level but shows differences in emphasis, scope, and granularity. Most variations are intentional and linked to domain requirements and the relevant specific use cases, rather than contradictions. The analysis of AI/ML terminology revealed that several AI/ML terms were overlapping across Working Groups.

A set of unified terminologies has been developed during the study, providing consistent definitions for key concepts such as ML model, ML model training, ML model inference, ML model lifecycle management, Functionality-based lifecycle management, Federated Learning (including Horizontal and Vertical), Transfer Learning.

## 7.2 Detailed evaluation on AI/ML-related terminology

*-* **ML model:** Defined with differences across SA5, SA6, and RAN1 but all converge on a mathematical construct producing outputs from inputs.

*-* **ML model training**: Consistently described as iterative optimisation of parameters, though WG perspectives differ (e.g., lifecycle management in SA5, AI/ML service enablement via AIMLE in SA6, performance-driven framing in RAN1). These variations are complementary.

*-* **ML model distributed training**: Defined only by SA5 as distributing workload across training functions.

*-* **ML model re-training / model update**: SA5 defines re-training as generating a new version without altering structure, while RAN1 and SA6 use update with broader scope (parameter and/or structure). Unified definitions have been proposed to reduce redundancy: Re-training = generating a new version of a previous model. Model update = broader concept covering re-training, parameter adjustment, structural modification, or deployment of a new version.

*-* **ML model testing**: SA5 defines testing as a distinct lifecycle stage; RAN1 considers it a subprocess of training. SA5’s lifecycle distinction provides clearer management separation, while RAN1 highlights operational evaluation during training. Not contradictory but conceptually different.

*-* **ML model pre-specialised training / fine-tuning**: Defined only in SA5. Pre-specialised training produces a task-agnostic model with wide inference scope, while fine-tuning narrows this scope to a new single inference type. This layered paradigm differs from traditional training/re-training and supports modular reuse across domains.

**- Functionality-based lifecycle management:** Defined only in RAN1 and used by RAN2. Signalling procedure where network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signalling (e.g. RRC, MAC-CE, DCI); operates based on, at least, one configuration of AI/ML-enabled Feature / Feature Group or specific configurations of an AI/ML-enabled Feature/FG.

- **Federated learning**: SA5 explicitly contrasts it with distributed training (job-splitting across nodes), clarifying that federated learning is collaborative but privacy-preserving.

- **Reinforcement learning (RL):** Explicitly defined in SA5 (TS 28.105) as part of SupportedLearningTechnology. RL is also mentioned by other WGs inluding SA2 (TR 23.700-84) in the context of NWDAF assisted QoS policy generation, though without a formal definition. The term is consistently understood as a learning paradigm where agents optimise policies through interaction with an environment. No misalignment has been identified, but a baseline reference definition (e.g. in TR 21.905) would help ensure cross-WG consistency.

*-* **ML model activation / de-activation**: SA5 defines activation/deactivation at inference function level (capability enablement), while RAN1 defines them at model level (feature enablement). These are complementary perspectives. Unified definitions have been proposed to distinguish clearly between function-level and model-level activation.

*-* **AI/ML Inference emulation**: Defined only by SA5 as an optional lifecycle step to verify deployment suitability. It is not mandated, and no explicit definitions exist in other WGs. Release 19 introduces enhancements such as environment selection to expand its scope, but adoption remains optional.

- The term "output" is proposed as unified term for e.g. decision or prediction or statistic or recommendation.

## 7.3 Evaluation summary on AI/ML-related terminology

The evaluation of AIML-related terminology yielded the following observation:

- No critical inconsistencies in terminology have been identified that would block cross-domain AI/ML lifecycle management.

- Differences mainly reflect WG scope and perspective, not fundamental conflicts.

## 7.4 Evaluation summary on AI/ML-related features

The analysis of AI/ML‑related features has identified services for AI/ML model training, AI/ML model inference and AI/ML performance are specified across the SA2, SA5 and SA6 Working Groups.

Additionally, analysis of AI/ML‑related features has also identified that services for data collection for AI/ML are specified across the SA2, SA6 and RAN3 Working Groups.

No detailed analysis was conducted on potential misalignment, inconsistencies or overlap between these services.

# 8 Conclusions

Editor's note: This clause will provide information on any potential outcome from clause 5, clause 6 and clause 7 to the respective WGs (according to their Terms of Reference (ToR)) to resolve any issues with appropriate SA-level co-ordination as necessary.

The term "AI/ML" is to be used as a unified definition encompassing "AI/ML", "AI" and "ML" in all corresponding ML related TRs/TSs.

Interim conclusions:

The term "output" is proposed as unified term for e.g. decision or prediction or statistic or recommendation.

The following definitions are proposed for adoption by WGs, and will be documented in a CR to 3GPP TR 21.905:

**- ML model:** A mathematical algorithm that applies AI/ML techniques to generate a set of outputs based on a set of inputs. It may include metadata which consists of, e.g. information related to the model and applicable runtime context.

**- ML model training:** A process to train an ML Model by learning the input/output relationship in a data driven manner and obtain the trained ML Model for e.g. inference.

**- ML model inference:** A process of running a set of inputs through a trained ML model to produce a set of outputs.

**- ML model lifecycle management:** The management capabilities allowing a producer or consumer to manage different phases of the ML model lifecycle as defined in clause 6.2.1.7.

**- Functionality-based lifecycle management:** Signalling procedure where network indicates activation/deactivation/fallback/switching of AI/ML functionality via 3GPP signalling (e.g. RRC, MAC-CE, DCI); operates based on, at least, one configuration of AI/ML-enabled Feature / Feature Group or specific configurations of an AI/ML-enabled Feature/FG.

**- Federated Learning:** A distributed machine learning approach where the ML model(s) are collaboratively trained by multiple participants, including one acting as an FL server and multiple acting as FL clients, iteratively without exchanging data samples.

**- Horizontal Federated Learning:** A federated learning technique without exchanging/sharing local data set, wherein the local data set in different clients for local model training have the same feature space for different samples.

**- Vertical Federated Learning:** A federated learning technique without exchanging/sharing local data, wherein the local data set in different clients for local model training have different feature spaces for the same samples.

**- Transfer Learning:** A machine learning technique where the knowledge acquired from training one or more ML models is leveraged to enhance the performance or accelerate the training of another ML model.

Annex A:  
ML Model

# A.1 ML model life cycle management (LCM)

Rel-18 specification addressed the AI/ML LCM management capabilities (including wide range of use cases, corresponding requirements (stage 1) and solutions (stage 2 NRMs & stage 3 OpenAPIs) for the ML model, including ML model training (which also includes validation), testing, AI/ML inference emulation, deployment and AI/ML inference steps of the lifecycle as shown below for managing the entire lifecycle of the ML model.

## A.1.1 Observations and analyses: AI/ML LCM

- The AI/ML workflow defined by SA WG5 TS 28.105 [9] represents a general framework encapsulating the various life cycle management (LCM) operations for ML model (i.e. model training, testing, emulation, deployment and inference).

- The AI/ML LCM capabilities defined by SA WG5 for each of the operational steps are generic for managing of 3GPP system including the Management and orchestration, CN and RAN domains.

- It is important to recognise that "domain-specific" ML model life cycle related tasks can be developed for the specific domains by the relevant 3GPP WGs, e.g. the RAN WGs can specify data collection within the RAN domain needed to train the UE-side, network-side, or the two-sided UE/network ML models and specific LCM operations for UE-side model over air-interface.

- While ML model and AI/ML inference function life cycle can be specified by the relevant 3GPP WG for the specific domain (i.e. RAN, CN or Management & Orchestration), the "management aspects" of life cycle (i.e. life cycle management) remains to be primarily a "management task" that falls within the responsibility of SA WG5.

- The ML models and the associated "Life Cycle" can be a use case and/or domain specific, the management of the Life Cycle (i.e. LCM) is a higher layer task which is typically a role of the OAM that encompasses the process of e.g. the governance, automation and operational practices applied to the entire AI/ML lifecycle. It is therefore imperative to distinguish the difference between Life Cycle and Life Cycle Management.

- Where feasible, the ML model LCM workflow and associated management capabilities specified by SA WG5 in TS 28.105 [9] could be considered by 3GPP for the currently ongoing and future relevant specification development. The 3GPP WG(s) should potentially provide AI/ML LCM-related requirements, if any, to SA WG5 to avoid duplication and contention of effort.

NOTE: SA WG5 Rel-18 specification in TS 28.105 [9] on ML model LCM and the associated management capabilities does not address the UE-side and UE/Network-side Model LCM.

# A.2 ML model lifecycle management capabilities

Each step in the ML model lifecycle. i.e. the ML model training, ML model testing, AI/ML emulation, ML model deployment and AI/ML inference correspond to number of dedicated management capabilities. The specified capabilities are developed based on corresponding use cases and requirements. The management capabilities specified by SA WG5 TS 28.105 [9] are highlighted below:

## A.2.1 Observations and analyses: ML model lifecycle management capabilities

- ML model lifecycle management (LCM) capabilities are crucial for the effective deployment, operation and optimization of AI/ML-enabled features and capabilities in both the NG-RAN and 5GC. These capabilities ensure that AI/ML models are not only developed and trained correctly but also tested, deployed, evaluated and operated efficiently in the network environment.

- The management capabilities outlined in TS 28.105 [9] offer a structured approach to managing the various steps of the ML model lifecycle. This structured approach is applicable to AI/ML-enabled features and capabilities in NG-RAN, 5GC and management system, ensuring consistency and reliability in the deployment and operation of AI/ML technologies for different domains.

- The AI/ML LCM management capabilities are foundational for integrating advanced AI/ML features into 5G networks. By ensuring that ML models are effectively managed from the training step through to inference, these capabilities provide robust and reliable AI/ML-driven network enhancements.

- The AI/ML LCM workflow and associated management capabilities specified by SA WG5 in TS 28.105 [9] should be considered as the baseline for the AI/ML E2E framework for the 3GPP. These capabilities provide a comprehensive foundation for ensuring that AI/ML models and related processes are consistently managed across all steps of their lifecycle, promoting seamless integration and operation for all domain within the 5G system.

# A.3 AI/ML functionalities management scenarios

The Rel-18 specification TS 28.105 [9] also documented AI/ML functionalities management scenarios in relation with managed AI/ML features which describe the possible locations of ML training function and AI/ML inference function involving the various 3GPP system domains.

## A.3.1 Observations and analyses: AI/ML functionalities management scenarios

- The functional arrangement scenarios defined by SA WG5 specifications demonstrate that different part of the ML model life cycle can be managed depending on the use case.

- The functional arrangements represent management deployment scenarios where for example ML model training related tasks can either be a domain specific or as a cooperative multi-domain task involving for example RAN and management & orchestration or CN and management & orchestration (OAM) domains.

- The LCM workflow defined by SA WG5 serves as a management framework to accommodate and enable all the possible functional arrangement scenarios within or cross-domains in the 3GPP system.

- The functional arrangement scenarios, coupled with the ML model LCM as defined by SA WG5 in TS 28.105 [9], can be considered in the ongoing and any future 3GPP relevant specification development.

Annex B:  
Change history

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Change history | | | | | | | |
| Date | Meeting | TDoc | CR | Rev | Cat | Subject/Comment | New version |
| 2024-09 | TSG SA#105 | SP-241367 | - | - | - | Proposed skeleton agreed for FS\_AIML\_CAL at TSG SA#105 | 0.0.0 |
| 2024-09 | TSG SA#105 | - |  |  |  | Implementing following approved papers: SP-241407, SP-241395, SP-241408, SP-241397, SP-241409, SP-241410. | 0.1.0 |
| 2024-12 | TSG SA#106 | - | - | - | - | Implementing following approved papers: SP-241834, SP-241982, Sp-241839, SP-241983, SP-241965, SP-241984, SP-241985, SP-241986, SP-241987, SP-241988. | 0.2.0 |
| 2025-03 | TSG SA#107 | - | - | - | - | Implementing following approved papers: SP-250404, SP-250405, SP-250406, SP-250299, SP-250346, SP-250407, SP-250349, SP-250408, SP-250409, SP-250410, SP-250411, SP-250412. | 0.3.0 |
| 2025-06 | TSG SA#108 | - | - | - | - | Implementing following approved papers: SP-250578, SP-250725, SP-250753, SP-250827, SP-250828, SP-250829, SP-250843, SP-250844, SP-250846, SP-250847, SP-250885, SP-250886 | 0.4.0 |
| 2025-09 | TSG SA#109 | - | - | - | - | Implementing following approved papers: SP-251219, SP-251221, SP-251163, SP-251164, SP-251257, SP-251258, SP-251275, SP-251120, SP-251223. | 0.5.0 |