**3GPP TSG-SA5 Meeting #157 *S5-245996***

**Hyderabad, India, 14 - 18 October 2024**

**Source: Samsung, AT&T**

**Title: Managing ML models in use in a live network**

**Document for: Approval**

**Agenda Item: 6.19.1**

# 1 Decision/action requested

***In this box give a very clear / short /concise statement of what is wanted.***

# 2 References

None

# 3 Rationale

This provides the solution for the UC “Managing ML models in use in a live network”. The solution requires providing the additional information in the AI/ML Inference Report. This information will specify the potential network impacts due to the inference output result. This information can then enable an authorized consumer to a) take an informed decision about the inference output result b) identify the ML model that is causing a specific performance degradation in the network at some future point of time. The consumer can then decide to either deactivate the inference or update the inference function properties to mitigate the performance degradation.

The basic assumption here is that the recommendation provided as part of inference output result (i.e provided MDA recommendations) may cause some sub-optimal network conditions. The following table shows example of some of the configuration recommendations that various MDA analytics provides and what PM data it may effect.

|  |  |  |
| --- | --- | --- |
| InferenceName/MDT Type (28.104) | Potential Actions | Potential PM Impact |
| CoverageAnalytics.CoverageProblemAnalysis. | Creation of new beam(s), or cell(s);  Change the transmission power of the NR sector carrier;  Delete some unwanted beam(s) or cell(s). | Energy Consumption may increase. |
| MDAAssistedFaultManagement.FailurePrediction. | Update 5GC NF (e.g., AMF and SMF) profile | Updates to servingScope may result in coverage hole. |
| ResourceAnalytics.virtualizedResourceUtilizationAnalysisNF | scale in a list of NFs;  scale out a list of NFs. | In case of scale out Energy Consumption may increase.  In case of scale in, considering the traffic projections, the throughput may decrease. |
| ResourceAnalytics.PhyiscalResourceUtilizationAnalysisNF | optimising the capacity of gNB (e.g., increasing or decreasing physical resources | In case of increasing resources, Energy Consumption may increase. |
| ResourceAnalytics.5GCControlPlaneCongestionAnalysis | scale out a list of 5GC NFs | Energy Consumption may increase. |
| MDAAssistedEnergySaving.EnergySavingAnalysis | For ES on NR cells. It may contain a set of:  - Recommended NR Cell (ES-Cell) to enter energySaving state.  - Recommended candidate cells with precedence for taking over the traffic of the ES-Cell.  - The time to enter and terminate the energy saving state.  - The load threshold to enter and terminate the energy saving state for the ES-Cell.  For ES on UPFs. It contains a set of:  - Recommended UPF (ES-UPF) to conduct energy saving.  - Recommended candidate UPFs with precedence for taking over the traffic of the ES-UPF.  - The time to conduct energy saving for the ES-UPF. | Switching energy saving state ON my reduce Throughput and increase latency. |

# 4 Detailed proposal

|  |
| --- |
| **First Change** |

### 5.5.4 Managing ML models in use in a live network

#### 5.5.4.1 Description

#### 5.5.4.2 Use Cases

##### 5.5.4.2.1 Handling of underperforming ML trained models in live networks

Actions may need to be taken by a network operator once an ML trained model has been identified that is contributing towards non-optimal running of the network. These actions may involve for example, without service interruptions, reverting to running of the network without ML based optimizations or replacing current ML model with an earlier model one that was performing better.

##### 5.5.4.2.2 Performance monitoring of Network Functions with ML trained models in live networks

Several trained ML models maybe in use in an operator network with each one of them influencing network performance. The Network Functions with these ML models need to be monitored to ensure network is running optimally with these models in use. KPIs for evaluating runtime performance of Network Functions using ML models should be provided for this purpose as part of the model and these KPIs would need to be collectable by the network operator.

#### 5.5.4.3 Potential requirements

**REQ-DATA-ACT-1:** The 3GPP management system should have a capability enabling an authorized consumer to identify an ML Model that caused performance measurement and/or KPI degradation.

**REQ-DATA-ACT-2:** The 3GPP management system should have a capability to recommend remedial actions to address performance measurement and/or KPI degradation caused by an ML model.

**REQ-DATA-PERF-1:** The 3GPP management system should be able to collect, from the network, performance data pertaining to Network Functions actively utilizing ML models.

**REQ-DATA-PERF-2:** The 3GPP management system should support performance KPIs that can be monitored for performance assurance purpose.

#### 5.5.4.4 Possible solutions

The solution requires providing the additional information in the AI/ML Inference Report. This information will specify the potential negative network impacts of the execution of the inference output result. This will include impacted network scope (e.g identifier of the network function, geographical location, impact time etc.) and the performance measurements/KPI. This information can then enable an authorized consumer to a) take an informed decision about executing the inference output result b) identify the ML models that is/are causing a specific performance degradation in the network at some future point of time. The consumer can then decide to either deactivate the inference or update the inference function properties to mitigate (i.e stop and then take actions to resolve) the performance degradation.

Note: the proposed solution is specific for the cases where the value of the attribute aIMLInferenceName indicates the values of the MDA type (see 3GPP TS 28.104 [2]).

Editor’s Note: The solution where the value of the attribute aIMLInferenceName does not indicates the values of the MDA type is FFS.

Editor’s Note: The need to have use case specific effected performance data is FFS.

#### 5.5.4.5 Evaluation

|  |
| --- |
| **End of Change** |