**3GPP TSG-SA5 Meeting #158 *S5-246595***

Orlando, USA, 18 - 22 November 2024

**Source: Samsung**

**Title: Solution enhancement and evaluation for ML model in live network.docx**

**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 enhancements and evaluation for an existing solution. Enhancements are regarding better clarity.

# 4 Detailed proposal

|  |
| --- |
| **Next Change** |

#### 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 affected 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]).

The required information will be added as part of InferenceReport as defined in 3GPP TS 28.105.

* AffectedScope: This will specify the scope of affect the inference output may have. This may include
  + Identifier of the network functions that may be affected by the output result of the inference function. This will be in form of a DN.
  + A Geographical location indicating that all the network function in that location may be affected by the inference output result.
  + A time duration indicating that all the related network function may be affected during this time duration by the inference output result.

Existing attribute ManagedActivationScope may be used for this purpose.

* Affected Performance Data: This will identify the potential performance data that may be affected in a non-optimal way due to the recommendations/configurations provided as part of inference output result.
  + PMMetrics: This will identify the performance data or the KPI that may be affected. This will be the name of PM and KPI as defined in 3GPP TS 28.552 and 28.554 respectively.
  + ExpectedPMValues: This will specify the potential non-optimal value of the performance data.



**Figure 1: Procedure for Managing ML models in use in a live network**

1. The ML model gets trained, tested and deployed as per the mechanism defined in 3GPP TS 28.105.
2. Consumer send the request to activate the ML inference at the node where the ML model was deployed.
3. The inference gets initiated.
4. Producer send the acknowledgment.
5. Producer generates the Inference report with the information on the potential network impact as defined in of the embodiment here.
6. Consumer is notified about the availability of the report
7. Consumer send a query request to read the inference report generated by the producer.
8. Producer send the response with the information
9. A performance degradation is detected by the consumer utilizing the existing performance assurance mechanism defined in 3GPP TS 28.622.
10. Consumer checks the historical inference report information to identify the ML model whose inference may be causing the performance degradation based on the information on the potential network impact provided as part of inference report.
11. Once the ML model is identified the consumer deactivate the ML inference.
12. Producer sends the acknowledgment.
13. Alternatively, consumer may choose to update the ML inference
14. Producer send the acknowledgment.

Note: The solution where the value of the attribute aIMLInferenceName does not indicates the values of the MDA type is not included in this version of the document.

The need to have use case specific effected performance data is to be done as part of normative work..

Considering that multiple AI/ML models may be concurrently deployed for generating inferences in the network, more than one AI/ML model may simultaneously impact the same PM(s)KPI(s). Thus, it may be challenging for an authorized consumer to hold an AI/ML model accountable for PM/KPI improvement (or drop). In this case, the solution assumes that multiple AI/ML model will be held responsible for PM/KPI improvement (or drop).

#### 5.5.4.5 Evaluation

The solution described in clause 5.5.4.4 proposes information objects that can enable identification of the ML model(s) that potentially be causing the performance degradation. Enabling consumer to either deactivate the inference or update the inference function properties to mitigate the performance degradation. Therefore, the solution described in clause 5.5.4.4 is a feasible solution before further development in the normative specifications.

Determining the actual impact of ML models on PM(s)/KPI(s) needs to be further investigated during the normative phase.

|  |
| --- |
| **Last Change** |