**Agenda item:** 9.8

**Source:** Interdigital Finland Ory

**Title:**  [FS\_AI4Media] Definitions and AI/ML model composition

**Document for** Discussion andAgreement

# Summary

SA4 has agreed a new study item on AI/ML for Media (FS\_AI4Media) at its SA4 #117-e meeting and a permanent document (S4-220500) for this study was created during the SA4 #118-e meeting.

This document provides considerations about AIML split operations and proposes to add text in section 3 of the Permanent Document. More specifically, we address objective 2 of the FS\_AI4Media study item on identifying model operation configurations.

# Discussion

Annex B of TR 22.874 [1] describes the general principle of split AI/ML operations between AI/ML endpoints different mode for split AI/ML operations between device and network. This is demonstrated in Figure 1 below.



Figure 1 Split AI/ML inference modes over endpoints

We propose to define:

* An AI/ML model composition describing the partitioning of the model,
* The different topologies in use with respect to the different split endpoints (network, device),
* Data exchanged between the different split endpoints, including upstream and downstream intermediate data.

We propose to add some definitions and two sub-clauses, “3.X.1 AI/ML model composition” and “3.X.2 Split model topologies”, that describe the different split model deployments.

# Proposal

## X. Definitions of terms, symbols, and abbreviations

## X.X. Terms

For the purposes of the present document, the terms given in 3GPP 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 3GPP TR 21.905 [1].

**AI/ML model**: A trained AI/ML model.

**AI/ML model subset:** An elementary element of an AI/ML model that can be inferred independently.

**AI/ML model composition:** The composition of an AI/ML Model into one or more AI/ML model subsets.

**AI/ML model split points:** The points in an DNN AI/ML model where it is split into multiple AI/ML model subsets.

**AI/ML inference endpoint:** An AI/ML endpoint that infers an AI/ML model, or a part of it.

**Split AI/ML model**: An AI/ML model composed of AI/ML subsets that is distributed to, and inferred on different inference endpoints.

### 3.X.1 AI/ML model composition

Figure 3.X.1 depicts an AI/ML model composed of different AI/ML subsets based on various split points. Several compositions of the same AI/ML model are represented with AI/ML subsets (M0, M1), (M’0, M’1), or (M “0, M “1, M “2) with split points highlighted in red. The same AI/ML subset may be used in different compositions depending on the configurations of the model composition (e.g. M’0 and M ’00 according to figure 3.X.1).

In this figure, (a) and (b) are examples of AI/ML inference endpoints running an AI/ML model M composed of two subsets M0, M1. A endpoint (network/UE) may run the AI/ML model subset M0 while downloading the other subset M1.

Examples (c) and (d) demonstrate AI/ML split models where M0, M’0 run on the UE while M1, M1’ run on the network respectively.



Figure 3.X.1 AI/ML model composition example

### 3.X.2 Split AI/ML model inference topologies

### 3.X.2.1 UE as media data source

Figure 3.X.2.1 depicts examples of split AI/ML model inference topologies where the UE is the media data source, such as in use-case sub clause 2.1 (object recognition). Assuming that the necessary AI/ML model subsets are already available at each endpoint, figure 3.X.2.1 shows the data exchanged between the different split inference endpoints, including input media data, intermediate data and inference result.

This result can be a textual indication of the recognized object, an output score, a bounding box, enhanced media data, etc.



Figure 3.X.2.1: Split AI/ML model inference where the UE is the media data source

### 3.X.2.2 Provider/network media data source

Figure 3.X.2.2 depicts examples of split model topologies where the network or the AI/ML provider ingests the media data, such as in use-case sub clause 2.2 (video quality enhancement).



Figure 3.X.2.2: Split AI/ML Model inference where the network/ AI/ML service provider ingests the media data

# References

[1] 3GPP TR 22.874, Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS