3GPP TSG-SA WG4 Meeting #131-bis-eS4-250603r03

Online, 11 – 17 April 2025

**Source: Xiaomi**

**Title: [FS\_AI4Media] Pseudo-CR on conclusion related to media data**

**Spec: 3GPP TR 26.927 v1.0.0**

**Agenda item: 9.6**

**Document for: Agreement**

**1. Introduction**

This contribution proposes changes in the conclusions to clarify aspects there were not in scope but not studied.

**2. Reason for Change**

Clarifying the conclusion that traffic characteristics and media interoperability of inference output media data has not been studied.

**3. Motivation**

In TR 26.927, the media data resulting from AI inference is defined in Clause 6.5. This media data is ether generated by the UE or in the network and then sent to the UE.

In the architecture for AI/ML model delivery with inference in the UE (Figure 5.2.2-1) and split inference between the network and UE (Figure 5.2.3-1), the UE generates media data locally and the architecture assumes local consumption, that media data.

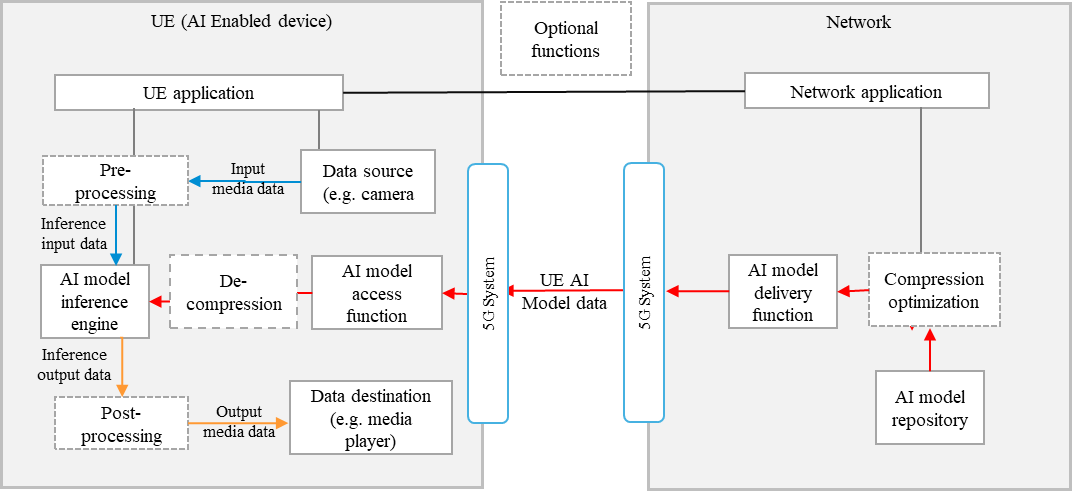


Figure 5.2.2-1: Basic architecture for AI/ML model delivery with inference in the UE

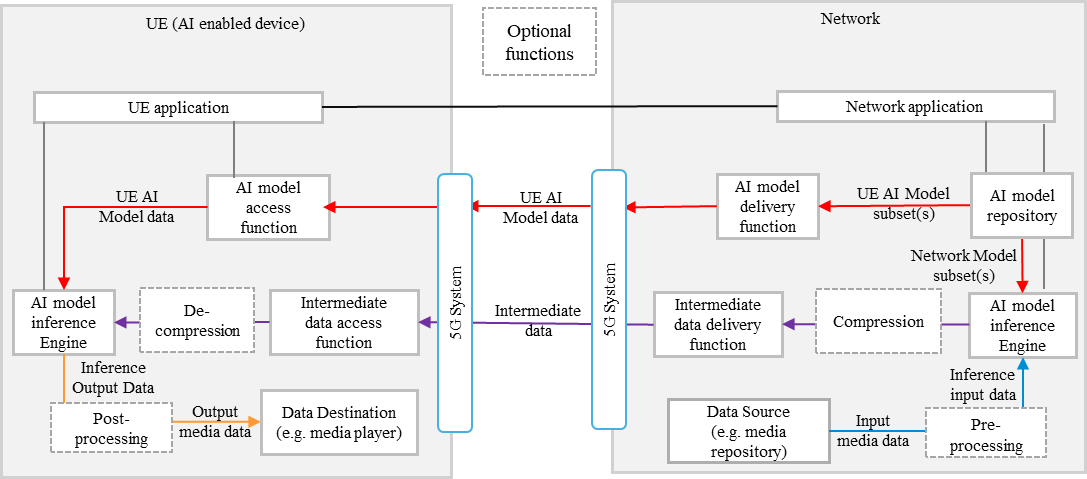


Figure 5.2.3-1: Basic architecture for split inference between the network and UE, with media data source in the network or from the UE via the network

However, it is common today that AI applications allow the user to further share the generated media data to other users (UE to UE scenario) and local-only consumption seems very restrictive. Instead, the receiving UE should be able to interpret the media data generated by the sending UE. Specific media data of AI inference task are image segmentation place, depth map estimation, alpha map estimation, etc.

In the architecture for split inference between the UE and network (Figure 5.2.3-2), the UE receives the generated media data from the network where the AI inference takes place.

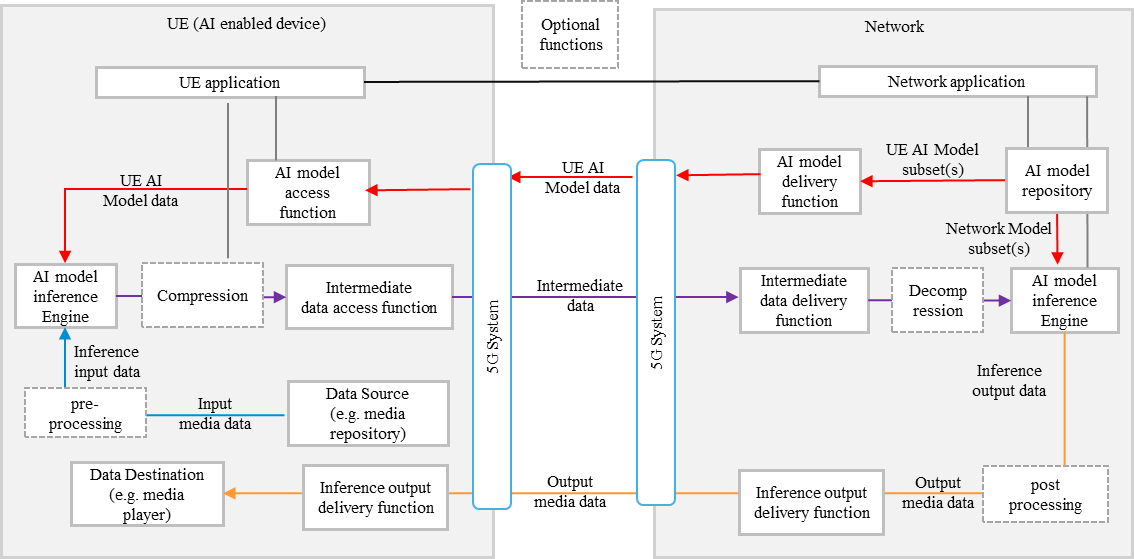


Figure 5.2.3-2: Basic architecture for split inference between the UE and network, with media data source in the UE

In this case, it is also essential that the UE is able to interpret the media data generated by the network. The same example of media data from AI inferences are also here applicable. In addition, knowing the traffic characteristics of such output media data would be relevant. Both aspects were not studied yet in the current study.

As a result, there seems to be a gap for interoperable UE to UE and Network to UE scenario, when it comes to media data generated AI inference in the context of a specific task.

For background, concrete examples of media data generated by AI inference tasks are:

- Image segmentation ,e.g. <https://ai.google.dev/edge/mediapipe/solutions/vision/image_segmenter>

- Depth map estimation, e.g. <https://aihub.qualcomm.com/mobile/models/depth_anything>

**4. Conclusion**

It is proposed to agree the following changes to 3GPP TR 26.927 v1.0.0

\* \* \* First Change \* \* \* \*

# 7 Conclusion

AI/ML in media services involve the use of AI/ML models to perform media processing, typically with video or audio media as the input into an AI/ML model, resulting in an output which may be a version of processed video or audio media (e.g. picture enhancement, audio translation), a specific description of the input media itself (e.g. labelling in object recognition) or a completely new media (e.g. sign/text translation to speech or video). In order to support such AI/ML based media processing, three scenarios have been documented:

- UE device AI inferencing

- AI inferencing in the network

- Split AI inferencing between the UE and the network.

In this study, the broad findings for AI/ML model transfer in TR 22.874 [2] have been further analysed with specific focus on media-based AI/ML use cases and scenarios. In particular this document describes how AI/ML models and data may be distributed over the 5G system and documents the split AI/ML operations between different AI/ML endpoints (noticeably the UE and the network), and the compression of AI/ML model data and intermediate data. Due to the broad range of applications for AI/ML based media processing, as well as the wide diversity of different AI/ML models available for same application, feasibly evaluations for a given set of scenarios are documented in TR 26.847 [20] as part of this study.

Based on the core use cases, functional architectures are presented for basic AI/ML model distribution, split AI/ML operation and distributed/federated learning. Different AI user plane data components have been identified and documented (AI model data, intermediate data, inference input and output data), and a set of logical AI functions have been defined. Regarding inference output media data, the traffic characteristics and media interoperability aspects while potentially relevant aspects to study have not been covered in this study since the focus was primarily on the AI/ML inference process itself.

The identified logical AI functions are further mapped to the 5G system, addressing the underlying 5GMS/RTC and IMS DC architectures. The mapping of such AI media use cases to the different architectures and their relevant procedures describes the provisioning, capability discovery/negotiation and delivery session support for the delivery of AI data components and the use of required AI media functions at different endpoints according to the service configuration negotiated. Architecture variants for three different collaboration scenarios are also introduced, each with a different level of MNO network support for AI/ML functions.

Based on the details in the report, the following next steps are identified:

Normative in release-20:

For collaboration scenario 3 IMS services:

- Recommend stage 3 normative work on the support of AI/ML model distribution and operation in IMS.

- Extend TS 26.114 and TS 26.264 specifications to support AI/ML data delivery and AI/ML media processing in IMS services, as identified in clauses 5.4 and 5.5 respectively.

- Specify support for AI/ML data signalling and negotiation, including support for split.

- Select interoperable formats for AI model data and intermediate data.

- Define the support of the configuration, delivery, compression, and processing of AI data.

* NOTE: Additional conclusions will be added for collaboration scenarios 1 OTT) and 2 (Hosting), scenario 3 non-IMS.

\* \* \* End of Changes \* \* \* \*