(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 February 2024 (15.02.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/031697 A1

(51) International Patent Classification:
H04W 8/22 (2009.01)

(21) International Application Number:
PCT/CN2022/112294

(22) International Filing Date:
12 August 2022 (12.08.2022)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **ZTE CORPORATION** [CN/CN]; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN).

(72) Inventors: **ZHENG, Guozeng**; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN). **LU, Zhaohua**; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN). **LIU, Wenfeng**; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN). **XIAO, Huahua**; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN). **LI, Lun**; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN). **LI, Yong**; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN). **WANG, Yuxin**; ZTE Plaza, Keji Road South, Hi-Tech Industrial Park, Nanshan, Shenzhen, Guangdong 518057 (CN).

(74) Agent: **JIAQUAN IP LAW**; No. 910, Building A, Winner Plaza, No. 100 West Huangpu Avenue, Tianhe District, Guangzhou, Guangdong 510627 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *of inventorship (Rule 4.17(iv))*

Published:
— *with international search report (Art. 21(3))*

(54) **Title:** DEVICE CAPABILITY AND PERFORMANCE MONITORING FOR A MODEL

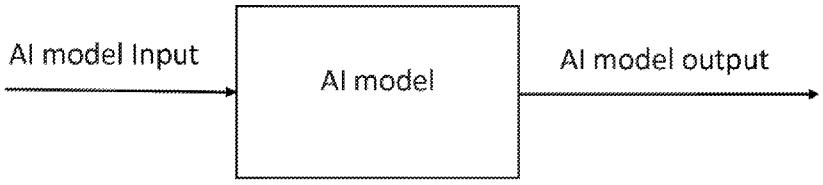

FIG. 1

(57) **Abstract:** Techniques are described for using Artificial intelligence/Machine Learning (AI/ML) models to increase the accuracy of channel state information (CSI). Due to the limited model generalization and dynamic wireless environment, several methods are proposed to monitor the performance of the models. A wireless communication method, comprising transmitting, by a communication device, a capability message indicating that the communication device is capable of using a model by the communication device to perform one or more wireless communication operations.

WO 2024/031697 A1

DEVICE CAPABILITY AND PERFORMANCE MONITORING FOR A MODEL

TECHNICAL FIELD

[0001] This disclosure is directed generally to digital wireless communications.

BACKGROUND

[0002] Mobile telecommunication technologies are moving the world toward an increasingly connected and networked society. In comparison with the existing wireless networks, next generation systems and wireless communication techniques will need to support a much wider range of use-case characteristics and provide a more complex and sophisticated range of access requirements and flexibilities.

[0003] Long-Term Evolution (LTE) is a standard for wireless communication for mobile devices and data terminals developed by 3rd Generation Partnership Project (3GPP). LTE Advanced (LTE-A) is a wireless communication standard that enhances the LTE standard. The 5th generation of wireless system, known as 5G, advances the LTE and LTE-A wireless standards and is committed to supporting higher data-rates, large number of connections, ultra-low latency, high reliability and other emerging business needs.

SUMMARY

[0004] Techniques are disclosed for report UE capabilities involving AI/ML models. Several solutions are discussed to monitor the model performance in some embodiments.

[0005] A wireless communication method, comprising transmitting, by a communication device, a capability message indicating that the communication device is capable of using a model by the communication device to perform one or more wireless communication operations.

[0006] In some embodiments, the capability message includes time requirements of the model. In some embodiments, the time requirements are associated with a model identifier.

[0007] In some embodiments, the model can be received by the communication device from a network.

[0008] In some embodiments, the capability message include one of the following: a computation time, a computation time offset, or a model activation time. In some embodiments, the computation time is a time the communication device needs to conduct an operation, wherein the computation time offset is an additional time the communication device needs to conduct an operation, wherein the model activation time is a time the communication device needs to activate an operation.

[0009] In some embodiments, the computation time can be a constant independent of the model selected. In some embodiments, the value of the computation time offset depends on the model. In some embodiments, the capacity message indicates a measurement report is created after a time lapse T based on a reference signal received by the communication device, wherein T depends on the computation time offset and the computation time.

[0010] In some embodiments, the method further comprising transmitting, by the communication device, an identifier message indicating the model is activated.

[0011] In some embodiments, the identifier message can be sent through a PUCCH. In some embodiments.

[0012] In some embodiments, the method further comprising receiving, by the communication device, a response indicating the identifier message was received successfully.

[0013] In some embodiments, the method further comprising deactivating, by the communication device, the model when the response was not received by the communication device.

[0014] In some embodiments, the capability message includes a capacity of the communication device, wherein the capacity includes concurrent capability for model-related and non-model related.

[0015] Another wireless communication method, comprising reporting, by a communication device, a performance indication of a model.

[0016] In some embodiments the performance indication includes a relationship between an assistance information with an actual measurement information conducted by the communication device.

[0017] In some embodiments, the relationship can be a similarity or a distance, wherein the similarity includes cosine similarity, generalization cosine similarity or square generalization cosine similarity, wherein the distance includes Euclidean distance or normalized mean square error.

[0018] In some embodiments, the communication device is configured with at least a dedicated occasion used for the performance indication of the model. In some embodiments, the performance indication includes a predicated measurement and an actual measurement of the dedicated occasion.

[0019] In some embodiments, the actual measurement includes RSRP(s) of a plurality of reference signal, wherein the predicted measurement includes predicted RSRP(s) of the reference signals based on the model.

[0020] In some embodiments, the actual measurement includes codebook-based precoding matrix indication (PMI).

[0021] In some embodiments, the performance indication further includes a relationship between the output of the model and an input of the model.

[0022] In some embodiments, the performance indication includes a relationship between M-N remaining measurement results of N measurement results and a part of the model output, wherein the N measurement results are based on the reference signals, wherein the M measurement results are based on the N measurement results and are used for an input of the model, wherein $M < N$.

[0023] In some embodiments, the relationship can be a similarity or a distance, wherein the similarity can be at least one of the following: cosine similarity, generalization cosine similarity or square generalization cosine similarity, wherein the distance can be at least one of the following: Euclidean distance or normalized mean square error.

[0024] An In yet another exemplary aspect, the above-described methods are embodied in the form of processor-executable code and stored in a non-transitory computer-readable storage medium. The code included in the computer readable storage medium when executed by a processor, causes the processor to implement the methods described in this patent document.

[0025] In yet another exemplary embodiment, a device that is configured or operable to perform the above-described methods is disclosed.

[0026] The above and other aspects and their implementations are described in greater detail in the drawings, the descriptions, and the claims.

BRIEF DESCRIPTION OF THE DRAWING

[0027] FIG. 1 shows an example of Artificial Intelligence/Machine Learning (AI/ML) model.

[0028] FIG. 2 shows an example of computation time and computation time offset.

[0029] FIG. 3 shows an example the relationship between the assistance channel information and the actual channel measurement.

[0030] FIG. 4 shows an example of indicating a dedicated channel measurement occasion to monitor the model performance.

[0031] FIG. 5 shows an example of adopting the relationship between the part of AI model output and the AI model input for model performance monitoring.

[0032] FIG. 6 shows an exemplary block diagram of a hardware platform that may be a part of a network device or a communication device.

[0033] FIG. 7 shows an example of wireless communication including a base station (BS) and user equipment (UE) based on some implementations of the disclosed technology.

INTRODUCTION

[0034] Section headings are used in the present document only to improve readability and do not limit scope of the disclosed embodiments and techniques in each section to only that section. Furthermore, some embodiments are described with reference to Third Generation Partnership Project (3GPP) New Radio (NR) standard ("5G") for ease of understanding and the described

technology may be implemented in different wireless system that implement protocols other than the 5G protocol. In addition, an AI/ML model is an exemplary scenario, and the technical solutions described in this application can be generalized or applicable to any model that determines a relationship between an input and an output.

[0035] **Initial Comments.**

[0036] Artificial Intelligence/Machine Learning (AI/ML) has been studied and applied in various fields.

[0037] Some studies in Artificial Intelligence and Machine Learning are conducted to improve the efficiency of wireless communication system, especially in physical layer. For example, an AI/ML model can be used to increase the accuracy of channel state information (CSI). Additionally, AI/ML models can predict beam information. To support channel measurement by using AI/ML models, UEs in the network should have corresponding capabilities.

[0038] This application discloses some procedures are proposed to report UE capabilities involving AI/ML models. Due to the limited model generalization and dynamic environments, the model may degrade its performance over time. Accordingly, in this application, several solutions are discussed to monitor the model performance.

[0039] **I. Introduction**

[0040] AI/ML has been studied and used in various fields to extract features that cannot be derived by other mathematical methods. In general, AI/ML model is a data driven algorithm that applies AI/ML techniques to generate a set of outputs based on a set of inputs, which includes at least three parts as shown in FIG. 1:

- AI model input: the data fed into an AI/ML model,
- AI model output: the output of an AI/ML model, and
- AI model: an algorithm to derive the relationship between AI model input and AI model output

[0041] To facilitate the discussion, the following terminologies are introduced with some general descriptions.

[0042] AI model training is a process to train an AI/ML model by learning the input/output relationship in a data driven manner and to obtain the trained AI/ML model for inference.

[0043] AI model Inference is a process of adopting a trained AI/ML model to produce a set of outputs based on a set of inputs.

[0044] In a measurement report, e.g., a CSI report or beam report). A UE may use an AI model to derive the measurement report. To support channel measurement by using AI/ML models, UE should have corresponding capabilities.

[0045] In this application, some procedures are proposed to report UE capabilities involving AI/ML models. Furthermore, due to the limited model generalization and dynamic environments, the model may degrade its performance over time. In other words, the data used for AI model training and that for AI model inference may be quite different, causing the AI model unable to get the expected AI model output. In this application, several solutions are discussed to monitor the model performance.

[0046] **II. Example Embodiments**

[0047] **A. Capability signaling or assistance signaling**

[0048] In some embodiments, a UE may report its model (e.g., AI/ML model) related information to a network through a UE capability signaling or an assistance information signaling.

[0049] In some embodiments, the model related information can be reported after a network sends a request message to a UE, where the request message may include a model identifier.

[0050] In some embodiments, model related information can be reported without firstly requesting a message from the network.

[0051] In some embodiments, the model related information includes time requirements for a model.

[0052] In some embodiments, the model related information may include a model identifier to associate with the corresponding time requirements.

[0053] The time requirements may include one of the following:

[0054] **1) Computation time**

[0055] In an example, the computation time may include the time that a UE needs to conduct/operate/execute a model and/or the time a UE needs to prepare a measurement report associated with the model.

[0056] In some embodiments, the computation time is the same to all models.

[0057] In some embodiments, the computation time is different for different report quantities in a measurement report. In other words, the computation time depends on the content included in the measurement report. For example, the computation time for beam report and that for CSI report may be different.

[0058] In some embodiments, the computation time is not required to be reported by a UE. The computation time reuses the value defined for non-model based measurement report.

[0059] **2) Computation time offset**

[0060] In an example, computation time offset may include the additional time that a UE needs to conduct/operate/execute for a model.

[0061] In some embodiments, each model has its own computation time offset.

[0062] In some embodiments, the computation time offset is not required to be reported by a UE. In this scenario, the value of computation time offset is set zero by default. Alternatively, a UE can report a zero value for the computation time offset.

[0063] In some embodiments, as shown in FIG. 2, a UE may only report a measurement report based on reference signal(s) that is received T time units earlier than the time when the measurement report is transmitted, where the T time units are computation time plus computation time offset.

[0064] **3) Model activation/application time**

[0065] In some embodiments, when a UE receives a Media Access Control (MAC) signaling including an indication command to use a model, the UE may need some time to activate the model. For example, when a UE transmits a PUCCH with HARQ-ACK information in time n corresponding to the PDSCH carrying the indication command, the UE assumes that the corresponding model to be ready/activated for inference should be applied at time n + model activation time.

[0066] In some embodiments, when a UE receives a Downlink Control Indication (DCI) including an indication command to use a model, UE may need some time to activate the model. For example, when a UE receives a PDCCH carrying an indication command in time n, the UE assumes that the corresponding model to be ready/activated for inference should be applied at time n + model activation time.

[0067] In some embodiments, a UE can send a signaling to inform a network that a model is ready/activated for inference. In some embodiments, if a UE can send a signaling to inform a network that a model is ready/activated for inference and if the UE is indicated to use this model to do operations, the UE doesn't need extra time (e.g., activation time) to activate this model.

[0068] In some embodiments, the signaling can be transmitted by a PUCCH, where each model may be associated with a dedicated PUCCH resource.

[0069] In some embodiments, the signaling can be transmitted by a PRACH, where each model may be associated with a dedicated PRACH.

[0070] In some embodiments, a UE can receive the response from a network to confirm that the signaling has been received successfully by the network. In some embodiments, the response is transmitted by a MAC signaling. In some embodiments, if a UE detects a PDCCH scrambled by a dedicated RNTI, the UE is confirmed that the signaling has been received successfully by the network.

[0071] In some embodiments, if a UE doesn't receive the response, the UE may deactivate the model automatically.

[0072] In some embodiments, a UE may deactivate the model automatically when the waiting time, starting from the UE sending the signaling, exceeds a time threshold.

[0073] In some embodiments, a UE may report concurrent/mixed UE capabilities for model related and non-model related.

[0074] In some embodiments, a UE may report that UE can be configured with the number of measurement report(s) using model and the number of measurement report(s) not using model.

[0075] For example, a UE may report that when the UE is configured with M CSI reports using a model in a Bandwidth Part/Component Carrier (BWP/CC), the UE is not expected to be configured more than N other CSI reports not using model. Here, the value of N is either not required to be reported or can be reported by a value zero, both of which indicate that when configured with CSI reports using a model, the UE cannot be configured with other CSI reports not using model.

[0076] In another example, a UE may report that when the UE is configured with M beam reports using a model in a BWP/CC (Bandwidth Part/Component Carrier), the UE is not expected to be configured more than N other beam reports not using model. Here, the value of N

is not required to be reported or can be reported by a value zero, both of which indicate that when configured with beam reports using a model, the UE cannot be configured with other beam reports not using model.

[0077] In some embodiments, a UE may report that the UE can process simultaneously for measurement report(s) using model and measurement report(s) not using model.

[0078] For example, a UE can process simultaneously for M beam reports using model and N beam reports not using model in a CC (or across all CCs). The value of N is not required to be reported or can be reported by a value zero.

[0079] For example, UE can process simultaneously for M beam report using model and N CSI reports not using model in a CC (or across all CCs). The value of N is not required to be reported or can be reported by a value zero.

[0080] **B. Model monitoring procedure**

[0081] At the UE side, the performance of a deployed model may not always be good. Due to the limited model generalization and dynamic environment, the performance of the model may degrade. Therefore, both the network and the UE should constantly monitor the performance of the model.

[0082] Here we propose three solutions to conduct the model performance monitoring.

[0083] **Method 1: Network provides assistance channel information to assist model performance monitoring.**

[0084] In some embodiments, a UE may report the relationship between the assistance channel information and the actual channel measurement.

[0085] In an example shown in FIG. 3, an assistance channel information and an actual channel measurement are denoted by H_{assist} and H respectively. The first AI model output is $\widehat{H}_{assist}$ and its AI model input is based on assistance channel information. The second AI model output is $\widehat{H}$ and its AI model input is based on actual channel measurement

[0086] One metric to evaluate the model performance is similarity, e.g., a cosine similarity, a generalization cosine similarity, or a square generalization cosine similarity. Assume the similarity between H_{assist} and H is S_{input} and the similarity between $\widehat{H}_{assist}$ and $\widehat{H}$ is S_{output}. In an example, if the value of S_{input} equals to (or approximately equals to) that of S_{output}, the AI model still works well with some guaranteed performance.

[0087] Another metric to evaluate the model performance is distance, e.g., Euclidean distance or normalized mean squared error. Assume the distance between H_{assist} and H is D_{input} and the distance between $\widehat{H}_{assist}$and $\widehat{H}$ is D_{output}. In an example, if the value of D_{input} equals to (or approximately equals to) that of D_{output}, the AI model still works well with some guaranteed performance.

[0088] In some embodiments, UE needs to report a performance indication to network about the result of performance monitoring.

[0089] In some embodiments, the performance indication includes an indicator in a measurement report.

For example, the value of 0 and 1 mean the AI model is invalid and valid respectively. More specifically, the value 0 means S_{output}/S_{input} or D_{output}/D_{input} is lower than a threshold.

[0090] In some embodiments, the performance indication includes the value of similarity (S_{input} and S_{output}) or a distance (D_{input} and D_{output}) in a measurement report. In some embodiments, the performance indication includes the value of S_{output}/S_{input} or D_{output}/D_{input}.

[0091] In some embodiments, network may provide a bunch of assistance channel information. In this case, each assistance channel information may be associated with an performance indication.

[0092] Method 2: Dedicated channel measurement occasion is used to monitor the model performance.

[0093] In some embodiments, a network may indicate at least one dedicated channel measurement occasion to monitor the model performance.

[0094] In one example where an AI model is adopted for channel prediction, the AI model input is based on actual channel measurements in some previous occasions (e.g., observation occasions). The AI model output is the predicted channel measurements that can be used in some future occasions (e.g., predication occasions). Therefore, a UE doesn't have to acquire actual channel measurements of prediction occasions. However, if an occasion is indicated as a dedicated channel measurement occasion (e.g., monitoring occasion) to monitor the model performance, as shown in FIG. 4, a UE may need to acquire both actual and predicated channel

measurements. Then, a UE/network can compare actual channel measurement and predicated channel measurement to check whether the channel prediction is accurate enough.

[0095] In some embodiment, a UE has to report both actual channel measurement and predicated channel measurement of the monitoring occasion in a measurement report.

[0096] In one example, an AI model is adopted for beam prediction.

[0097] In some embodiments, the actual channel measurement includes RSRP(s) of a plurality of reference signals, and the predicated channel measurement includes RSRP(s) of the plurality of reference signals.

[0098] In some embodiments, a network may indicate the numbers of RSRP(s) included in actual and predicated channel measurement

[0099] In some embodiments, the number of RSRP(s) included in actual channel measurement and the that included in predicated channel measurement are the same.

[0100] In some embodiments, each RSRP is associated with a reference signal index.

[0101] In another example, the AI model can be adopted for CSI prediction.

[0102] In some embodiments, the actual channel measurement includes a traditional codebook-based precoding matrix indication (PMI).

[0103] **Method 3: Dedicated module of an AI module is for model performance monitoring**

[0104] In some embodiments, a part of AI model output can be used for model performance monitoring.

[0105] In some embodiments, the part of AI model output at least includes an indicator showing whether the AI model is valid or not. The indicator can be reported in a measurement report.

[0106] In some embodiments, the relationship between the part of AI model output and the AI model input can be used for model performance monitoring.

[0107] In an example as shown in FIG. 5, the AI model input is channel information H, the AI model output of first sub-module is the compressed channel information $\widehat{H}_1$ and the AI model input of second sub-module is the reconstructed channel information $\widehat{H}_2$.

[0108] In some embodiments, a measurement report may include an indication of the relationship.

[0109] In some embodiments, the relationship can be the similarity/distance between the channel information and the reconstructed channel information.

[0110] In some embodiments, a measurement report may include both an indication to the relationship and the compressed channel information.

[0111] In some embodiments, a UE gets N measurement results based on the received reference signal(s). An AI model input may include only M measurement results of the N measurement results with M<N. In this scenario, the remaining N-M measurement results can be used for model performance monitoring.

[0112] In some embodiments, the N measurement results are N RSRP values, where each RSRP value corresponds to a beam between a network and a UE. An AI model input only includes M RSRP values out of the N RSRP values with M<N. Accordingly, the remaining N-M RSRP values can be used for model performance monitoring. The remaining N-M RSRP values correspond to a set of beams between a network and a UE. Furthermore, the part of AI model output includes the predicted N-M RSRP values correspond to the set of beams between a network and a UE.

[0113] In some embodiments, a UE may report a validity indicator based on the remaining N-M RSRP values with the predicted N-M RSRP values.

[0114] In some embodiments, the validity indicator can be the similarity between the remaining N-M RSRP values and the predicted N-M RSRP values.

[0115] In some embodiments, the validity indicator can be the difference (or distance) between the remaining N-M RSRP values and the predicted N-M RSRP values.

[0116] In some embodiments, a UE may report both the remaining N-M RSRP values and the predicted N-M RSRP values in a measurement report.

[0117] In some embodiments, one of the predicted N-M RSRP values is reported relative to one of the remaining N-M RSRP values corresponding the same beam between a network and a UE.

[0118] In some embodiments, the N measurement results are N channel measurement results corresponding to a N-port reference signal. An AI model input only includes M (M<N) channel measurement results out of the N channel measurement results. Accordingly, the remaining N-M channel measurement results can be used for model performance monitoring. The remaining N-M channel measurement results correspond to a set of ports of the N-port reference signal.

Furthermore, the part of AI model output includes the predicted N-M channel measurement results correspond to the set of ports of the N-port reference signal.

[0119] In some embodiments, a UE may report a validity indicator based on the remaining N-M channel measurement results and the predicted N-M channel measurement results.

[0120] In some embodiments, the validity indicator can be the similarity between the remaining N-M channel measurement results and the predicted N-M channel measurement results.

[0121] In some embodiments, the validity indicator can be the difference (or distance) between the remaining N-M channel measurement results and the predicted N-M channel measurement results.

[0122] FIG. 6 shows an exemplary block diagram of a hardware platform 600 that may be a part of a network device (e.g., base station) or a communication device (e.g., a user equipment (UE)). The hardware platform 600 includes at least one processor 610 and a memory 605 having instructions stored thereupon. The instructions upon execution by the processor 610 configure the hardware platform 600 to perform the operations described in FIGS. 1 to 5 and 7 and in the various embodiments described in this patent document. The transmitter 615 transmits or sends information or data to another device. For example, a network device transmitter can send a message to a user equipment. The receiver 620 receives information or data transmitted or sent by another device. For example, a user equipment can receive a message from a network device.

[0123] The implementations as discussed above will apply to a wireless communication. FIG. 7 shows an example of a wireless communication system (e.g., a 5G or NR cellular network) that includes a base station 720 and one or more user equipment (UE) 711, 712 and 713. In some embodiments, the UEs access the BS (e.g., the network) using a communication link to the network (sometimes called uplink direction, as depicted by dashed arrows 731, 732, 733), which then enables subsequent communication (e.g., shown in the direction from the network to the UEs, sometimes called downlink direction, shown by arrows 741, 742, 743) from the BS to the UEs. In some embodiments, the BS send information to the UEs (sometimes called downlink direction, as depicted by arrows 741, 742, 743), which then enables subsequent communication (e.g., shown in the direction from the UEs to the BS, sometimes called uplink direction, shown by dashed arrows 731, 732, 733) from the UEs to the BS. The UE may be, for

example, a smartphone, a tablet, a mobile computer, a machine to machine (M2M) device, an Internet of Things (IoT) device, and so on.

[0124] A wireless communication method, comprising transmitting, by a communication device, a capability message indicating that the communication device is capable of using a model by the communication device to perform one or more wireless communication operations.

[0125] In some embodiments, the capability message includes time requirements of the model. In some embodiments, the time requirements are associated with a model identifier.

[0126] In some embodiments, the model can be received by the communication device from a network.

[0127] In some embodiments, the capability message include one of the following: a computation time, a computation time offset, or a model activation time. In some embodiments, the computation time is a time the communication device needs to conduct an operation, wherein the computation time offset is an additional time the communication device needs to conduct an operation, wherein the model activation time is a time the communication device needs to activate an operation.

[0128] In some embodiments, the computation time can be a constant independent of the model selected. In some embodiments, the value of the computation time offset depends on the model. In some embodiments, the capacity message indicates a measurement report is created after a time lapse T based on a reference signal received by the communication device, wherein T depends on the computation time offset and the computation time.

[0129] In some embodiments, the method further comprising transmitting, by the communication device, an identifier message indicating whether the model is activated.

[0130] In some embodiments, the identifier message can be sent through a PUCCH. In some embodiments.

[0131] In some embodiments, the method further comprising receiving, by the communication device, a response indicating the identifier message was received successfully.

[0132] In some embodiments, the method further comprising deactivating, by the communication device, the model when the response was not received by the communication device.

[0133] In some embodiments, the capability message includes a capacity of the communication device, wherein the capacity includes concurrent capability for model-related and non-model related.

[0134] Another wireless communication method, comprising reporting, by a communication device, a performance indication of a model.

[0135] In some embodiments the performance indication includes a relationship between an assistance information with an actual measurement information conducted by the communication device.

[0136] In some embodiments, the relationship can be a similarity or a distance, wherein the similarity includes cosine similarity, generalization cosine similarity or square generalization cosine similarity, wherein the distance includes Euclidean distance or normalized mean square error.

[0137] In some embodiments, the communication device is configured with at least a dedicated occasion used for the performance indication of the model. In some embodiments, the performance indication includes a predicated measurement and an actual measurement of the dedicated occasion.

[0138] In some embodiments, the actual measurement includes RSRP(s) of a plurality of reference signal, wherein the predicted measurement includes predicted RSRP(s) of the reference signals based on the model.

[0139] In some embodiments, the actual measurement includes codebook-based precoding matrix indication (PMI).

[0140] In some embodiments, the performance indication further includes a relationship between the output of the model and an input of the model.

[0141] In some embodiments, the performance indication includes a relationship between M-N remaining measurement results of N measurement results and a part of the model output, wherein the N measurement results are based on the reference signals, wherein the M measurement results are based on the N measurement results and are used for an input of the model, wherein $M < N$.

[0142] In some embodiments, the relationship can be a similarity or a distance, wherein the similarity can be at least one of the following: cosine similarity, generalization cosine similarity or square generalization cosine similarity, wherein the distance can be at least one of the following: Euclidean distance or normalized mean square error.

[0143] An In yet another exemplary aspect, the above-described methods are embodied in the form of processor-executable code and stored in a non-transitory computer-readable storage medium. The code included in the computer readable storage medium when executed by a processor, causes the processor to implement the methods described in this patent document.

[0144] In yet another exemplary embodiment, a device that is configured or operable to perform the above-described methods is disclosed.

[0145] The above and other aspects and their implementations are described in greater detail in the drawings, the descriptions, and the claims.

[0146] In this document the term "exemplary" is used to mean "an example of" and, unless otherwise stated, does not imply an ideal or a preferred embodiment.

[0147] Some of the embodiments described herein are described in the general context of methods or processes, which may be implemented in one embodiment by a computer program product, embodied in a computer-readable medium, including computer-executable instructions, such as program code, executed by computers in networked environments. A computer-readable medium may include removable and non-removable storage devices including, but not limited to, Read Only Memory (ROM), Random Access Memory (RAM), compact discs (CDs), digital versatile discs (DVD), etc. Therefore, the computer-readable media can include a non-transitory storage media. Generally, program modules may include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract

data types. Computer- or processor-executable instructions, associated data structures, and program modules represent examples of program code for executing steps of the methods disclosed herein. The particular sequence of such executable instructions or associated data structures represents examples of corresponding acts for implementing the functions described in such steps or processes.

[0148] Some of the disclosed embodiments can be implemented as devices or modules using hardware circuits, software, or combinations thereof. For example, a hardware circuit implementation can include discrete analog and/or digital components that are, for example, integrated as part of a printed circuit board. Alternatively, or additionally, the disclosed components or modules can be implemented as an Application Specific Integrated Circuit (ASIC) and/or as a Field Programmable Gate Array (FPGA) device. Some implementations may additionally or alternatively include a digital signal processor (DSP) that is a specialized microprocessor with an architecture optimized for the operational needs of digital signal processing associated with the disclosed functionalities of this application. Similarly, the various components or sub-components within each module may be implemented in software, hardware or firmware. The connectivity between the modules and/or components within the modules may be provided using any one of the connectivity methods and media that is known in the art, including, but not limited to, communications over the Internet, wired, or wireless networks using the appropriate protocols.

[0149] While this document contains many specifics, these should not be construed as limitations on the scope of an invention that is claimed or of what may be claimed, but rather as descriptions of features specific to particular embodiments. Certain features that are described in this document in the context of separate embodiments can also be implemented in combination in a single embodiment. Conversely, various features that are described in the context of a single embodiment can also be implemented in multiple embodiments separately or in any suitable sub-combination. Moreover, although features may be described above as acting in certain combinations and even initially claimed as such, one or more features from a claimed combination can in some cases be excised from the combination, and the claimed combination may be directed to a sub-combination or a variation of a sub-combination. Similarly, while operations are depicted in the drawings in a particular order, this should not be understood as

requiring that such operations be performed in the particular order shown or in sequential order, or that all illustrated operations be performed, to achieve desirable results.

[0150] Only a few implementations and examples are described and other implementations, enhancements and variations can be made based on what is described and illustrated in this disclosure.

CLAIMS

What is claimed is:

1. A wireless communication method, comprising:

transmitting, by a communication device, a capability message indicating that the communication device is capable of using a model by the communication device to perform one or more wireless communication operations.

2. The method of claim 1, wherein the capability message includes time requirements of the model.

3. The method of claim 2, wherein the time requirements are associated with a model identifier.

4. The method of claim 3, wherein the model can be received by the communication device from a network.

5. The method of claim 2, wherein the capability message include one of the following: a computation time, a computation time offset, or a model activation time,

wherein the computation time is a time the communication device needs to conduct an operation, wherein the computation time offset is an additional time the communication device needs to conduct an operation, wherein the model activation time is a time the communication device needs to activate an operation.

6. The method of claim 5, wherein the computation time can be a constant independent of the model selected.

7. The method of claim 5, wherein the value of the computation time offset depends on the model.

8. The method of claim 5, wherein the capacity message indicates a measurement report is created after a time lapse T based on a reference signal received by the communication device, wherein T depends on the computation time offset and the computation time.

9. The method of claim 1 further comprising transmitting, by the communication device, an identifier message indicating that the model is activated.

10. The method of claim 9, wherein the identifier message can be sent through a PUCCH.

11. The method of claim 9 further comprising receiving, by the communication device, a response indicating the identifier message was received successfully.

12. The method of claim 11 further comprising deactivating, by the communication device, the model when the response was not received by the communication device.

13. The method of claim 1, wherein the capability message includes a capacity of the communication device, wherein the capacity includes concurrent capability for model-related and non-model related.

14. A wireless communication method, comprising:
reporting, by a communication device, a performance indication of a model.

15. The method of claim 14, wherein the performance indication includes a relationship between an assistance information and an actual measurement information conducted by the communication device.

16. The method of claim 15, wherein the relationship can be a similarity or a distance, wherein the similarity includes cosine similarity, generalization cosine similarity or

square generalization cosine similarity, wherein the distance includes Euclidean distance or normalized mean square error.

17. The method of claim 14, wherein the communication device is configured with at least a dedicated occasion used for the performance indication of the model.

18. The method of claim 14, wherein the performance indication includes a predicated measurement and an actual measurement of the dedicated occasion.

19. The method of claim 18, wherein the actual measurement includes RSRP(s) of a plurality of reference signal, wherein the predicted measurement includes predicted RSRP(s) of the reference signals based on the model.

20. The method of claim 18, wherein the actual measurement includes codebook-based precoding matrix indication (PMI).

21. The method of claim 14, wherein the performance indication further includes a relationship between the output of the model and an input of the model.

22. The method of 14, wherein the performance indication includes a relationship between M-N remaining measurement results of N measurement results and a part of the model output, wherein the N measurement results are based on the reference signals, wherein the M measurement results are based on the N measurement results and are used for an input of the model, wherein $M < N$.

23. The method of claim 21 or 22, wherein the relationship can be a similarity or a distance, wherein the similarity can be at least one of the following: cosine similarity, generalization cosine similarity or square generalization cosine similarity, wherein the distance can be at least one of the following: Euclidean distance or normalized mean square error.

24. An apparatus for wireless communication comprising a processor, configured to implement a method recited in one or more of claims 1 to 24.

25. A non-transitory computer readable program storage medium having code stored thereon, the code, when executed by a processor, causing the processor to implement a method recited in one or more of claims 1 to 24.

FIG. 1

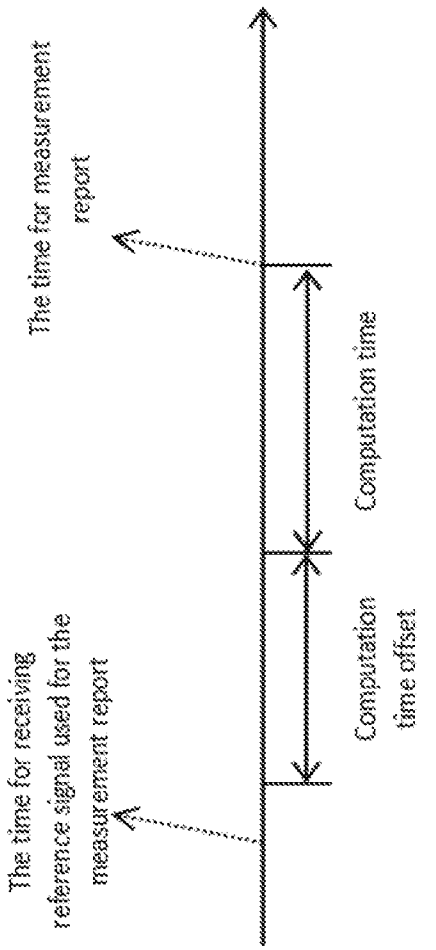
The time for measurement report
Computation time
Computation time offset
The time for receiving reference signal used for the measurement report

FIG. 2

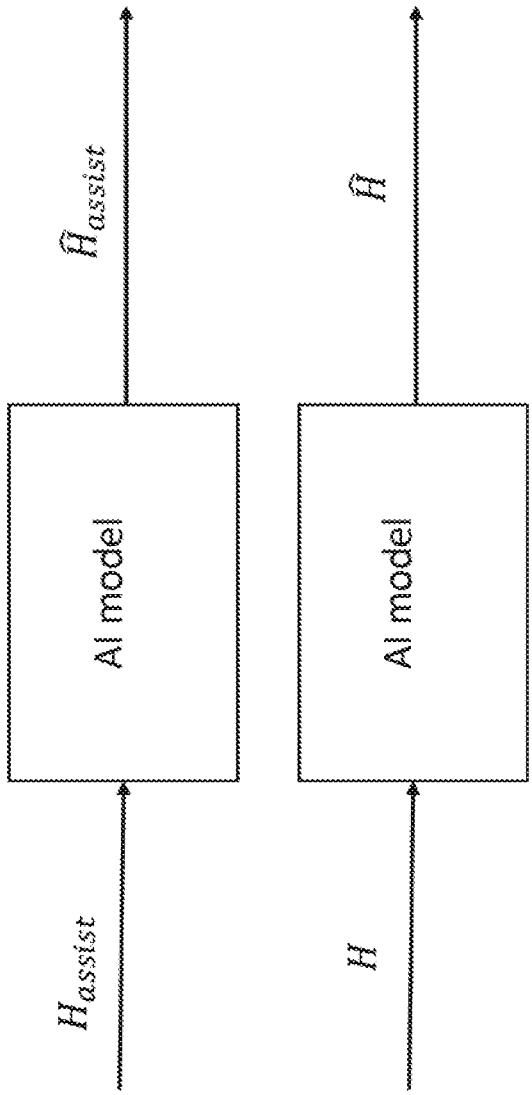

FIG. 3

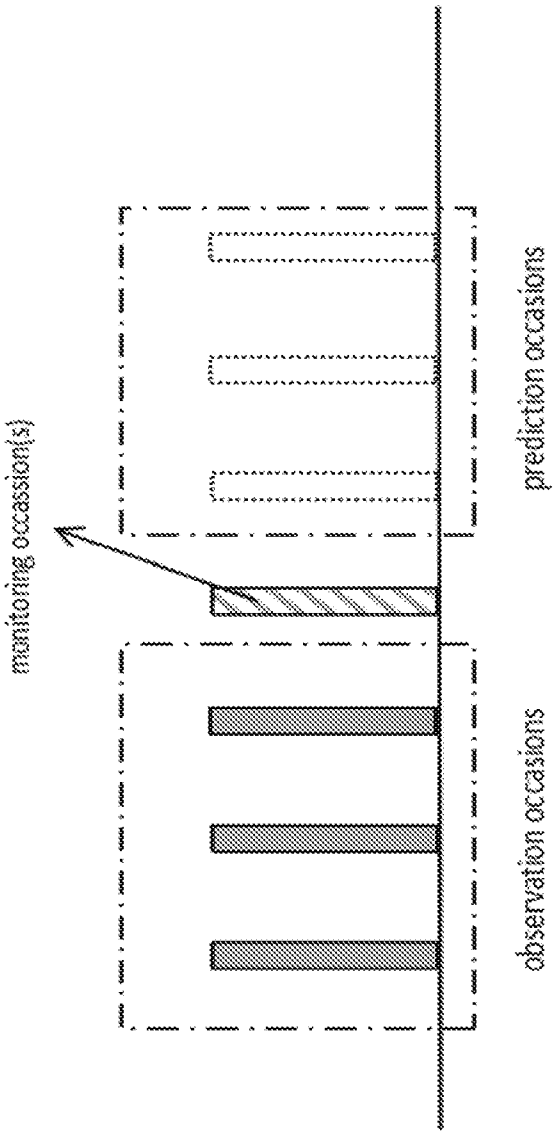
monitoring occasion(s)
observation occasions
prediction occasions

FIG. 4

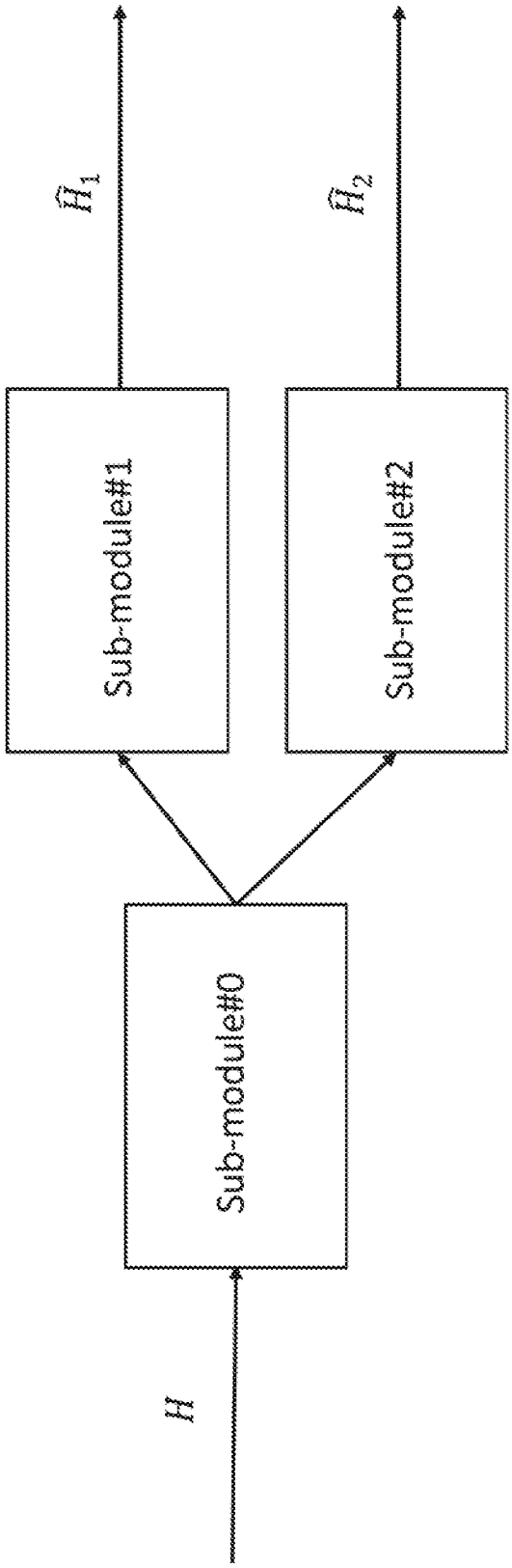

FIG. 5

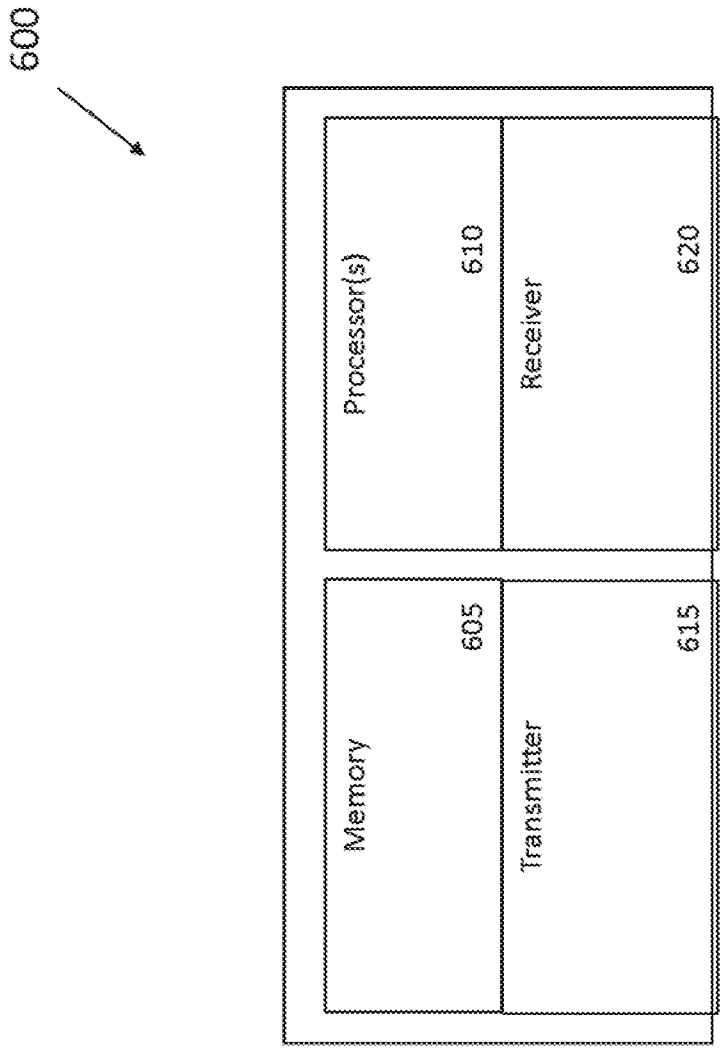
600
Processor(s)
610
Receiver
620
Memory
605
Transmitter
615

FIG. 6

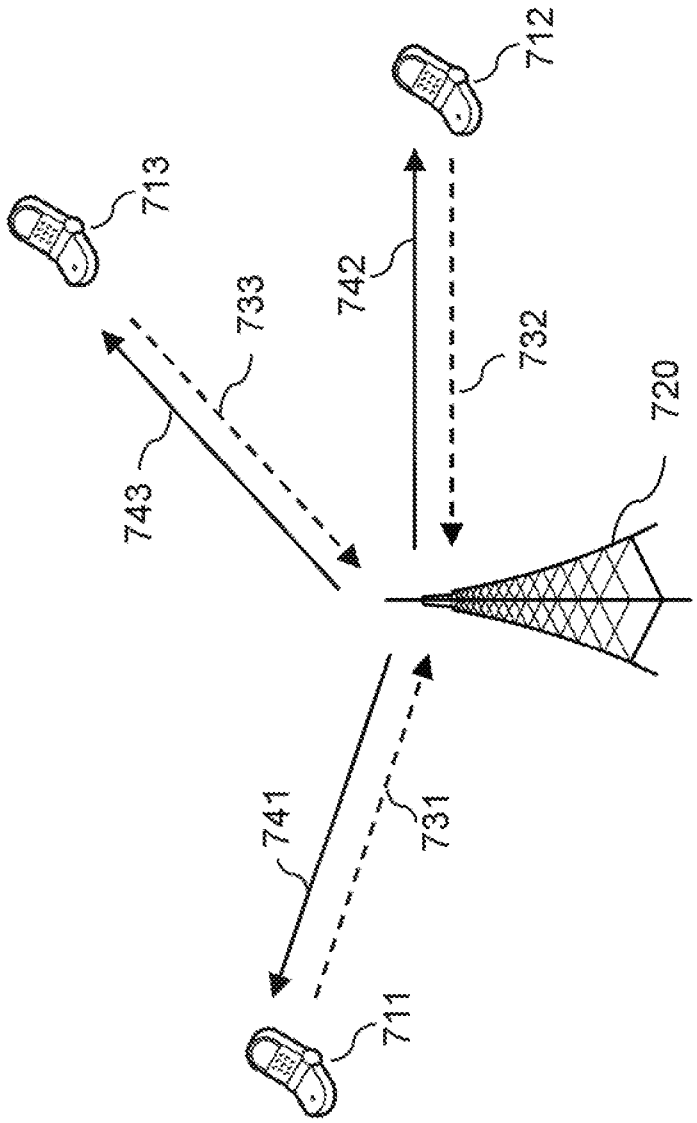

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/112294

A. CLASSIFICATION OF SUBJECT MATTER

H04W8/22(2009.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC:H04B, H04W, H04L, G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, ENTXTC, 3GPP: capability, indicate, model, report, performance indication, AI/ML

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2021142609 A1 (GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.) 22 July 2021 (2021-07-22) description, page 3 line 20 to page 11 line 25, figures 1-16	1-13, 24-25
X	CN 114443556 A (INTEL COMPANY) 06 May 2022 (2022-05-06) description, paragraphs 19-122	14-25
A	CN 114091679 A (HUAWEI TECHNOLOGIES CO., LTD.) 25 February 2022 (2022-02-25) the whole document	1-25
A	WO 2022008037 A1 (NOKIA TECHNOLOGIES OY) 13 January 2022 (2022-01-13) the whole document	1-25
A	INTEL CORPORATION. "R3-213468 High level principle and Functional Framework of AI/ML enabled NG-RAN Network" *3GPP TSG-RAN WG3 Meeting #113-e*, 26 August 2021 (2021-08-26), the whole document	1-25

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 April 2023	**21 April 2023**

Name and mailing address of the ISA/CN	Authorized officer
CHINA NATIONAL INTELLECTUAL PROPERTY ADMINISTRATION **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China**	**TANG,GuangQiang** Telephone No. **(+86) 010-53961734**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/112294

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
WO 2021142609 A1	22 July 2021	EP 4087343 A1	09 November 2022
		US 2022342713 A1	27 October 2022
		CN 114930945 A	19 August 2022
CN 114443556 A	06 May 2022	None	
CN 114091679 A	25 February 2022	WO 2022041947 A1	03 March 2022
WO 2022008037 A1	13 January 2022	None	