Appendix 5. Specific descriptive information to be provided for medical device functions relying on machine learning processes (technologies falling within the scope of artificial intelligence)

Preliminary observations

If your MD is based on at least one machine learning process, you should complete this grid to provide the committee members with the information needed in this area of your MD. Included in the submission guide in September 2020, it should be amended as needed in line with technological upgrades.

Depending on the case, you should construct one or more grids, the principle being that you complete one grid for each "smart" function of the device:

- where there is only one function relying on machine learning processes: you should complete a single grid. This particularly applies when the interlinking, or succession, of several processes can justify their grouping in the same grid when they contribute to the same "smart" function.
- in the case of an MD including several functions of this type, you should complete one grid per function.

Depending on the type of technology, some items may not be adapted. In these cases, you should specify this, providing a justification. Conversely, you can also supplement the descriptive information listed with any information deemed useful.

Descriptive grid

		Information to help you complete the grid
Purpose		
1	Note the claimed use and the envisaged scope of the medical device (MD) including one or more machine learning algorithms	Is it used for example to: – help the patient adjust the dosage of their treatment? – predict or provide early detection of the occurrence of a clinical event? You should specify the pathologies or clinical scenarios addressed, or the multidisciplinary nature of the MD, where applicable. You should also systematically specify the user (patient or professional).
2	Specify the benefit of the information provided or decisions made by machine learning processes	In this section, specify the "smart" function in which machine learning has played a direct role. For example: – Determining a severity score? – Calculating a dose for treatment adaptation?
3	Note the characteristics of the target population and, where applicable, the	These may be: – Demographic (age groups, sex, etc.)

	characteristics for which use of the MD is unsuitable, due to non-indication, contraindication, or factors influencing the product result	– Physiopathological (pregnancy, diabetics or asthmatics, etc.) or morphological (lower limb amputees, etc.) – Clinical or biological (disease stage, etc.)
4	Describe the operating environment of the smart system	Particularly specify the environmental conditions (meteorological, brightness, temperature, ground conditions, etc.) used to characterise the operating range.

Data		
Description of samples used for initial model learning or relearning		
5	Specify the characteristics of the population on which the initial model learning or relearning data are based	These may be: Demographic (age groups, sex, etc.) Physiopathological (pregnancy, diabetics or asthmatics, etc.) or morphological (lower limb amputees, etc.) Clinical or biological (disease stage, etc.) Differentiate the population on which the initial learning data are based (training, validation, and testing) from that used during the relearning phase (retraining, validation, and testing of updated system), where applicable.
6	Specify the characteristics of each sample used for the initial model learning or relearning data .	Expected: their function, size and composition. Included variables must be cited. The manner in which rare events are taken into account must be described. Differentiate the databases of the initial learning phases (training, validation, and testing) and in the relearning phase (retraining, validation, and testing of updated system), where applicable
7	Specify the methodology for separating or segmenting samples	For example, specify the procedures for separating (methods used and proportions) and segmenting (random, by date, by subject, etc.) the training, validation, and test data sets Differentiate the databases in the learning and relearning phases, where applicable.
Description of input data involved in initial model learning or relearning		
8	Specify the characteristics of the variables (variable type, distribution, etc.)	Differentiate the training, validation and test corpus where applicable.

9	Indicate the method of acquisition of the variables and their origin during the learning process	For example, was a variable entered by a patient? Does it come from a sensor? Was it generated from virtual patient models? Specify whether the variables were extracted from corpora of open or purchased data, and indicate which, where applicable, as well as whether they are long-term or not. Specify the types of sensors used during variable acquisition, where applicable.
10	Describe the pre-processing applied to the data.	For example, tasks to clean, transform, reduce, increase data (additions of artificial noise, artificial interference simulating weather variations or sensor faults, etc.) Specify the data concerned and the proportion of data modified by these pre-processing operations
11	Indicate the proportion of missing data, among the raw data, and describe their management.	Specify the types of missing data (random or anticipated).
12	Explain the procedures in place to detect and manage outliers, where applicable	In particular, specify how outliers (e.g. physiologically impossible data) are distinguished from atypical values (e.g. rare events)
13	Justify the representativeness of the samples used for the initial learning (training, validation, and testing) of the algorithm in relation to the data to which this algorithm will be exposed once deployed	A justification of the representativeness criteria is expected. Particularly specify the tools and methods used to verify the representativeness of the samples and detect potential bias. In the case of incremental or continuous learning, indicate the potential impact of updates on all learning data.
Description of input data involved in decision-making (once the medical device has been deployed)		
14	Specify the characteristics of the variables (type, distribution, etc.)	Indicate the main sources of difference between the training, validation and test data, and the data involved in decision-making, once the system has been deployed (different sensors, different environmental conditions, etc.).
15	Indicate the method of acquisition of the variables and their origin	For example, was a variable entered by a patient? Does it come from a sensor? Indicate the measurement range and sensitivity settings of the measuring devices, where applicable.

16	Describe the pre-processing applied to the data used for decision-making	For example, tasks to clean, transform, reduce data, etc.
17	List the output variables (model prediction objects) and their characteristics (type, unit, etc.)	Specify the variables that will be processed in relation to the objective. Specify whether they are processed by another component of the MD or whether they are communicated to the user (if so, how)

Model: description of training, validation, and testing, before and after MD deployment		
18	Describe the type of learning used	Is this machine learning process: – supervised, – semi-supervised, – unsupervised, – by reinforcement, – federated, – centralised, – other? These suggestions are not mutually exclusive.
19	Describe the type of task automated by the algorithm	supervised classification (determine the ranking criteria), unsupervised classification (define classes), ranking (rank in classes), regression (quantitative projection), segmentation, other?
20	Specify the update frequency	Is learning: – continuous (system learning autonomously after deployment)? – initial (algorithm designed based on learning then fixed after MD deployment)? – or incremental (algorithm for which updating of the structure and/or settings after MD deployment is supervised by a human and involves prospective and/or retrospective validation)?
21	Describe the model selection criteria	For example, the error rate, computing time, the number and nature of the data available, explainability or embeddability, etc.

		Do not go into detail on the system input data (covered in questions 5 to 17), or the test methods used (covered in questions 26 to 32)
22	Describe the various training, validation, and test phases, prior to MD deployment	Indicate the various training, validation, and test phases, particularly specifying whether they are based on individual or collective data. Do not go into detail on the test methods in place (covered by questions 26 to 32).
23	Describe the training, validation and test strategies for updates, if applicable	Indicate the various training, validation, and test phases applied once the MD is deployed, particularly specifying whether they are based on individual or collective data. Specify in particular the retraining frequency, the variables involved and the data inclusion period, the retraining computation location (locally on the MD or on server). Do not go into detail on monitoring and/or human intervention in these phases (already covered in questions 24 and 25), or the update test methods (already covered in questions 26 to 32).
24	Describe how parties involved in system development are referenced	Specify whether the human managers or legal entities involved at each stage of the life-cycle of the smart MD (data collection, development, qualification, use and feedback for MDs with AI capability) can be identified.
25	Where applicable, state in which cases a human being is involved in the retraining process	For example, in the case of active learning, specify the frequency and qualification of the person involved. In the case of operator annotation, specify the operator's qualification and role.

Functional characteristics		
Performance and qualification		
26	Describe and justify the choice of metrics used to measure performance …	For example: Root-mean-square deviation, Area Under Curve, F1-score, ZoneMap, Jaccard
27	Describe the processing operations applied which have had a substantial impact on performance	For example, in the case of class imbalances in the context of supervised classification, indicate whether class rebalancing has been carried out, as well as the method used.
28	Describe the identified risks of over- and under-learning and the methods in place to remedy this	A link may particularly be established with the responses to question 7 on data separation/segmentation.

29	Specify whether the system returns a confidence rating for each of its decisions	This could for example indicate, for an image classifier, whether it returns the probabilities of the input image to belong to each of the classes
30	Describe the qualification methods of the machine learning system	Particularly specify the test protocol in place and the procedures used to ensure performance measurement repeatability and test reproducibility. If using formal methods to qualify the machine learning system, justify the choice of methods used and how the ranges on which the formal methods were applied were defined.
31	Indicate the performance measurement results on the different data sets	For example, the error rates supplied by the metrics on the training, validation, and test databases, according to the distribution applied. Specify whether a separate database from the training, validation, and test databases was used to qualify the model. Specify, in the case of formal proof analysis, the results obtained and the validity range of these results.
32	Specify the performance thresholds selected (limit values, maximum error rate, etc.) and explain the choice of these thresholds	
System robustness		
33	Specify the tools in place to generate antagonistic examples in the performance evaluation and qualification phase	
34	Specify the tools in place to monitor the performances of the smart system after its deployment	Particularly specify the mechanisms in place to measure model degradation and/or concept drift (regular evaluation campaigns, etc.), as well as performance degradation logging, archival and analysis
35	Specify the thresholds selected (limit values, maximum error rate, etc.) for tracking model degradation and/or concept drift and explain the choice of these thresholds	
36	Specify the measures in place in the case of automatic or user detection of model degradation or concept drift	For example: information sent to the user, substitution of the learning algorithm by an expert system, retraining, etc.

System resilience		
37	Describe the system in place for input data anomaly detection in operational use	This could for example concern the detection of data outside the nominal operating range of the smart system
38	Describe the potential clinical and technical impacts induced by anomalies on the input data of the machine learning system	For example, what will happen: In the event of non-correction of outliers? In the event of a declarative value input error by the patient? Due to the level of uncertainty associated with the input data (physiological, environmental data, etc.)? In the event of data unavailability? In the event of data integrity loss?
39	Specify the measures in place in the case of automatic or user error detection (e.g. malfunction damaging the input data)	For example: information sent to the user, degraded mode, substitution of the learning algorithm by an expert system, clinician or technician intervention, etc.
Explainability and interpretability		
40	Indicate the explainability elements provided by the smart device	Specify, where applicable, the explainability technique(s) in place to help understand the main factors leading to the decision taken or proposed by the machine learning algorithm. Specify the recipient of these explanations: user (caregiver or patient), developer, etc. Also indicate whether the explanations are recorded for retrospective analysis by experts (users and/or developers).
41	Indicate the interpretability elements, i.e. the parameters (input variables, weightings, etc.) influencing decision-making, as well as the method used to identify them	For algorithms with initial or incremental learning, are these parameters identified (e.g. by means of influence functions)?
42	Specify whether the decisions and actions of the smart device are compared to professional guidelines	Particularly indicate whether the machine learning algorithm outputs are compared to professional guidelines in real time or retrospectively. Specify whether these comparisons are made accessible to users. For example, are the machine learning algorithm outputs compared to those of an expert system modelling care guidelines?

Glossary

This glossary is solely intended for use alongside this descriptive grid of machine learning algorithms in the context of CNEDiMTS medical device evaluation.

Term	Definition	Source
Machine learning	Process whereby an algorithm evaluates and improves its performances without programmer intervention, by repeating its execution on data sets, until appropriate results are regularly obtained.	11
Unsupervised learning	Machine learning in which the algorithm uses a raw data set and obtains a result based on the detection of similarity between some of these data items.	11
Supervised learning	Machine learning in which the algorithm practises a defined task using a data set each accompanied by an annotation indicating the expected result	11
Ranking	Action of ranking objects, persons in a certain order.	12
Supervised classification	Technique consisting of categorising data according to their proximity thus making it possible to differentiate among two or more discrete classes.	13
Concept drift	A machine learning algorithm in which the parameters are fixed becomes inconsistent with its environment if the latter has been updated.	14
Range of use	Description of the environment and target population, for which the algorithm or program is designed.	-
Data	Representation of the observation of a variable on an element, individual, or instance of a population, intended to facilitate its processing.	-
Raw data	Data having undergone no transformation since the initial observation.	-
Input data	Data used for model learning or decision-making.	-
Output data	Value representing all or part of the decision made by the algorithm based on the input data.	-

[11] Official Journal of 09/12/2018
[12] https://www.larousse.fr/dictionnaires/francais/classement/16405
[13] Based on ISO definition (drafting in progress)
[14] Tsymbal, A. (2004). The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin, 106(2), 58.

Sample	Representative fraction of a population or a statistical universe	15
Training	Machine learning process through which the artificial intelligence system builds a model from data.	13, 16
Antagonistic example	Borderline case placing the system under evaluation in difficulty.	-
Explainability	Ability to link and explain the elements taken into account by the algorithm, for example the input variables, and their consequences, for example, on the prediction of a score, and thus on the decision. The explanations must be adapted to the comprehension level of the person for whom they are intended.	-
Hyperparameter	Parameters tweaked during successive runs of training of a model in order to check under- and over-learning in particular.	17
Information	Knowledge element expressed by a data set according to a defined code, with a view to being stored, processed, or communicated. An item of information is obtained from the interpretation of one or more pooled data items.	18
Interpretability	Ability to render the operation of an artificial intelligence system comprehensible. An algorithm is "interpretable" when its operation is accurately understood, for example, when an expert system models a decision tree.	13
Data set	Collection of data	-
Model	Mathematical construction generating an inference or a prediction from input data.	13
Parameter	Coefficient of a model that the machine learning system estimates or trains on its own and which impacts the output data.	17
Resilience	Ability of the system to maintain its conformity with performance and/or security requirements in the presence of input data outside its range of use (e.g. due to a sensor fault).	-
Robustness	Ability of a system to maintain its performance level whatever the circumstances.	13

[15] Centre National de Ressources Textuelles et Lexicales www.cntrl.fr
[16] From the Montreal Declaration for a Responsible Development of Artificial Intelligence
[17] https://developers.google.com/machine-learning/glossary
[18] https://www.dictionnaire-academie.fr/article/A9I1218

Segmentation "Data segmentation" "Automatic segmentation task"	Data segmentation: division of a corpus of data into several bases (e.g. training, validation, and testing) either based on objective criteria (date, age, etc.), or at random. Automatic segmentation task: extraction and automatic recognition of zones of interest from input data (e.g. an image).	19
Test	Process consisting of detecting errors associated with running an algorithm or a program based on input data sets not used during the training phase.	-
Validation	Process consisting of testing, observing and optimising (hyperparameters) system behaviour during running so as to ensure, in the range of use, that the output data are in line with the expected results.	13
Variable	Observable characteristics (qualitative or quantitative) of an element.	-

[19] Rakoto–Ravalontsalama, M. (1990). Méthodes de segmentation automatique d'image. Analyse quantitative des formes, Télédétection, pp251-260.