|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-J-038 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | DAISAM | E-meeting, 30 September – 2 October 2020 | |
| **DOCUMENT** | | | | |
| **Source:** | | WG-DAISAM | | |
| **Title:** | | WG-DAISAM Metrics and measures paper questionnaire | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Alixandro Werneck | | Email: alixandrowerneck@outlook.com |

|  |  |
| --- | --- |
| **Abstract:** | This is a questionnaire to elicit details about the data and development process of your algorithm. These details are important to make useful qualitative conclusions about the applicability and safety of the algorithm. We enabled to elaborate on every answer. If you feel like a yes/no answer is sufficient to answer a question, you can also just write this into the answer field. You can also type 'NA' if a question is not applicable to your use case. |

|  |  |  |
| --- | --- | --- |
| DAISAM AI4H Measures and Metrics – Questionnaire | | |
| Purpose: This is a questionnaire to elicit details about the data and development process of your algorithm. These details are important to make useful qualitative conclusions about the applicability and safety of the algorithm. We enabled to elaborate on every answer. If you feel like a yes/no answer is sufficient to answer a question, you can also just write this into the answer field. You can also type 'NA' if a question is not applicable to your use case.  Asterisk (\*) -Mandatory | | |
| Section 1- Basic ML Model Information | | |
| 1 | Who developed the ML model? (e.g. details of developer, organization, etc.) |  |
| 2 | When was the ML model developed? (e.g. time period) |  |
| 3 | Which type of ML model was used? (e.g. architecture, i.e. CNN, Decision tree, etc.) |  |
| 4 | Which version of the ML model is this, and how is it different from previous releases? |  |
| 5 | How does the ML model output look like? (i.e. disease class, probability of having the disease, risk score, etc.) |  |
| 6 | How many distinct tasks does the ML model perform? (e.g. detection task followed by classification task) |  |
| 7 | Are decision thresholds being used for classification? If yes, specify the thresholds and the ‘thresholding rule’. Can you also state the clinical significance of the selected operating threshold, if any? |  |
| 8 | Did you apply any technique to speed up the ML model training computational process? (e.g. transfer learning technique for deep neural networks). If, yes, did the technique improve the ML model performance? And by how much did it improve the performance? |  |
| 9 | Were patients and clinicians involved or consulted during ML algorithm selection stage, algorithm development stage or algorithm acceptance and adoption stage? |  |
| 10 | Where can we find more detailed information on the ML model? (e.g. repository URL or cite publication) |  |
| 11 | How should the ML model be cited? \* |  |
| 12 | Are there other relevant citations that were used for the ML model creation? If yes, please list them here. \* |  |
| 13 | Does the ML model have a license? (Yes / No) \* |  |
| 14 | Where can we send questions or comments about the ML model? ( e.g. corresponding author’s email address) \* |  |
| Section 2- Intended Use of the ML Model | | |
| 1 | Primary intended use: For which purpose was the ML model developed for? |  |
| 2 | Describe how the ML model fits into the intended health intervention workflow? (e.g. as autonomous tool, assistive tool, augmentative tool, as add-on unit to existing system/workflow, as replacement unit for existing system/workflow component, as new stand alone application, etc) |  |
| 3 | Is your ML model optimized for a specific local clinical operational setting (i.e. specialization)? If not (i.e. generalization), has the model effectiveness been reassessed / re-evaluated/ re-calibrated for multiple clinical settings of different health environments (e.g. with variability in workflow, demographics, etc)? |  |
| 4 | For ML model generalization, was it ensured that input data were representative of variations in data acquisition and reconstruction parameters, target population, operating time scales (i.e. for the purpose of simulating the real-world protocol differences across multiple clinical settings in different health systems, geographic boundaries)? |  |
| 5 | Could the ML model be applied in different contexts and if yes, give details? |  |
| 6 | Are there similar ML models that can be confused with the proposed ML model? Are there contexts in which the proposed ML model is not recommended / advisable to be applied? |  |
| 7 | Which of the following clinical considerations apply to your ML model outcome?  a. Avoiding ‘False Positives’ is having HIGHER PRIORITY over allowing ‘False Negatives’  b. Avoiding ‘False Negatives’ is having HIGHER PRIORITY over allowing ‘False Positives |  |
| Section 3- ML Model Development | | |
| Information about data used for developing the ML model | | |
| 1 | Where was the data collected? \* |  |
| 2 | Who created this dataset? (e.g. company, consortia, hospital, etc.) \* |  |
| 3 | For what purpose was the dataset created? \* |  |
| 4 | When was the data collected? (e.g. time frame) \* |  |
| 5 | Who funded the creation of the dataset? (e.g. company, consortia, hospital, etc....) |  |
| 6 | What kinds of data modalities are used? (e.g. CT-Scans, text reports, data frame, etc. …) |  |
| 7 | What kinds of predictors are used as ML model input? (e.g. images, other variables i.e. biomarker, age, etc.) |  |
| 8 | What kinds of instruments were used to capture the model input data? (i.e., camera type, image resolution, etc.). \* |  |
| 9 | Are there any errors, sources of noise, redundancies present in the dataset? If so, please provide a description. \* |  |
| 10 | How many total data instances are there in the original dataset? |  |
| 11 | Is there a label associated with data instances? If yes, how and by whom were these annotated? In case of manual annotation, specify the competency of the annotation specialist? |  |
| 12 | How many instances of each label class were present in the raw dataset? |  |
| 13 | Does the data identify any subpopulations (age-group, gender, ethnicity, etc.)? If yes, specify the type \* |  |
| 14 | Did you encounter any ‘missing data’ problem for any predictor variables? If yes, what could be reasons for missing data? How was this problem handled? |  |
| 15 | What kinds of preprocessing/cleaning steps have been done to prepare data for ML model development? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, imputation, etc). \* |  |
| 16 | Are there archived versions of the raw/original dataset? (Yes/ No) \* |  |
| 17 | How was the original dataset split into training, test and validation set? Please indicate the number and percentages, stratifications (i.e. matched case-control) |  |
| 18 | Specify the inclusion and exclusion criteria for instances in the training dataset? Were any instances from the original dataset excluded from the ML model training or evaluation? If yes, provide the reason? \* |  |
| 19 | How many instances of each label class were present in the training dataset?(e.g. proportionate sample size of different classes) |  |
| 20 | What kind of technique was applied to reduce ML model over-fitting? (i.e. k-fold cross validation) |  |
| 21 | Is there anything significant about the way how the data was preprocessed/cleaned/selected that might impact future use of the developed model? \* |  |
| Section 4- Legal Aspects | | |
| 1 | Does the dataset contain confidential/personal information? (i.e. sensitive variables such as ethnicity, religion, etc) (Yes/ No) |  |
| 2 | Is it possible to identify individuals from the dataset? Were the datasets de-identiﬁed / anonymised ? (Yes / No) \* |  |
| 3 | If the dataset relates to people, were they informed about for what purpose the dataset would be used for? Did you obtain consent from them? If so, how? Were they provided with any mechanism to revoke their consent in the future or for specific uses? \* |  |
| Section 5- Ethical Considerations | | |
| 1 | Is the ML model intended to inform decisions about matters central to human life or flourishing – e.g., health or safety? Or could it be used in such a way? |  |
| 2 | Which kind ethical considerations did you follow in your ML model development? |  |
| 3 | Does the ML model use any sensitive attributes to make predictions? (e.g. ethnicity, gender, etc) Yes / No |  |
| 4 | What kind of risks may be present in model usage? Try to identify the potential recipients, likelihood, and magnitude of harms of these risks. If these cannot be determined, note that they were considered but remain unknown. \* |  |
| 5 | Did you apply any mitigation strategies to overcome risk of bias across sensitive attributes during ML model development? |  |
| Section 6- ML Model Evaluation and Metrics | | |
| 1 | Specify the type of ML prediction model evaluation that was carried out \*   1. Development only – ML Prediction model development without external validation. (Includes internal validation methods, such as bootstrapping and cross-validation techniques. 2. Development and validation – ML Model dev + external validation in other participants 3. Other |  |
| 2 | If validation on external data was carried out: Was the data source (location, i.e. same hospital, collection and selection method) of the validation data, etc. the same as for the training data? (Obtained by splitting the original dataset) Yes / No |  |
| 3 | If you ticked 'No' for the previous question, please specify the differences between the training and evaluation data. |  |
| 4 | Specify the inclusion and exclusion criteria for instances in the evaluation dataset? \* |  |
| 5 | How many instances of each label class were present in the evaluation dataset? |  |
| 6 | What measures of ML model performance are reported, and specify their selection criteria in comparison with other measures of ML model performance ? \* |  |
| 7 | What are foreseeable salient factors for which ML model performance may vary, and how were they determined? (e.g. sex, age, ethnicity, race, and other attributes of the population of interest) |  |
| 8 | Which among these population types are considered relevant to your use case: Check all that is applicable. | * Diverse demographics * Persons with disabilities * Ethnic minorities * Women * Children * Geriatric * Persons facing risk of exclusion or discrimination in social status and educational status * other |
| 9 | Were clinicians or other domain experts involved or consulted during the selection of ML model performance metric? Was the performance metric chosen purely based on its statistical merits? Or Did the underlying ‘clinical decision process’ considerations influence the metric selection? |  |
| 10 | How did the ML model perform for each evaluation metric? Please specify whether performance evaluation was done on training or evaluation (test) data. \* |  |
| 11 | Please provide the ML model performance plots? (e.g. ROC- AUC plot) (Please specify the file location)\* |  |
| 12 | Are there output classes or disease types for which the ML model performed worse than others? Please provide the confusion matrix. |  |
| 13 | Did you investigate ML model performance variations across different groups? If yes, specify the groups and the corresponding ML model performances for these groups? \*  (Note: groups can be defined by cultural, demographic, and phenotypic variables.) |  |
| 14 | Have you applied statistical testing to compare ML model performance across different groups? If yes, specify the tests and significance level of p-values applied. Please provide performance results across groups with confidence intervals. |  |
| 15 | Did you perform an input feature importance analysis on your ML model? If yes, how was it done and what were the results obtained? Based on this, which of the input features was ranked as the most important one? And which of the features have high positive influence and which of the features have least influence on performance results? |  |
| 16 | Did you use approaches to assess uncertainty and variability in model performance? If yes, which ones? |  |
| 17 | Did you use any approaches to assess uncertainty and variability of ML model performance? If yes, specify them? |  |
| Section 7- Caveats and Recommendations | | |
| 1 | Did the ML model evaluation results suggest any further testing? |  |
| 2 | Are there relevant subgroups that were not represented in the evaluation dataset? |  |
| 3 | Are there additional recommendations for the ML model use? |  |
| 4 | Is the ML model performance comparable to the performance scores or the level of competence of the clinician/specialist/user in the clinical setting? Specify the competency skill sets for the clinician/specialist? |  |
| 5 | Under what conditions the ML model outperforms the clinician/specialist/user?  Under what conditions were the performances of both the ML model and clinician/specialist comparatively similar? (e.g. statistically significant differences in the ROC-AUCs  Observed or not)  Under what conditions the ML model performs worse than that of the clinician/specialist/user? |  |
| 6 | Do you recommend your ML model as a complete replacement for the proposed clinical solution? If not, what are the gaps observed as a result of replicating the clinical setting with your experimental design? Did the ML model fail to address any relevant clinically important findings? |  |
| 7 | Is the model performance comparable to the performance scores or the level of competence of the clinician/specialist/user in the clinical setting? What are the competency specifications for the clinician/specialist? For what conditions the model performed worse than that of the clinician/specialist/user? And for what conditions were the performances of both the model and clinician/specialist comparable? (I.e. where are the statistically significant differences in the ROC-AUCs observed). |  |
| 8 | Can adoption of your proposed ML model reduce the overall clinical practice cost (or enhance the clinical practice savings)? If yes, in what way? (e.g. faster patient diagnosis / treatment, percentage reduction in clinician cognitive workload, degree of automation / semi-automation introduced, degree of smartness/intelligence augmentation new knowledge discovery, enabling replacement or redefinition of  existing gold standard, etc) |  |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_