|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-J-036 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | E-meeting, 30 September – 2 October 2020 |
| **DOCUMENT** |
| **Source:** | Editors |
| **Title:** | Updated DEL06: AI training best practices specification (E-meeting, 30 September – 2 October 2020) |
| **Purpose:** | Discussion |
| **Contact:** | Sim XinmingAI Singapore | Email: xinming@aisingapore.org |
| **Contact:** | Stefan WinklerAI Singapore | Email: stefan@aisingapore.org |

|  |  |
| --- | --- |
| **Abstract:** | Machine learning models for AI in Health are deployed in high-impact tasks. As a result, it is important to follow best practices for training and documentation so as to achieve maximum performance and transparency. The first part of this documentprovides a review of best practices for proper AI model training. The second part of this document provides guidelines for model reporting. This document was first submitted as I-032 at the FG-AI4H meeting I (e-meeting), 7-8 May 2020. |

|  |  |
| --- | --- |
|  | **International Telecommunication Union** |
|  |  |
| **ITU-T** | **FG-AI4H Deliverable** |
| TELECOMMUNICATIONSTANDARDIZATION SECTOROF ITU | (draft 30 September 2020) |
|  |  |
|  | DEL06AI training best practices specification |
|  |  |

Summary

Machine learning models for AI in Health are deployed in high-impact tasks. As a result, it is important to follow best practices for training and documentation so as to achieve maximum performance and transparency. The first part of this documentprovides a review of best practices for proper AI model training. The second part of this document provides guidelines for model reporting.

Keywords

Machine learning, deep learning, healthcare, model training

Change Log

This document contains Version 1 of the Deliverable DEL06 on "AI training best practices specification" [approved at the ITU-T Focus Group on AI for Health (FG-AI4H) meeting held in (draft 30 September 2020)].

|  |  |  |
| --- | --- | --- |
| **Editor:** | Sim XinmingAI Singapore | Email: xinming@aisingapore.org |
|  | Stefan WinklerAI Singapore | Email: stefan@aisingapore.org |

**CONTENTS**

|  Page |
| --- |
| [1 Scope 1](#_Toc52183842)[2 References 1](#_Toc52183843)[3 Terms and definitions 2](#_Toc52183844)[3.1 Terms defined elsewhere 2](#_Toc52183845)[4 Abbreviations 2](#_Toc52183846)[5 Conventions 3](#_Toc52183847)[6 AI pipeline: a brief description 3](#_Toc52183848)[7 Best practices for proper model training 3](#_Toc52183849)[7.1 Feature selection and extraction 3](#_Toc52183850)[7.2 Treatment of missing data 4](#_Toc52183851)[7.3 Model architecture 4](#_Toc52183852)[7.4 Hyperparameter optimization 4](#_Toc52183853)[7.5 Overfitting and underfitting 5](#_Toc52183854)[7.6 Model generalisability 6](#_Toc52183855)[7.7 Model selection 7](#_Toc52183856)[8 Model reporting framework 7](#_Toc52183857)[8.1 Basic model details 8](#_Toc52183858)[8.2 Use case and users 8](#_Toc52183859)[8.3 Model performance metrics 9](#_Toc52183860)[8.4 Model performance factors 9](#_Toc52183861)[8.5 Data 9](#_Toc52183862)[8.6 Analysis 9](#_Toc52183863)[8.7 Ethical considerations 10](#_Toc52183864)[Bibliography 11](#_Toc52183865) |

**List of Figures**

|  Page |
| --- |
| [Figure 1: A brief description of an AI pipeline 4](#_Toc51952578) |

ITU-T FG-AI4H Deliverable DEL06

AI training best practices specification

Summary

Machine learning models for AI in Health are deployed in high-impact tasks. As a result, it is important to follow best practices for training and documentation so as to achieve maximum performance and transparency. The first part of this documentprovides a review of best practices for proper AI model training. The second part of this document provides guidelines for model reporting.

# Scope

This deliverable provides a general introduction and background to model training across a typical AI pipeline. It explains basic key concepts, considerations, practices, and limitations on the different aspects of model training in the healthcare domain and points readers to cited sources or studies where advanced materials can be found.

# References

1. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2017;24(1):198-208.
2. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837–1847.
3. G. Luo, A Roadmap for Semi-automatically Extracting Predictive and Clinically Meaningful Temporal Features from Medical Data for Predictive Modeling, Global Transitions.
4. Rajkomar, A., Oren, E., Chen, K. et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 1, 18 (2018).
5. Wong, W.K., Boscardin, W.J., Postlethwaite, A.E., & Furst, D.E. (2011). Handling missing data issues in clinical trials for rheumatic diseases. Contemporary clinical trials, 32 1, 1-9.
6. Little RJA. Pattern-mixture models for multivariate incomplete data. J Am Stat Assoc 1993;88:125–34.
7. Diggle PJ, Kenward MG. Informative drop-out in longitudinal analysis (with discussion). Appl Stat 1994;43:49–93.
8. I. Zion, S. Ozuomba and P. Asuquo, "An Overview of Neural Network Architectures for Healthcare," 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), Ayobo, Ipaja, Lagos, Nigeria, 2020, pp. 1-8.
9. Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine, 101822..
10. Frazier, P.I. (2018). A Tutorial on Bayesian Optimization. ArXiv, abs/1807.02811.
11. Bellemo, V., Lim, Z. W., Lim, G., Nguyen, Q. D., Xie, Y., Yip, M. Y., ... & Lee, M. L. (2019). Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. The Lancet Digital Health, 1(1), e35-e44.
12. Kelly, C.J., Karthikesalingam, A., Suleyman, M. et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17, 195 (2019).
13. Liang, Shiyu & Li, Yixuan & Srikant, R.. (2017). Principled Detection of Out-of-Distribution Examples in Neural Networks.
14. Hendrycks, D., & Gimpel, K. (2017). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. ArXiv, abs/1610.02136.
15. De Fauw, J., Ledsam, J.R., Romera-Paredes, B. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 24, 1342–1350 (2018).
16. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–1246. doi:10.1093/bib/bbx044.
17. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), 1589-1604.
18. Mitchell, Margaret et al. "Model Cards for Model Reporting." Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT\* '19 (2019): n. pag. Crossref. Web.
19. Trisha Mahoney et al. "AI Fairness: How to Measure and Reduce Unwanted Bias in Machine Learning", IBM Corporation, O'Reilly, 2020.
20. Weng, T. W., Zhang, H., Chen, P. Y., Yi, J., Su, D., Gao, Y., ... & Daniel, L. (2018). Evaluating the robustness of neural networks: An extreme value theory approach. arXiv preprint arXiv:1801.10578.

# Terms and definitions

## Terms defined elsewhere

This document uses the following terms defined elsewhere:

**3.1.1 Alarm fatigue [Wikipedia]:** Alarm fatigue or alert fatigue occurs when one is exposed to a large number of frequent alarms (alerts) and consequently becomes desensitized to them.

# Abbreviations

|  |  |
| --- | --- |
| AI | Artificial intelligence |
| AutoML | Automated machine learning |
| EHR | Electronic health records  |
| LOCF | Last observation carried forward |
| LR | Learning rate |
| MAR | Missing at random |
| MCAR | Missing completely at random |
| NMAR | Not missing at random |
| OHDSI | Observational Health Data Sciences and Informatics |
| OMOP | Observational Medical Outcomes Partnership |
| RNN | Recurrent Neural Networks |
|  |  |

# Conventions

This document does not use any particular conventions.

# AI pipeline: a brief description

While there are many different aspects to consider throughout the AI pipeline, in this paper we will be focusing on the few points highlighted in Figure 1.



Figure 1: A brief description of an AI pipeline

# Best practices for proper model training

Machine learning for healthcare is one that requires a high level of customisation, given the complexities of diseases as well as the human body. As with all types of machine learning applications, proper training needs to be ensured for the model to give accurate and high-performance results.

## Feature selection and extraction

Healthcare data exist in many forms, such as Electronic Health Records (EHR), emails, physicians' notes, medical scan images etc. Even just within EHR data itself, the number of predictor variables that could be used to predict patients' medical outcomes are large and complex. To effectively improve the outcome of a prediction, or expand the limitations of an AI model, the correct features, or predictor variables, must be carefully selected. Yet, employing traditional clinical models to fully utilize all features within the dataset would be tedious and time-consuming. There have been a few different pre-processing methods that could be used for feature selection.

* Traditional approaches. In traditional prediction models, predictor variables are typically a predefined list [1], such as blood pressure or cholesterol in the Framingham Heart Study [2]. In most cases, these variables are selected based on expert judgement or knowledge, but the downside being that the depth of patients' data is not completely utilized, and that they may fall short in model accuracy when compared to newer methods and approaches (see below).
* Semi-automatic approaches. Such approaches may consist of two components; one which relies on automatic data-driven techniques to identify and extract additional clinically meaningful features from large datasets for building machine learning models, while complementing existing pre-defined features selected manually based on expert judgement, established clinical studies or industry standards, potentially resulting in a model that performs better [3].
* Deep learning approaches. One of the key benefits of deep learning techniques is the model's ability to execute feature engineering without being explicitly programmed. This removes the need to pre-define specific predictor variables to be used. Deep learning models are also able to handle large datasets with messy or incomplete data, which is often the case for EHR data. In a study by Google AI on deep learning for Electronic Health Records [4], the utilisation of the full spectrum of EHR data using deep learning techniques has been shown to outperform traditional clinical models in prediction tasks such as in-hospital mortality, 30-day unplanned readmission, prolonged length of stay and final discharge diagnoses.

## Treatment of missing data

As with most publicly collected datasets, it is not realistic to assume that the data comes in a form which can be readily consumed by the AI model. In many cases, it is common to expect missing data within healthcare records. How the missing data (e.g. missing true positive alarms) is managed or rationalized could drastically impact the final model performance (e.g. resulting in false positives or false negatives), so this is of critical importance.

Missing data could be classified into three types: (1) missing completely at random (MCAR); (2) missing at random (MAR); and (3) not missing at random (NMAR). The easiest way to handle missing data is of course to simply ignore and remove either the value or the entire case itself. When doing so an underlying assumption is made, which is that the missing data is independent of both the observed and unobserved data, and this may not be the case in reality. Wong et al. (2010) have documented several extensive methods below which could be used to address missing data in various healthcare context [5], summarized below:

* Last observation carried forward (LOCF). LOCF carries forward observation from the last observed time point to fill the missing data at the endpoints.
* Complete-case analysis and available-case analysis. Complete-case analysis uses only data from patients with a complete record of all visits and ignores all patients with any missing data. Available-case analysis on the other hand uses all available data.
* Mean imputation, hot-deck imputation, regression imputation. Mean imputation uses the mean across all data at the time point to fill in the missing data. Hot-deck imputation fills the missing data with observed responses from another randomly selected but similar data point. Regression imputation uses a regression model to predict and fill the missing value.
* Mixed effects models and generalized estimating equations. For mixed effects models, a statistical distribution (e.g. gaussian distribution) is used to account for the missing data. On the contrary, an estimating function could also be used to predict the missing data.
* Inference for NMAR data. There are two methods that can be used to infer for NMAR data; pattern mixture models and selection models. Pattern mixture models [6] express the joint distribution of the responses and missing indicators using the distribution of all possible missingness patterns and the distribution of the responses given a specific missingness pattern. Selection models [7] express the joint distribution of the responses and the missing indicators using the opposite decomposition; the distribution of the responses and the distribution of the missing indicators given the responses.

## Model architecture

The wide spectrum of AI models available is daunting, and the complexity, number and types of models will only proliferate further over time. To begin with zero knowledge of AI, it would be laborious, particularly for a clinician, to even begin to understand which types of models to use for a particular use case, much less evaluate their performance. Zion et al (2020) [8] recently summarized a comprehensive list of different neural network architectures and their limitations for the healthcare domain based on their general use cases for bioinformatics, medical informatics, medical imaging, and public health below:

## Hyperparameter optimization

Hyperparameter are a set of parameters that cannot be learnt by the model, and whose values are specified before the machine learning process begins. They are used to control the model learning process, such as learning rates, dropout rates, batch sizes, epochs etc. Specific hyperparameters may also lead to possible overfits. For instance, overfitting could occur if the learning rate (LR) is too small. However, this is a trade-off to balance, as having large learning rates may run the risk of divergence in the model training phase.

Optimization of these hyperparameters ensures that the resultant model optimally minimises the predefined loss function for a given set of data. For healthcare professionals with little or no experience in machine learning, the process of optimizing these hyperparameters may be challenging. Furthermore, it becomes increasingly resource-intensive to do hyperparameter optimisation as the datasets increase in volume and complexity. Here we discuss some common manual and automated techniques used for hyperparameter optimization which could be used.

* Grid Search. One of the most basic methods of optimisation techniques, Grid Search does an exhaustive search of a manually specified subset of the hyperparameter space. This is one of the simplest techniques and is easy to implement. However, because it explores the space exhaustively, it spends substantial time evaluating hyperparameter combinations which are unpromising and is therefore less efficient that Random Search. At greater dimensions of hyperparameters, the number of function evaluations grows exponentially, which inevitable makes this technique computationally expensive.
* Random Search. Similar to the Grid Search, Random Search evaluates random combinations of hyperparameters within the search space and defined number of iterations and is also relatively easy to implement. As not all combinations are evaluated over the search space, this makes hyperparameter optimization computationally less demanding when optimizing for higher dimensions of hyperparameters as compared to Grid Search.
* Bayesian Optimization. Bayesian optimization is a commonly adopted optimisation framework for many automated machine learning (AutoML) system [9]. It is an approach to optimizing objective functions that take a long time (minutes or hours) to evaluate. It is best suited for optimization over continuous domains of less than 20 dimensions and tolerates stochastic noise in function evaluations. It builds a surrogate for the objective and quantifies the uncertainty in that surrogate using a Bayesian machine learning technique, Gaussian process regression, and then uses an acquisition function defined from this surrogate to decide where to sample [10]. The difference between Bayesian Optimization methods from Grid Search or Random Search is that it uses past evaluation results to determine the subsequent values to evaluate. In essence this makes the optimization process more efficient as it limits iterations with poor hyperparameter combinations while focusing on promising hyperparameter combinations obtained from past results.

## Overfitting and underfitting

Overfitting occurs when the AI model or algorithm fits the data too well and unnecessarily captures the noise present within which affects the performance of the model when generalising to data beyond that used in its training. Underfitting, on the contrary, is when the AI model or algorithm cannot capture the trends or features of the data well enough. Both poor overfitting and underfitting results in poor performance of the AI model. Both of these could happen due to factors such as model parameters/weights, algorithm and model selection.

* Selection of model or algorithms. It is important to ensure that the complexity of the model is befitting of the dataset size. In general, models or algorithms which are overly restrictive limits the learning of the dataset, and is likely to lead to an underfit, while models or algorithms which are overly complex captures unnecessary randomness within the training data and is likely to product an overfit. Neural networks can sometimes be prone to overfitting. While there is no single hard rule or guideline to follow, in instances of an overfit, one should aim to reduce the complexity of the network. This could be achieved by either reducing the number of hidden layers, reducing the number of neurons or adding dropout layers within the hidden layers.
* Recognizing the warning signs. Recognizing when the data is overfitted or underfitted is crucial in addressing the issue, and more often than not, an overfit is typically the case compared to an underfit. One of the ways to know when the best fit can be achieved for a given setting is by analysing the bias-variance trade-off. Bias is an error due to erroneous assumptions and is an indication of a possible underfit. Conversely, variance refers to the model or algorithm's sensitivity to perturbations within the datasets. High variance generally occurs when the model or algorithm is overly complex and captures the randomness with the sample, giving an indication of a possible overfit. To ensure a well- performing model, one needs to find right balance between bias and variance which minimizes the total error function.
* Cross-validation to prevent overfits. While there are other methods which could detect and prevent overfitting such as back-testing or regularization, one of the more popular methods is cross validation. Though there are many different variations of cross validation methods such as k-fold or stratified k-fold cross validation, the underlying principles are the same, which is to partition the data into (1) training subset and (2) validation subset. At any iteration, the model only trains and tunes on the training subset. The validation on the validation subset gives an indication of the performance on unseen (test) data. The final evaluation of the model on the completely unseen test dataset gives an unbiased assessment of its performance. If the model outperforms on the training data subset compared to the test set, it is likely to be an indicator that the model is producing an overfit.
* Number of model parameters/weights. The number of parameters affects the resulting performance of the model. Having too few leads to underfitting, while conversely having too many may lead to an overfit. It is recommended for parameter optimization to be carried out to determine what is the optimal number to be used to train the model. Underfitting can be resolved by increasing the number of parameters or using weaker regularization during model training.

## Model generalisability

The generalisability of AI models has always been a key challenge in the community, and many are far from achieving reliable generalisability. In the healthcare context, AI models are relatively constrained in their deployment to populations and settings similar to those they were trained on. For example, diseases may manifest itself slightly different across ethnicity. What happens if the AI models were trained on a particular ethnic distribution and tested on a different one? While there has been research showing some success in developing limited generalisable AI models across ethnic distributions [11], this remains a key concern for healthcare professionals in ensuring that the diagnosis made by the AI model is as accurate as it is expected to be.

* Generalisability across different populations. Kelly et al (2019) proposed several measures which could help to overcome the issue of AI generalisability across different populations [12]:
* Site-specific training to adapt an existing system for a new population, particularly for complex tasks like EHR predictions. Methods to detect out-of-distribution inputs [13][14]and provide a reliable measure of model confidence to prevent clinical decisions being made on inaccurate model outputs. For medical image classification, this problem may overcome by the curation of large, heterogenous, multi-centre datasets. However, the limitation to this method is that re-training becomes computationally demanding and expensive should the number of out-of-distribution cases far exceed a certain threshold.
* Generalisability across different settings. Healthcare institutions around the world vary in their protocols, hardware such as medical equipment and medical equipment software. This results in vastly different healthcare data generated by each unique healthcare institution, which makes it a nightmare to generalise AI system from one institution to another. While there have been efforts by organisations such as Observational Health Data Sciences and Informatics (OHDSI) to standardize a common data model called Observational Medical Outcomes Partnership (OMOP) for electronic healthcare recording, this remains a barrier to the interoperability of AI models across institutional and international boundaries.
	+ One possible way of overcoming this issue lies with the design of the AI model and training. Splitting a disease diagnosis step into segmentation and measurement may enable easier generalization to new imaging hardware by retraining only the segmentation model, which is more data- efficient [15].
* Generalisability across time. Besides the issue of adapting an AI system to new populations, one has to also consider the temporality and everchanging nature of diseases. In a study on Deep Learning for Healthcare, Miotto et al (2018) highlighted that diseases are always progressing and changing over time in a nondeterministic way [16]. However, many existing deep learning models, including those already proposed in the medical domain, assume static vector-based inputs, which cannot handle the time factor in a natural way.
* To address this, it is recommended to design deep learning approaches that can handle temporal health care data. In this aspect, Recurrent Neural Networks (RNN) is a good choice to utilize when healthcare data is sequentially ordered, as they are well designed to handle temporal dependencies [17].

## Model selection

The most commonly understood and used metric to evaluate the AI model performance for a use case is model accuracy; i.e. accuracy of the model predictions with the ground truth. Beyond that however, there are other important considerations that should be taken into account as well when selecting the AI model.

* Model validation and generalisability. Not all data are created equal. How the model was validated may be a factor to consider in the healthcare context. Healthcare data from prospective studies generally have fewer biases and are more favourable for model validation than from retrospective studies. This however remains a challenge, as the number of established prospective studies are limited. Models that are trained and validated from clinical studies spanning multiple centres are also more likely to be generalisable, allowing the model to be adopted for a different population group significant model retraining effort.
* Workflow integration and implementation. Developing a superior AI model is worthless if it doesn't fit well within the clinician's workflow and eventually ends up being side-lined; i.e. requiring additional user interface, software or hardware to integrate the AI solution within the hospital. If the AI model excessively prompts the clinician with alarms, it may lead to "alarm fatigue", leading to these alerts being ignored or switched off.
* Clinical applicability and impact. AI models are typically assessed against a baseline performance. However, for the healthcare context, such systems should be evaluated against a human clinician to determine if it does indeed result in an impact to the level of care received by a patient. Even so, a well performing model may not necessarily translate to an overall higher clinical impact. The system has to be assessed prospectively in practice to determine the level of impact it can deliver to clinicians and patients.

While there are many other important aspects that need to go into the evaluation of an AI model for healthcare deployment, such as model privacy, security, and especially ethical considerations, these will not be discussed in detail here.

# Model reporting framework

Now that the pipeline for AI model training has been described in clause 7, responsible development protocols dictate that proper recording and accounting framework be in place to ensure that the AI model is used within the boundaries that come with it. This helps ensure that users fully appreciate the AI models' purpose, performance, and particularly, limitations so that they are appropriately deployed.

This section describes a framework based on the "model cards" proposed by Mitchell et al. (2019) to encourage transparent reporting about a trained machine learning model [18]. The purpose of this reporting framework is to enable those considering the use of a specific trained model in a particular context to better understand the systematic impacts of the model before deploying it. A standardized reporting format not only makes it easier for different stakeholders to assess and compare candidate models, but also encourages forward-looking model analysis in the development phase. Such a reporting framework is particularly important in the healthcare context, where model failures or misuse can have serious repercussions

When preparing the model description, it is important to consider the different stakeholders that may rely on this information. These include other model developers, software engineers, executives, policymakers, and the general public.

The following subsections discuss various aspects of a model that should be addressed in the model reporting framework. Note that there will be cases where some of this information may be sensitive; for e.g., the amount of detail a company can disclose might be different from academic research groups.

## Basic model details

* **Persons and/or organization developing model**: Who developed the model? This can be used by all stakeholders to infer details pertaining to model development and potential conflicts of interest.
* **Funding**: Who funded the development of the model?
* **Model date**: When was the model developed? This is useful for all stakeholders to become further informed on what techniques and data sources were likely available then.
* **Model version**: Which version of the model is it, and how does it differ from previous versions? This is useful for all stakeholders to track whether the model is the latest version, associate bugs to the correct model versions, and aid in model comparisons.
* **Model type**: This includes basic model architecture details, such as Naive Bayes classifier, Convolutional Neural Network, etc. This is particularly relevant for AI practitioners to highlight what kinds of assumptions are encoded in the model.
* **Sources**: Papers or other resources with more information.
* **Citation details**: How should the model be cited?
* **License**: Licensing information as appropriate.
* **Feedback channels**: for example, an email address that people can write to for further information.
* **Model output**: How can potential users interact or integrate the AI model with existing systems or workflows? Does the model also provide unnecessary outputs beyond what is required to achieve its intent? For example, if the confidence score is not required, but provided as an output, potential adversaries may be able to reverse engineer its training data or other aspects (e.g. security, privacy) about the AI model.

## Use case and users

* **Intended uses**: What general or specific tasks was the model developed for?
* **Intended users**: Was the model developed for entertainment purposes, for hobbyists, or enterprise solutions? This helps users gain insight into how robust the model may be to different kinds of inputs.
* **Out-of-scope uses**: Similar in spirit to warning labels or disclaimers found on products. What are the known limitations of the model?

## Model performance metrics

* **Metrics**: What measures of model performance are being reported, and why were they selected? For classification systems, the error types that can be derived from a confusion matrix are useful, as their relative importance depends on context. More recently, research attention has also been directed to addressing how can AI models be evaluated for the fairness [19] or robustness [20] against adversarial attacks?
* **Thresholds**: If decision thresholds are used, what are they, and how were they chosen?
* **Uncertainty**: What metrics are used to estimate model uncertainty and variability? How were they estimated?

## Model performance factors

Which factors are being reported, and why were these chosen? What are foreseeable (other) salient factors for which model performance may vary?

* **Groups**: Groups refers to distinct categories with similar characteristics that are present in the evaluation data instances. For human-centric models, these are people who share one or multiple characteristics. Determining which groups to include depends on the intended use of the model and the context under which it may be deployed.
* **Instrumentation**: The performance of a model can vary depending on what instruments were used to capture the input data to the model.
* **Environment**: Helping the user understand how model performance depends on different environmental conditions.

## Data

All referenced datasets would ideally point to a set of documents that provide visibility into the source and composition of the dataset. Evaluation datasets in particular should include datasets that are publicly available for third-party use.

* What datasets were used to train and evaluate the model?
* **Motivation**: Why were these datasets chosen?
* **Pre-processing**: How was the data pre-processed for training or evaluation? What augmentation was performed? How was the data split for training, validation and testing subsets and assessed to still be representative of the original data?

The datasets should not only be representative of the model's typical use cases but also anticipated test scenarios and challenging cases.

## Analysis

The quantitative analyses should provide the results of evaluating the model according to the chosen metrics, providing measures or uncertainty or confidence where possible. The analysis should be broken down by the chosen factors.

* **Unitary analysis**: How did the model perform with respect to each factor?
* **Intersectional analysis**: How did the model perform with respect to the intersection of evaluated factors?

## Ethical considerations

Provide information on ethical considerations that went into model development, surfacing ethical challenges and solutions to stakeholders.

* **Data**: Is the model based on any sensitive data?
* **Risk management**: What risk mitigation strategies were used during model development?
* **Risks and harms**: What risks may be present in model usage? Try to identify the potential recipients, likelihood, and magnitude of harms.
* **Use cases**: Are there any known model use cases that are especially sensitive?
* **Review process**: Was the work reviewed by an external board?

Bibliography

1. Qayyum, Adnan & Qadir, Junaid & Bilal, Muhammad & Al-Fuqaha, Ala. (2020). Secure and Robust Machine Learning for Healthcare: A Survey.
2. Zinkevich, M. (2019). Rules of Machine Learning: Best Practices for ML Engineering.
3. Luo, Gang. (2015). MLBCD: a machine learning tool for big clinical data. Health Information Science and Systems. 3. 10.1186/s13755-015-0011-0.
4. Probst, Philipp & Bischl, Bernd & Boulesteix, Anne-Laure. (2018). Tunability: Importance of Hyperparameters of Machine Learning Algorithms.
5. Center for Open Data Enterprise (CODE), Sharing and Utilizing Health Data for AI Applications, Roundtable Report, 2019

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