|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-G-207-A02 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | Plenary | New Delhi, 13-15 November 2019 |
| **DOCUMENT** |
| **Source:** | Editor |
| **Title:** | DEL07B: AI test specification |
| **Purpose:** | Discussion |
| **Contact:** | Auss Abbood, Robert Koch Institute, Germany | Email: AbboodA@rki.de  |

|  |  |
| --- | --- |
| **Abstract:** | The present document proposes an outline for the future deliverable “AI Test Specification”. Background: The ITU/WHO Focus Group on Artificial Intelligence for Health (AI4H) has proposed a list of deliverables (at meeting “G” in New Delhi in November 2019), including this “AI Test Specification”, which belongs to a set of four deliverables under the umbrella of the “AI4H Evaluation Specification”. |

Call for Participation

If you are interested in contributing to the *AI4H Test Specification*, please contact the editor of this document (Auss Abbood, AbboodA@rki.de) and the Secretariat of the Focus Group (tsbfgai4h@itu.int) using “ITU/WHO Deliverable - AI4H Test Specification” as e-mail-subject, briefly introduce yourself, describe your relevant expertise, and explain your interest.

Expertise profile of potential contributors

Participants should be familiar with the implementation of AI algorithms and/or testing of software products.

Time plan (first draft, release 1)

TBD.

Target audience

Engineers of AI algorithms or others technical stuff who want to make sure to follow best practices in AI testing.

**Table of Contents**

1 Objectives 3

2 Functional Testing 3

2.1 Hardware and Software Compatibility 3

2.2 Pipelines 3

2.3 In- and output 3

3 Non-functional Testing 3

3.1 Performance Measures 3

3.2 Best Practices: Design 3

3.3 What not to Test 3

4 Bibliography 3

AI Test Specification (Outline)

# Objectives

This document specifies how an AI can and should be tested *in silico*. Among other aspects, best practices for test procedures known from (but not exclusively) AI challenges will be reviewed in this document. Important testing paradigms that are not exclusively related to AI applications should be mentioned too.

# Functional Testing

This part contains all relevancies related to testing the functionality of the AI algorithm

## Hardware and Software Compatibility

Especially algorithms that utilize “special” hardware/hardware specification (e.g. graphics card/16-bit floating operations) might only run (or produce the exact expected results) on an exhaustive list of hardware-software combinations.

## Pipelines

AI algorithms are most likely linked to data and therefore require a wide variety of computation until the algorithm comes into play. This process needs to be streamlined to be tested and reproduced (e.g., MLFlow, Sacred, …).

## In- and output

Certain algorithms rely on assumptions on the (input) data. To assure its functionality, input data needs to be tested for these assumptions to assure that the output of the AI works as expected including edge cases in the input.

# Non-functional Testing

While the AI algorithm might run without errors, we have certain expectations to the algorithm that target the

## Performance Measures

A learning goal is mostly linked to some metric or other performance measure (A/B testing, human-based testing, …). It is important to specify metrics and tests whether an algorithm and/or dataset achieves the required amount of data. Learning goals needs to be tested avoiding typical pitfalls such as data leakage, bias, etc…

## Best Practices: Design

Add a reference paper or implementation that shows that the planned AI system tackles the problem adequately. Much time can be wasted not handling known problems or using known solutions.

## What not to Test

Should we treat third party AI-output any differently than other third-party output?

# Bibliography

TBD. Format: APA bibliography style with DOI links. References like [Name, Year].

Suggested reading:

* Anderson‐Cook et al. (2019). How to Host An Effective Data Competition: Statistical Advice for Competition Design and Analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(4), 271-289. [[doi](https://doi.org/10.1002/sam.11404)]
* Reinke et al. (2018). How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 388-395). Springer, Cham. [[doi](https://doi.org/10.1007/978-3-030-00937-3_45)]

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