DOI 10.5162/SMSI2023/A1.2

Digital Testing Platform for Artificial Intelligence: A modular and scalable concept

M. Liebl[1], D. Hutzschenreuter[1], A. Kofler[1], C. Kolbitsch[1], S. Eichstädt[1] and F. Härtig[1]
[1] Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany
Maik.Liebl@ptb.de

Summary:
Artificial intelligence (AI) technologies in medicine require advanced testing approaches to assess and ensure the quality of applications. Here, we present a modular and scalable platform concept for quality assessment of AI technologies to accomplish this in a fast automated and digital way. We describe the design and functionality of the platform and its application to AI-based image reconstruction in accelerated magnetic resonance imaging (MRI) as a use-case example.

Keywords: Artificial Intelligence, Digital platform for AI testing, Medical AI applications

Introduction

Artificial Intelligence (AI) is a rapidly growing field of innovation which is producing powerful software solutions, e.g. in the healthcare sector. However, for their successful application in clinical routine, quality assessment is essential. For this purpose, the European Information Technology for the Future of Cancer (ITFoC) consortium demands independent tests that go beyond the common internal testing and validation during development [1]. In the future, regulation on AI-based software in the medical sector is likely as these applications are ranked as “high-risk” in the AI Act of the EU commission [2]. Thus, a key question will be how the existing quality infrastructure can be adjusted to carry out independent tests without slowing down the innovation potential.

AI testing platform (ATP)

Developing a digital AI testing platform (ATP) is an important step towards a fast and automated assessment of AI based software. The single modules of the ATP concept we propose are depicted in Figure 1. An IT platform handling client interactions, documentation and quality management forms its foundation. Five modules are connected to the IT platform that are used to perform and evaluate results of an AI software test. Data creation, e.g. using collection of existing data, provides independent, and application-specific test data serving as ground truth or ‘numerical reference artifact’. The Application Programming Interface (API) randomly picks a test data set and performs data preprocessing to emulate unknown test cases for the client. By the given ground truth, the API estimates the predefined performance metrics specified in the test criteria. Ideally, a non-AI-based reference procedure is available to determine a baseline for the performance metrics. Additionally, the test protocol is validated on AI test models.

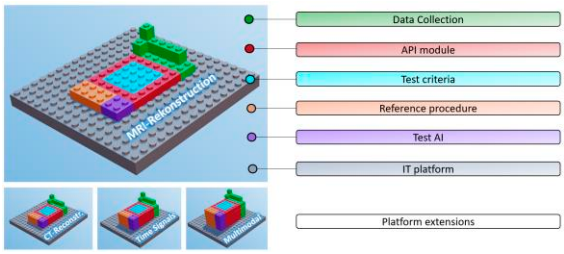

Figure 1: Modular and scalable design principle of the ATP. An IT platform carries five components: data collection, Application Programming Interface (API), test criteria, reference procedure, and test AI, required to develop and perform the AI software test.

While IT platform and API module can serve for various AI applications, the other modules are application-specific, e.g. to test AI-based image reconstruction techniques in accelerated magnetic resonance imaging (MRI). An easy and fast exchange of these modules is envisioned to tailor the AI testing platform to multiple AI-based applications as currently applied for computed tomography (CT) reconstruction, time signals as the electrocardiogram (ECG) or data sets from multiple modalities.

Service chain of the ATP

The service chain of the ATP is depicted in Figure 2. After registration on the IT platform, the client can order an AI software test. Triggered by an incoming order, the ATP provides test data for the client to download. The client uses this test data as input for the AI under test and

uploads the output to the ATP. The ATP then compares the test result with the ground truth to generate a test report. Finally, the test report is sent to the client.

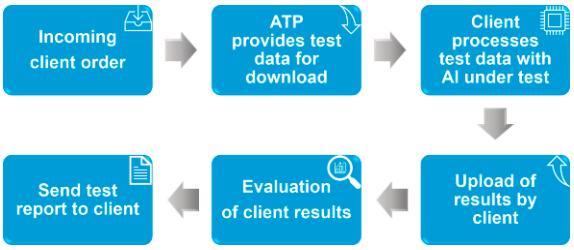

Figure 2: Service chain of the ATP. Triggered by an incoming client order, the ATP provides a test data set for the client to download. The client processes this data with his AI under test and uploads the results. The ATP then compares the test results with the ground truth and sends the test report to the client.

ATP applied to the MRI use-case

AI-based approaches enable MR image reconstruction with fewer measurements of the patient [3, 4]. This allows to speed-up the image acquisition process up to a factor of ten compared to conventional sampling schemes. Quality assessment with independent test data can help to increase the confidence of the society in AI-based methods. One possibility to accomplish this are so-called “grand-challenges” [3], that are, however, resource-binding, and competitive. With the ATP, we aim to provide an alternative way for the industry to test their AI-based software with independent data. Therefore, the concept of the ATP is applied to the example as shown in Figure 3: A data collection provides fully sampled MRI data serving as ground truth. Further, the data collection contains predefined parameters to add noise and emulate under-sampled (fast) MRI data by k-space masks.

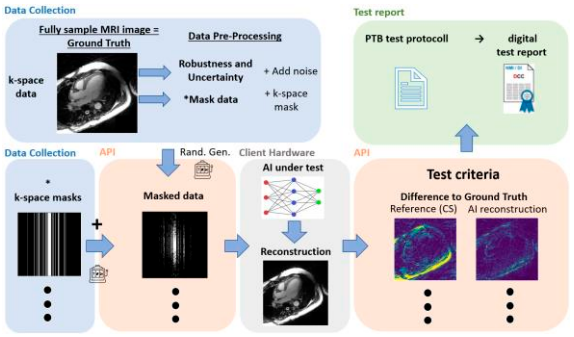

Figure 3: Workflow of the ATP applied to fast MRI reconstruction techniques. The client is provided with masked MRI data, where only the ATP knows the ground truth. By comparing the output of the AI-based client software with the ground truth, the ATP generates a digital test report listing the achieved results of the performance metrics. MRI images are adapted from Kofler et al. [4] with permission under the Creative Commons Attribution 3.0 license.

The API automatically generates the masked data using a random generator and sends this data to the client. The client reconstructs the under-sampled MRI data with their AI under test and sends the reconstructed images to the ATP. The API then calculates the difference between reconstruction and ground truth to determine the quality measures. In image reconstruction, possible performance metrics are normalized mean square error, peak signal-to-noise ratio, structural similarity, and L2 Error [3]. In this and many other AI use-cases, a non AI-based reference procedure is available to determine a baseline that the results of the AI under test should not undergo.

Finally, a test report is generated that contains the estimated results of the performance metrics for the AI under test and the reference procedure. This test report is sent to the client. After the ATP service is online, also benchmarking information might be sent confidentially to the client, e.g. the average scores obtained over all AI software tests performed so far.

Discussion and Conclusion

We presented the concept of a digital ATP that builds a framework for advanced AI software testing and quality assessment. In the design process, we considered extensions and advancements of the ATP by a modular structure. With this, we aim to provide the digital infrastructure and basis for AI software testing with independent data as currently recommended by the ITFoC.

While the ATP can provide the infrastructure, its success and possible extensions will depend on two key components: First, the expertise in the application scenarios to develop performance metrics. Second, creation of data with well characterized quality and uncertainty for each application. For future development, we plan to explore further possibilities to use existing IT platforms such as TraCIM [5] for implementing the ATP service chain.

References

[1] R. Tsopra, X. Fernandez, A. Burgin et al., A framework for validating AI in precision medicine: considerations from the European ITFoC consortium, *BMC Medical Informatics and Decision Making* 21, 247 (2021); doi: 10.1186/s12911-021-01634-3

[2] European Commission, Proposal for a Regulation of the European parliament and of the council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, (2021); https://artificialintelligenceact.eu/the-act/

[3] J. Zbontar, F. Knoll, Y. W. Lui et al., fastMRI: An Open Dataset and Benchmarks for Accelerated MRI (2019), arXiv:1811.08839v2 [cs.CV]

[4] A. Kofler, T. Schaeffter, C. Kolbitsch et al., Neural networks-based regularization for large-scale medical image reconstruction, *Physics in Medicine and Biology* 65, 135003 (2020); doi: 10.1088/1361-6560/ab990e

[5] TraCIM PTB, https://tracim.ptb.de/