Consistency of XAI Models against Medical Expertise: An Assessment Protocol

Emilien Arnaud [1,2], Mahmoud Elbattah [2,4], Amandine Pitteman [2], Gilles Dequen [2], Daniel Aiham Ghazali [1,3,5], Pedro A. Moreno-Sánchez [6]
[1] Department of Emergency Medicine, Amiens-Picardy University Hospital, Amiens, France
[2] Laboratoire MIS, Université de Picardie Jules Verne, Amiens, France
[3] Université de Picardie Jules Verne, Amiens, France
[4] College of Art, Technology and Environment, University of the West of England, Bristol, United Kingdom
[5] University of Paris Cité, Paris, France
[6] Faculty of Medicine and Health Technology, Tampere University, Finland
arnaud.emilien@chu-amiens.fr, pedro.morenosanchez@tuni.fi

***Abstract*— Despite the significant advances made by Artificial Intelligence (AI) models in enhancing medical diagnostics and prognostics, their opacity poses a hurdle to widespread clinical adoption. In this regard, Explainable AI (XAI) aims to demystify these complex models, such as neural networks, by revealing the reasoning behind predictions. However, a notable gap exists in enabling non-experts to verify these explanations, necessitating human-in-the-loop evaluation. This paper introduces a systematic protocol, including a novel "consistency" metric, to evaluate the SHAP-based explanations of XAI, comparing them against the clinical knowledge of expert clinicians. We demonstrate how this metric could facilitate both global and feature-specific analyses, operating at the level of individual instances, and thus enhancing AI transparency. It is conceived that the implications of this work may extend beyond the medical context, offering a standardized methodology that could potentially improve the interpretability and acceptance of AI systems in diverse domains.**

Keywords—Explainable AI, Interpretability, Human in the Loop, Trustworthy AI.

I. Introduction

emerged as powerful and precise methodologies for the development of computer-aided diagnostic systems, revolutionizing the approach to patient care and clinical decision-making. Despite their proven effectiveness, the underlying mechanisms of AI/ML algorithms are frequently intricate and opaque, making them challenging for users to interpret. This complexity poses a significant barrier, particularly in healthcare settings where decisions directly affect human lives. In such critical domains, the demand for transparency and understandability in the algorithmic outputs becomes paramount. Ensuring that these outputs are comprehensible not only to the technical community but also to end-users, including clinicians and patients, is crucial for fostering trust and facilitating the integration of AI/ML solutions into healthcare practices. Transparent AI/ML systems empower users by providing insight into the decision-making process, thereby enhancing their confidence in leveraging these technologies for diagnostic and therapeutic purposes.

Explainable AI (XAI) has been developed as a crucial response to the opacity of AI/ML models, which are often criticized for their "black box" nature. XAI encompasses methodologies that enable the generation of insights into the underlying factors influencing the predictions made by these models [1]. The primary objective of XAI is to enhance the transparency, accountability, and reliability of AI systems, making them more approachable, equitable, interactive, and informative [2]. Such explanations are crafted to be comprehensible to the intended audience within the healthcare sector, including both clinicians and patients, thereby bridging the gap between complex AI decision-making processes and practical clinical applications. The significance of XAI is particularly pronounced in healthcare, a sector characterized by stringent regulatory standards. For instance, the General Data Protection Regulation (GDPR) by the European Union mandates that patients are entitled to receive clear explanations regarding the automated decisions affecting their care. This regulatory context underscores the necessity for XAI in ensuring that AI-driven healthcare solutions are not only advanced and efficient but also aligned with legal and ethical standards, promoting a transparent and patient-centric approach to medical care [3].

A central ambition of Explainable AI (XAI) is to bolster the trust in AI technologies, especially in sectors that are either stringently regulated or demand high levels of precision, such as healthcare. Achieving trustworthy AI involves adhering to several key principles, including human agency and oversight, system robustness, privacy and data governance, transparency, fairness and non-discrimination, and accountability [4]. XAI systems rely on the transparency principle that advocates for the decisions made by an AI system to be understood and traced by human beings. This transparency is critical, as it allows domain experts and stakeholders to review and affirm the decisions suggested by AI, ensuring these decisions are clear, justifiable, and in line with expected outcomes.

However, the challenge often arises in the capacity of non-experts users to assess the accuracy of these explanations, highlighting the importance of establishing universally recognized metrics within the field. This necessity arises from certain fundamental characteristics of explanations that require assessment, such as the sensitivity of the explanation to variations in input, the precision with which the explanation identifies the most critical features, the veracity of counterfactual explanations, and the fairness of the model. Each of these aspects plays a crucial role in evaluating the effectiveness and reliability of explanations provided by AI systems, ensuring they are robust, equitable, and accurately reflective of the underlying decision-making process.

Furthermore, the influence of users on the assessment of XAI systems is significant, shaping the methodology for evaluation. This perspective is highlighted by Doshi-Velez et al. through the introduction of a three-tiered framework for evaluation, encompassing application-grounded, human-grounded, and functionally-grounded approaches [5]. Numerous metrics and methodologies, though not initially designed for the assessment of XAI within the healthcare context, have proven to be highly applicable to the medical field. Nevertheless, there are occasions when a more tailored, domain-specific evaluation of XAI models becomes crucial. This is largely due to the specialized knowledge that medical experts possess regarding specific use cases, which may not be readily apparent to computer scientists. Consequently, incorporating healthcare professionals into the evaluation phase of XAI systems can significantly enhance the quality and relevance of the outcomes. Their expert insights can ensure that the models are not only technically proficient but also clinically pertinent, aligning more closely with the practical needs and complexities of the healthcare domain [6].

Due to this identified gap, different metrics are emerging in the scientific community to approach better the XAI explanation to the final users. Pietilä and Moreno-Sanchez propose a taxonomy where metrics are categorized into different classes based on the domain target as well as the involvement of non-experts users in the validation of the metric. More granular classes are also proposed aligned with the aim of the metric, including robustness, faithfulness, fairness, understandability, contrastivity or sparsity [7].

The main objective of this paper is to propose an assessment metric designed to validate the explanations provided by a medical XAI system. This validation is achieved by comparing the system's explanations to the expert clinical knowledge of clinicians, who represent the potential users of the system's decision-making outputs. We seek to contribute to the ongoing efforts of developing metrics embracing the human-in-the-loop approach to validate XAI explanations in the healthcare domain.

II. Material and Methods

A. 3P-U System

The Amiens Picardy University Hospital has developed a model of "Prediction of the Patient Pathway in Emergency Department" (3P-U) to predict the patient outcomes based on triage data [8, 14]. The predictive model using the structured data is based on a Neural Network (NN), which is considered as a black-box generating technique [9]. In cases of missing data, imputation was implemented using the "physiological value" for bio-variables and the mode value for administrative variables, as outlined in Table 1. It is crucial to emphasize that, particularly in emergency medicine, the absence of certain data points does not necessarily indicate a deliberate error; instead, they are categorized as "Data Not Collected Purposely" (DNPC) [10].

Despite its application in the COVID strategy [11] to improve unit organization, the individual acceptance of 3P-U for optimizing patient pathways remains limited. A local survey identified the primary reason for this skepticism as the perceived lack of explainability in the predictions. In response, we aimed to enhance the 3P-U's interpretability by developing an XAI model, in a currently under-review article. Factors indicating a lower likelihood of admission included younger age, limb trauma presentation, FRENCH [12] level 4 or 5 classification, arrival in a personal vehicle, and a normal heart rate. However, physicians are keen to ascertain whether the automatically generated explanation aligns with what an expert could derive from their clinical knowledge.

Table 1. Description of the 3P-U Dataset

Patient Records	319,460
Years of Inclusion	2018 to 2023
9 Categorical Variables	Arrival, Gender, Origin Arrival Modality, Accompaniers, Family Status, Waiting Modality, Reason for Encounter, Circumstances
17 Numerical Variables	Age, Oxygen Flow, Heart Rate, Respiration Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Pain Scale, Temperature, Oxygen Saturation, Capillary Blood Glucose, Capillary Blood Hemoglobin, Bladder volume, Capillary Blood Ketones, Breath Test of Alcohol, Nurse Triage Scale

B. Shapley Additive Values (SHAP)

SHAP (SHapley Additive exPlanations) represents a prominent framework in the realm of interpretability, aimed at elucidating the output of complex machine learning models. This framework, a key aspect of Explainable Artificial Intelligence (XAI), was introduced by Lundberg et al. and offers a comprehensive method to dissect the influence of individual features on the predictions made by models. SHAP values facilitate a granular understanding of how each feature contributes to specific predictions, shedding light on the rationale behind a model's decisions for individual data points. Additionally, SHAP provides an overarching analysis of a model's behaviour by aggregating the impact of features across all predictions, which is instrumental in uncovering broad patterns and insights within the dataset. Beyond offering insights into model behaviour, SHAP values are crucial for comparing and evaluating models based on how their predictions are influenced by different features, thus playing a significant role in model selection and refinement.

The foundation for calculating SHAP values is based on game theory's Shapley value formula, which ensures an equitable and mathematically sound attribution of contributions among all input features. The Shapley value is calculated as follows:

$$SHAP\ value = \sum_{S \in N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

In this formula, N is the set of all features, S is a subset excluding the feature 'i' for which the SHAP value is calculated. *f(S)* is the model's output with only features in S, *and* $f(S \cup \{i\})$ is the output when feature 'i' is added to S. This approach precisely allocates prediction output among input features, revealing their relative importance in the model's decision-making [13].

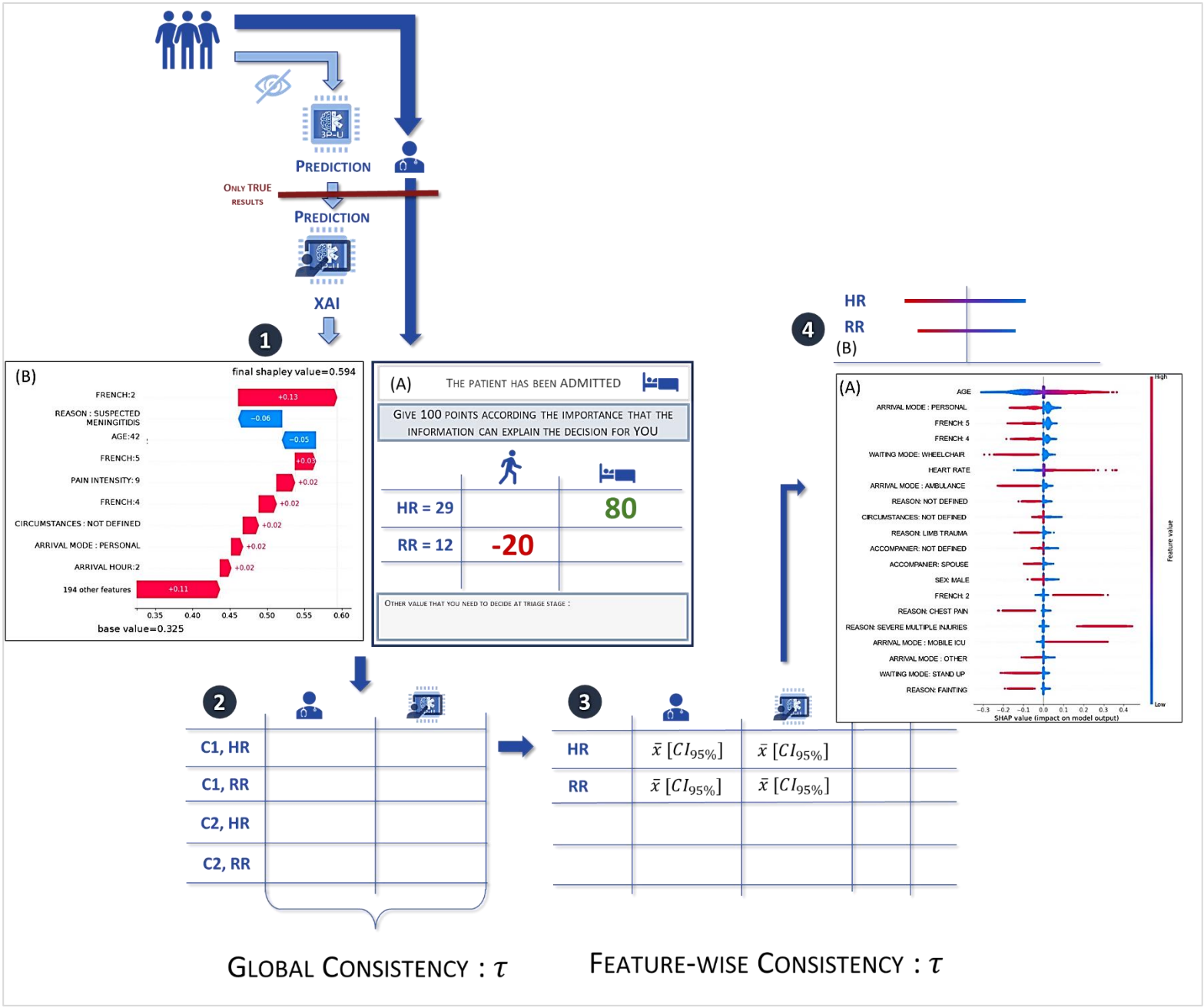

Figure 1. Protocol in four steps: (1) Scoring by physicians: 1A with XAI, 1B by physicians (2) Global consistency (3) Feature-wise consistency (4) Dot plot for visual comparison: 4A generated with physicians' scores, and 4B by XAI.

C. Consistency Metric

We aim to establish a quantitative metric, referred to as "consistency," to measure the disparity between explanations provided by clinicians and the 3P-U system. Our assessment specifically centres on cases where 3P-U aligns with the final medical decision, such as instances resulting in the admission of patients to the emergency unit.

On one hand, we request physicians to distribute a total of 100 points across the set of features identified by 3P-U as significant for an individual explanation. This distribution involves evaluating whether each feature value positively or negatively contributes to the likelihood of admission, as depicted in Figure 1-1A. On the other hand, we generate instance explanations represented by SHAP values arranged in a "waterfall" format (Figure 1-1B). The y-axis encompasses all the relevant features influencing the prediction. The SHAP values for each feature are then added to the base SHAP value, with red indicating a positive contribution to the probability of admission and blue indicating a negative contribution.

To facilitate a direct comparison, we normalize the SHAP values to the same range (100 points) as the physician's explanation using Equation 1. At this stage, we have established two datasets: one consisting of the scaled SHAP values, and the other reflecting the physician's expertise.

$$\sum \square\, abs(SHAP_{values}) \Leftrightarrow \underset{\square}{100}\ points \quad (1)$$

D. Integral Assessment of Explanations

To evaluate the overall consistency of the XAI explanation for an individual prediction, we conduct a paired-student test comparing scores assigned by physicians and the XAI on pairs of instances-features (Figure 1-2). At this stage, we globally compare the distribution of scores provided by XAI and physicians. If the test reveals no significant difference, we cannot reject the null hypothesis that the explanation from XAI is essentially the same as the physician's (and we can conclude here). However, if the test indicates a significant difference between the two groups ($p < 0.05$), rejecting the null hypothesis, we proceed to assess consistency for each feature considered in the explanation.

E. Feature-Based Assessment of Explanations

If a statistically significant difference emerges between the explanations provided by the 3P-U system and clinical experts concerning an individual instance prediction, we proceed to scrutinize the consistency of each feature included in the explanation. To evaluate the feature-wise consistency of the XAI explanation, we conduct a paired-Student test for each feature, comparing the distribution of explanation points given by the 3P-U system (scaled SHAP values) and those provided by physicians (Figure 1-3). Similar to the assessment of global consistency for an individual instance, for features where the test reveals no significant difference (thus unable to reject the null hypothesis), we can assert that the explanation generated by the XAI technique applied to 3P-U for these features is significantly consistent with the physician explanations, and vice versa. Significant differences will highlight features contributing to the disparities between the two global explanations.

F. Comparative Analysis of XAI and Clinical Perspectives

The global XAI explanation is visually depicted through a beeswarm plot, where all features are listed on the y-axis. The feature's contribution to the probability of admission is given along the x-axis, with negative contributions on the left and positive contributions on the right. Additionally, the feature's value is color-coded, ranging from low (blue) to high (red). For example, a higher age (in red) corresponds to an increased probability of admission (on the right), while arriving with personal mode (also in red) negatively contributes (on the left) to the probability of admission (Figure 1-4B).

III. DISCUSSION

The proposed protocol aims to evaluate the accuracy of XAI in comparison to medical expertise and address the fundamental question: Does the XAI explanation align with physicians' explanations? A prospective study is scheduled at Amiens Picardy University Hospital, where the XAI results will be compared with explanations from multiple physicians across different centers. However, several questions remain open within this protocol including:

- How many clinicians are required? We anticipate involving at least eleven physicians from a minimum of two centers, representing an ambitious goal.
- How many cases should each clinician assess to identify differences effectively? We aim for each physician to evaluate a minimum of twenty cases, also considered an ambitious target.
- How can a consensus among multiple clinicians be reached for a single case? Our plan involves utilizing the average score for each feature in every case.
- In the event of misalignment between XAI and physicians' explanations, is it feasible to incorporate this knowledge into a unified explainability model?

IV. CONCLUSIONS

In this study, we outlined the preliminary stages of a systematic approach for evaluating the alignment between XAI-based insights and those offered by medical professionals. The prospective study at Amiens Picardy University Hospital will aim to rigorously assess the XAI's capacity to provide explanations comparable to expert medical knowledge.

The potential of our work lies in bridging the gap between XAI and medical expertise. By emphasizing the need for refining XAI systems, we strive to enhance decision support in healthcare. As our study progresses, we remain committed to refining our protocol and contributing valuable insights to the evolving landscape of XAI in medicine.

Beyond its immediate medical application, the proposed protocol could have broader implications. It can contribute to the understanding of XAI's interpretability, acting as a foundational step in bridging the gap between AI and expert knowledge across diverse domains.

REFERENCES

[1] A. Barredo Arrieta *et al.*, « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », *Inf. Fusion*, vol. 58, p. 82-115, juin 2020, doi: 10.1016/j.inffus.2019.12.012.

[2] R. R. Hoffman, S. T. Mueller, G. Klein, et J. Litman, « Metrics for Explainable AI: Challenges and Prospects », févr. 2019.

[3] M. Mourby, K. Ó Cathaoir, et C. B. Collin, « Transparency of machine-learning in healthcare: The GDPR & European health law », Comput. Law Secur. Rev., vol. 43, p. 105611, nov. 2021, doi: 10.1016/j.clsr.2021.105611.

[4] European Commission Expert Group, « Ethics guidelines for trustworthy AI | Shaping Europe's digital future », avr. 2019. Consulté le: 21 février 2024. [En ligne]. Disponible sur: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[5] F. Doshi-Velez et B. Kim, « Towards A Rigorous Science of Interpretable Machine Learning ». arXiv, 2 mars 2017. doi: 10.48550/arXiv.1702.08608.

[6] R. V. Zicari et al., « Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier », Front. Hum. Dyn., vol. 3, 2021, Consulté le: 25 février 2024. [En ligne]. Disponible sur: https://www.frontiersin.org/articles/10.3389/fhumd.2021.688152

[7] E. Pietilä et P. A. Moreno-Sánchez, « When an Explanation is not Enough: An Overview of Evaluation Metrics of Explainable AI Systems in the Healthcare Domain », in MEDICON'23 and CMBEBIH'23, A. Badnjević et L. Gurbeta Pokvić, Éd., in IFMBE Proceedings. Cham: Springer Nature Switzerland, 2024, p. 573-584. doi: 10.1007/978-3-031-49062-0_60.

[8] E. Arnaud, M. Elbattah, M. Gignon, et G. Dequen, « Deep Learning to Predict Hospitalization at Triage: Integration of Structured Data and Unstructured Text », in Proceedings of 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA: IEEE, dec. 2020, p. 4836-4841. doi: 10.1109/BigData50022.2020.9378073.

[9] A. I. F. Poon et J. J. Y. Sung, « Opening the black box of AI-Medicine », J. Gastroenterol. Hepatol., vol. 36, no 3, p. 581-584, mars 2021, doi: 10.1111/jgh.15384.

[10] E. Arnaud, M. Elbattah, C. Ammirati, G. Dequen, et D. A. Ghazali, « Predictive models in emergency medicine and their missing data strategies: a systematic review », Npj Digit. Med., vol. 6, no 1, Art. no 1, févr. 2023, doi: 10.1038/s41746-023-00770-6.

[11] E. Arnaud, M. Elbattah, C. Ammirati, G. Dequen, et D. A. Ghazali, « Use of artificial intelligence to manage patient flow in emergency department during the COVID-19 pandemic: a prospective, single-center Study », Int. J. Environ. Res. Public. Health, vol. 19, no 15, p. 9667, août 2022, doi: 10.3390/ijerph19159667.

[12] P. Taboulet et al., « Triage with the French Emergency Nurses Classification in Hospital scale: reliability and validity »:, Eur. J. Emerg. Med., vol. 16, no 2, p. 61-67, avr. 2009, doi: 10.1097/MEJ.0b013e328304ae57.

[13] S. M. Lundberg et S.-I. Lee, « A Unified Approach to Interpreting Model Predictions », in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Consulté le: 15 décembre 2022. [En ligne]. Disponible sur: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[14] E. Arnaud, M. Elbattah, M. Gignon, et G. Dequen, « NLP-based prediction of medical specialties at hospital admission using triage notes », in Proceedings of IEEE 9th International Conference on Healthcare Informatics (ICHI), IEEE, 2021. p. 548-553. doi: https://doi.org/10.1109/ICHI52183.2021.00103.