Contribution ID: **93** Type: **Contributed Oral**

Advancing Interpretability in Artificial Intelligence Disruption Prediction Models: A Cross-Tokamak Perspective

Tuesday, 3 September 2024 16:45 (25 minutes)

Artificial Intelligence (AI) techniques, such as Machine learning and Deep Learning, have been extensively investigated for the construction of disruptions predictive models in tokamaks. Although the excellent performance has demonstrated the applicability of the paradigm to the experimental machines currently in service, the development of cross-tokamak models is still in its infancy [1]. These models, trained on existing machines, should be applied to newly constructed or next-generation machines. To achieve this ambitious goal, several issues still need to be resolved, including the unavailability of experimental data for newly constructed machines and the scalability of the data available from existing machines to new ones. The choice of the appropriate AI model also plays a fundamental role in the transportability of a predictor in a cross-tokamak approach.

For JET the literature has proposed AI supervised approaches such as Support Vector Machines [2-4], Decision Trees [5,6], Convolutional Neural Networks (CNN) [7-10], and unsupervised Manifold Learning approaches such as Generative Topographic Maps (GTM) [11-13] and Self-Organizing Maps (SOM) [14-16]. However, in both these supervised and unsupervised approaches, it was necessary to provide information on the duration of the pre-disruptive phase in the disrupted discharges of the training set during model training. This information could only be obtained heuristically [2-7] or statistically [13]. Recently, the authors of this contribute proposed, for JET, an unsupervised predictor [17], based on SOMs that does not require this information, relying only on the termination condition of the discharge (disrupted or regularly terminated), with undeniable advantages for the transportability of the model to another machines.

In this work the authors will initially conduct an analysis of their AI disruption predictors developed for JET [8-10,12-15,17] with a view towards developing a cross-tokamak model, highlighting their strengths and limitations.

Another aspect of fundamental importance for the use of AI models in control systems is their interpretability and their ability to identify the chain of events leading to disruption, so that appropriate avoidance actions can be implemented. The predictors presented in the literature so far also differ in this aspect. Manifold learning methods, such as SOMs and GTMs, allow to track the trajectory of a new discharge on the 2D map of machine's operational space, associating each time instant with the risk of disruption and, potentially, with the type of preceding event. On the other hand, despite their undeniable proficiency in predicting disruptions, deep neural network models exhibit a black-box behavior. However, recently, Explainable AI techniques [18] have emerged, aimed at interpreting the model's responses in relation to its input. In [19] a contribution to the explainability of a CNN predictor for JET was proposed.

The current work will thoroughly delve into these aspects and briefly introduce an innovative method for extracting rules from the SOM map of JET's operational space, facilitating a clear interpretation of the model's decisions throughout the discharge evolution.

Speaker's title

Ms

Speaker's email address

giuliana.sias@unica.it

Speaker's Affiliation

Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

Member State or IGO

Italy

Primary authors: Prof. FANNI, Alessandra (Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy); Prof. CANNAS, Barbara (Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy); Dr AIMERICH, Enrico (Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy); Dr PISANO, Fabio (Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy); Prof. SIAS, Giuliana (ENEA -University of Cagliari)

Presenter: Prof. SIAS, Giuliana (ENEA -University of Cagliari)

Session Classification: Prediction & Avoidance

Track Classification: Prediction and Avoidance