INTERPRETABLE NEUTRON SPECTRUM RECONSTRUCTION BASED ON TWO-STAGE ANFIS LEARNING WITH SHAP REGULARIZATION

Chizhov K.A., Lebedev A.D.[1], Trofimov Yu.V.[2], Ilyin A.S.[1], Lebedev M.D.[3]

Joint Institute for Nuclear Research, Russia, 141980, Dubna, Joliot-Curie str., 6;
kchizhov@jinr.ru
[1]Dubna State University, Russia, 141980, Dubna; lad.24@uni-dubna.ru,
a.ilin@innopolis.university
[2]MLIT JINR, Russia, 141980, Dubna, Joliot-Curie str., 6; ura_trofim@bk.ru
[3]NUST MISIS, Russia, 119049, Moscow; lebedevmisha2003@yandex.ru

Reconstruction of the neutron energy spectrum from Bonner sphere detector readings is an ill-posed inverse problem critical for radiation safety, dosimetry, and experimental physics. Classical unfolding methods are sensitive to a priori assumptions, while neural network approaches often act as a ”black box”. Fuzzy-neural models like ANFIS possess a transparent rule structure, but without special tuning, they may utilize physically uninterpretable dependencies and get stuck in local minima.

The paper considers the problem of spectrum reconstruction using ten Bonner sphere channels (input vector $\mathbf{x} \in \mathbb{R}^{10}$) predicting the distribution over 60 energy bins ($\mathbf{y} \in \mathbb{R}^{60}$). An adaptive neuro-fuzzy inference system (ANFIS) with Gaussian membership functions and a linear part in the final layer is used as the base model. A two-stage learning pipeline is proposed, combining global parameter and structure optimization with the introduction of explainable constraints on feature usage.

In the first stage (Vanilla), ANFIS parameters are tuned using the Particle Swarm Optimization (PSO) method. This stage ensures stable initialization in a multimodal parameter space and forms a meaningful fuzzy rule base. In the second stage, gradient fine-tuning is performed using SHAP-like feature importances as a regularizer: a penalty on the SHAP weight profile is added to the main MSE function.

Experimental studies were carried out on real neutron spectrometry data using Bonner spheres. The ”vanilla” model (first stage without SHAP regularization), achieving $R^2_{\text{weighted}} \approx 0.86$ on the test set, is considered as a baseline. The proposed two-stage method with SHAP regularization demonstrates comparable accuracy ($R^2_{\text{weighted}} \approx 0.84$) with additional interpretability properties. Thus, interpretability becomes an intrinsic property of the trained model, rather than just a post-hoc ”black box” analysis.

The work was carried out within the framework of the state assignment of the Ministry of Science and Higher Education of the Russian Federation (theme No. 124112200072-2).