DECODING THE CAUSES OF HIGH-DENSITY DISRUPTION THROUGH INTERPRETABLE MACHINE LEARNING

Chengshuo Shen, Mingqiao Wen, Weijie Lin, Li Gao, Wei Zheng, Yonghua Ding and Zhongyong Chen

IFPP, HUST, Wuhan, China

shenchengshuo@hust.edu.cn

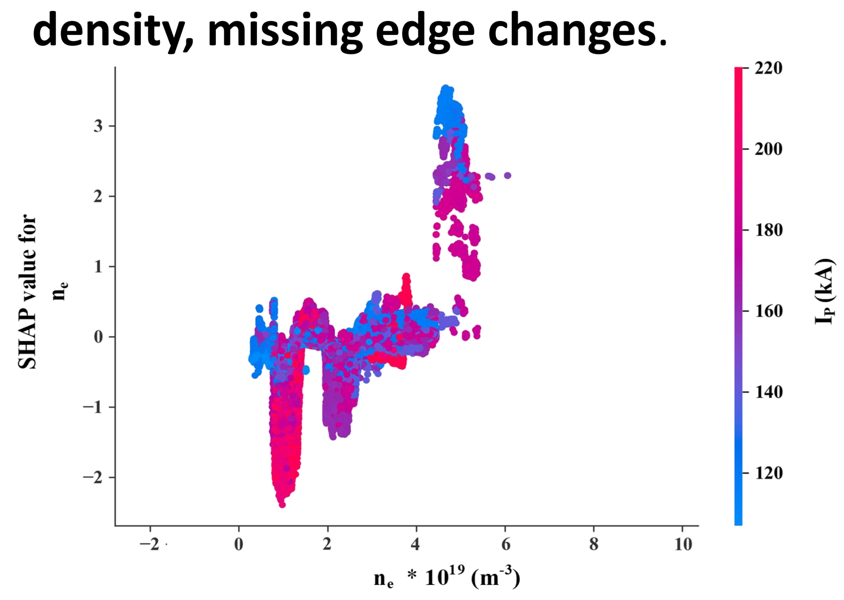
ID: IAEA-CN-123/45

A Physics-Guided Approach to Disruption Prediction

While machine learning enables accurate disruption prediction, poor interpretability limits physical insight and model transfer. We propose a hierarchical model for density-limit disruptions, **replacing Greenwald scaling with physics-guided features. SHAP analysis identifies edge density asymmetry and fluctuations as key drivers**.

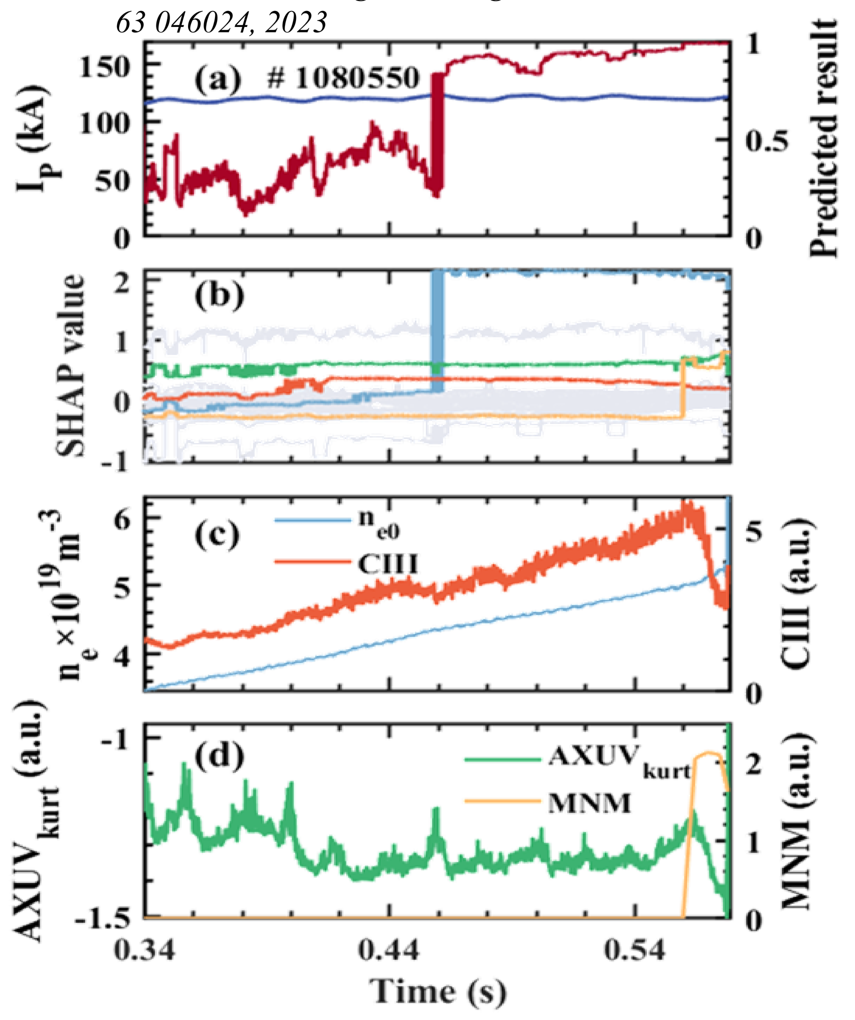
The limitations of experience-based calibration

- IDP-PGFE, which performs well on J-TEXT, **may have internalized the Greenwald scaling**.
- However, in RMP experiments, the model **focuses too much on core density, missing edge changes**.



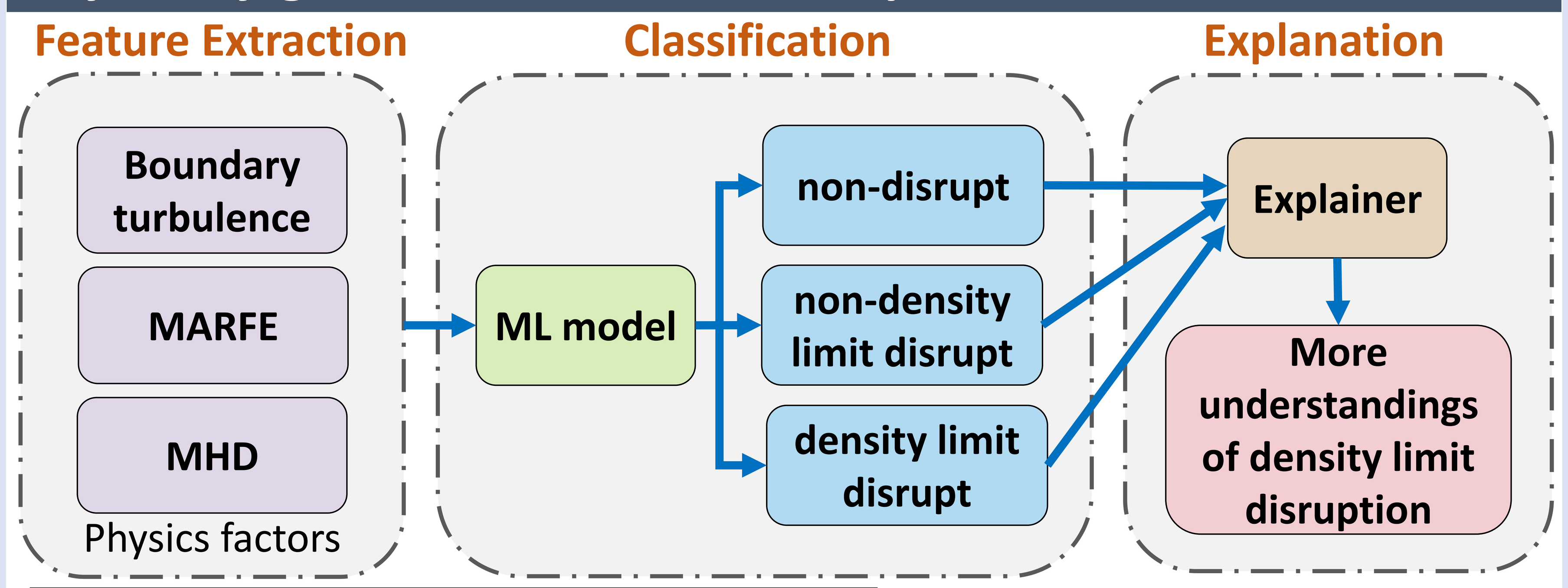
CS. Shen , W. Zheng, 49th EPS, 2023

CS. Shen , W. Zheng, YH. Ding et al. Nuclear Fusion, 63 046024, 2023



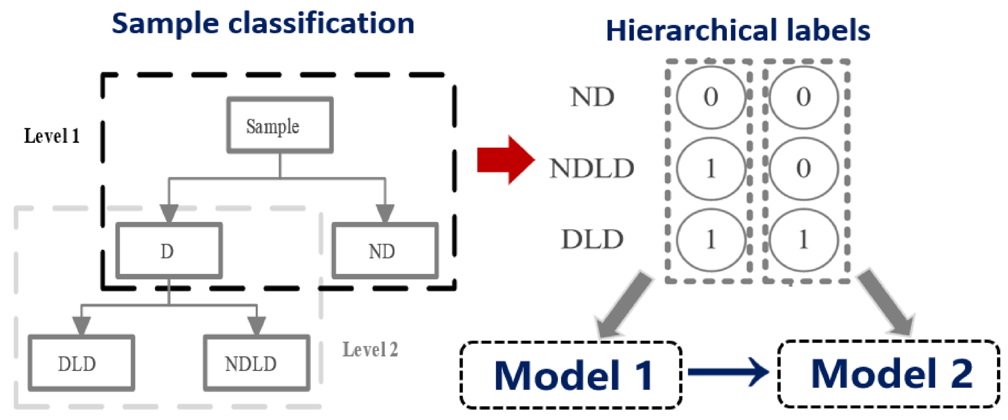
- **Greenwald scaling does not reflect the intrinsic physics of density-limit disruptions. Disruptions may occur before or beyond the limit.**
- **Can a machine learning model predict disruptions without relying on core density? And can it further distinguish density-limit disruptions from other types?**

Physically-guided hierarchical interpretable model



Physics Relation	Feature Name	Physical Meaning
MARFE	CIIIAsym (95/82/70)	Asymmetry of CIII Radiation
	HαAsym (95/82/70)	Asymmetry of Hα Radiation
Density Fluctuations	DensAsym (95/82/70)	Asymmetry of Line-Integrated Density
	Den_ngrad	Line-Integrated Density Normalized Gradient
	DenFlu_int (70,60)	Standard Deviation of Density Fluctuations
MHD	DensFlu_fre (70,60)	Density Fluctuations Frequency
	DensFlu_amp (70,60)	Density Fluctuations Amplitude
	MHD_fre	Mirnov Probe Frequency
	MHD_amp	Mirnov Probe Amplitude
	MNM	Average Poloidal Mode Number
PCS	bt	Toroidal Field
	dx	Plasma Horizontal Displacement
	dy	Plasma Vertical Displacement

- Develop a **hierarchical classification model** for disruption prediction
- Build a **SHAP-based interpreter** for the model architecture.



- **Avoiding core density as an input feature**
- **Incorporating physics-guided features such as MARFE, density fluctuations, and MHD activity.**

Hierarchy-aware loss function

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{(1)}\log\left(\hat{y}_i^{(1)}\right) + \left(1-y_i^{(1)}\right)\log\left(1-\hat{y}_i^{(1)}\right)\right) + \alpha \cdot \text{Penalty}$$

$$\text{Penalty} = \begin{cases} 1.5 \times L_1, & \text{if } y_i^{(1)} = 0 \text{ and } \hat{y}_i^{(1)} > 0.5 \\ L_1, & \text{otherwise} \end{cases}$$

Hierarchical accuracy rate

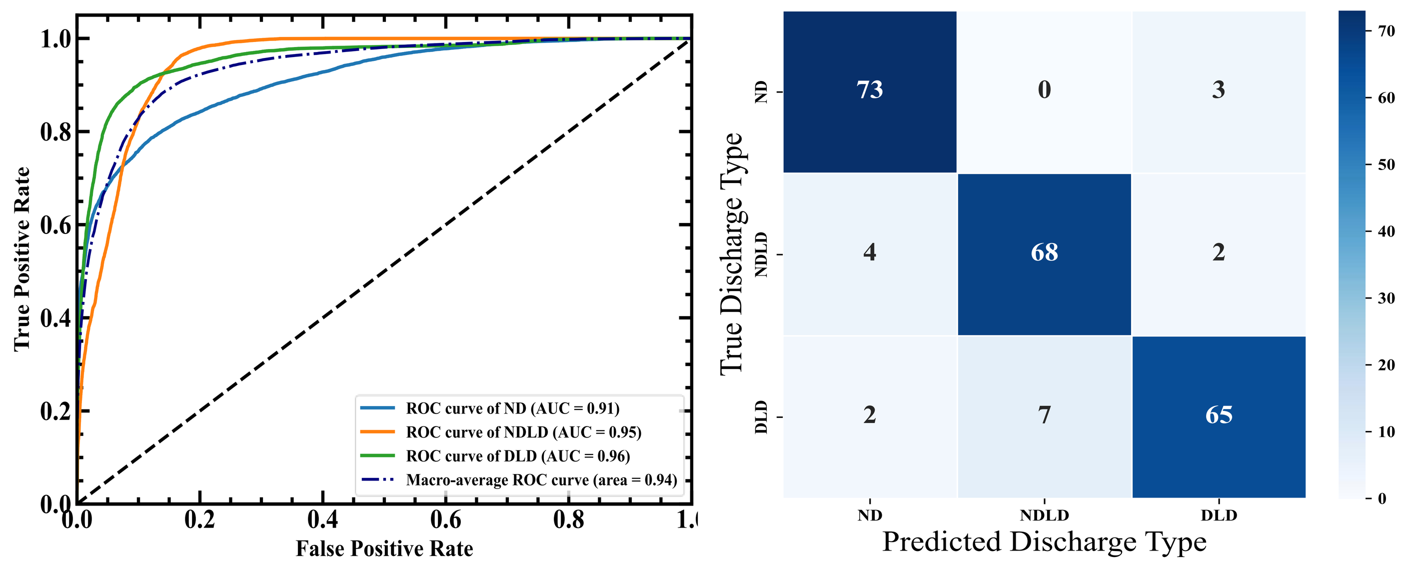
$$\text{HierarchicalAccuracy} = \frac{1}{N}\sum_{i=1}^{N} 1\left(\hat{y}_i^{(1)} = y_i^{(1)} \text{ and } \hat{y}_i^{(2)} = y_i^{(2)}\right)$$

	Shot No. of ND	Shot No. of NDLD	Shot No. of DLD
Training	262	254	253
Validation	38	36	36
Test	75	73	72

Model results and interpretability analysis

Model results

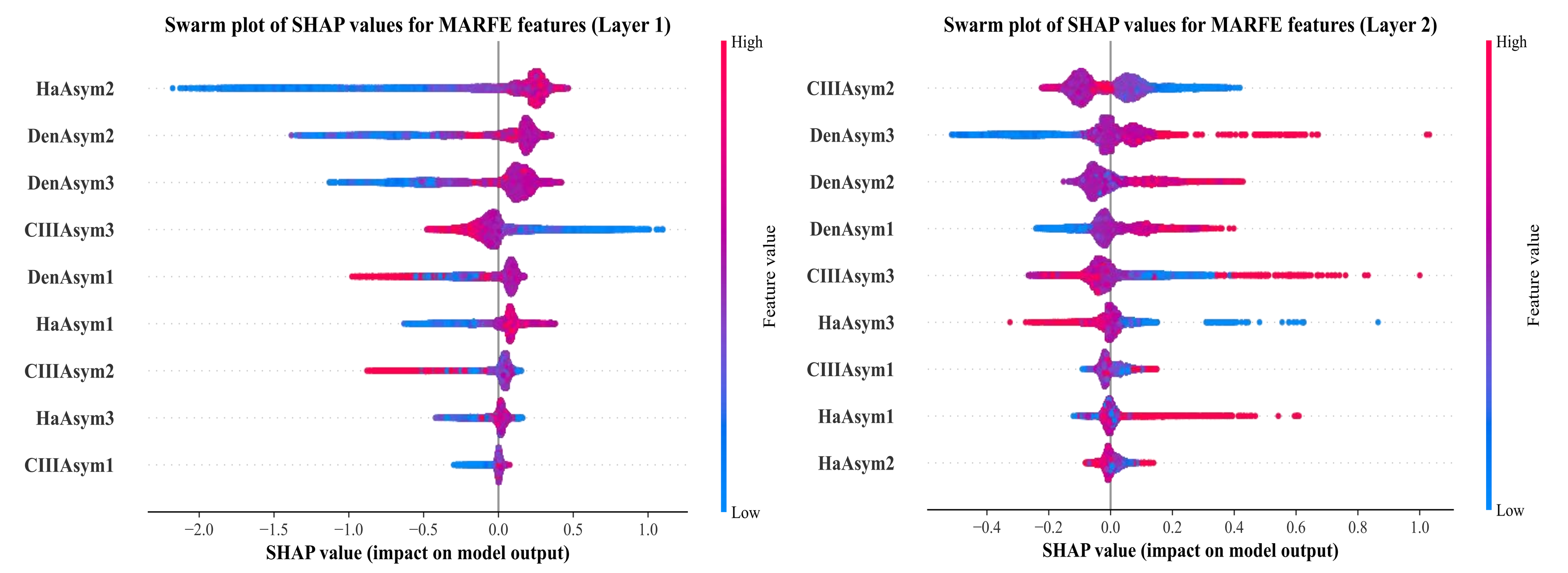
- One-vs-Rest ROC shows strong and balanced performance on all three classes.
- Confusion matrix indicates high and consistent accuracy across all discharge categories.



Prediction accuracy
- ND: 96.1%(73/76)
- NDLD: 91.9%(68/74)
- DLD: 87.8%(65/74)

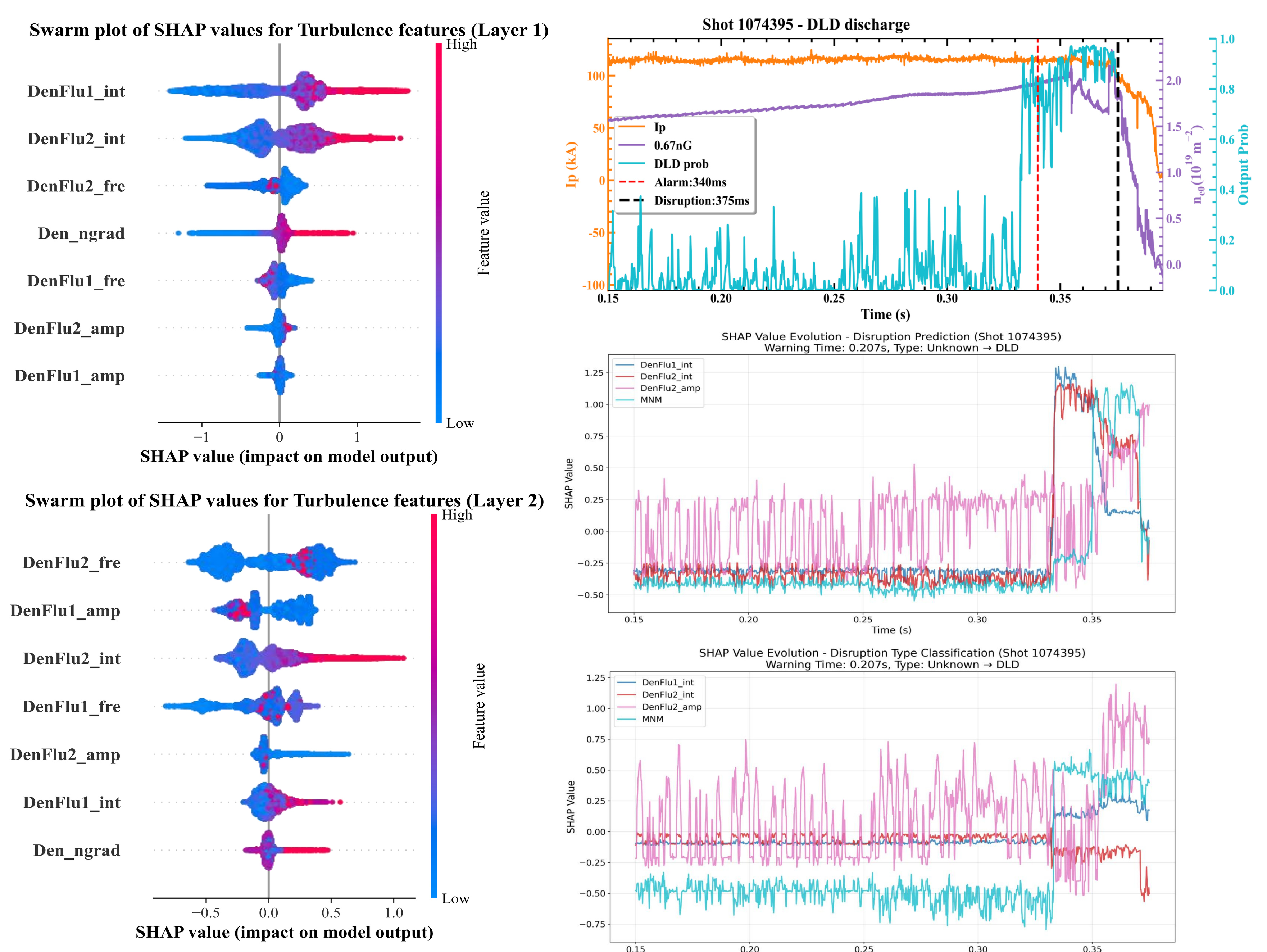
Research on interpretability

- **Edge MARFE may have limited impact on disruption onset, while stronger density asymmetry increases the likelihood of density-limit disruptions.**
- **CIII radiation asymmetry mitigates disruption prediction, in contrast to density asymmetry which enhances it—revealing competing roles in the process.**



1–3 denote 0.95a, 0.82a, and 0.7a on high/low-field sides

- **Stronger density fluctuations and steeper gradients raise disruption risk**, **reflecting turbulence-driven destabilization.**
- **Density-limit disruptions are identified by stronger fluctuations or higher gradients.**
- **Inward-shifted density fluctuations play a key role in triggering density-limit disruptions.**



1 and 2 denote 0.7a and 0.6a, near the q = 2 surface.

CONCLUSION

- **An interpretable hierarchical model is developed to classify DLD, NDLD, and ND, replacing the Greenwald fraction with physics-guided features.**
- The model achieves strong performance on J-TEXT data, **with 96.0% accuracy and a macro-average AUC of 0.94.**
- **SHAP analysis reveals that edge density asymmetry and turbulence near 0.6a–0.7a are key drivers of density-limit disruptions, while CIII asymmetry has a stabilizing effect.**
- **Physics-guided Machine learning offers reliable prediction and insight beyond empirical scaling.**

ACKNOWLEDGEMENTS / REFERENCES

- The authors would like to acknowledge the help from J-TEXT team. This work was supported by No. 2024YFE03230100, No. 2022YFE03040004, No. 2022EHB003, No. 2024040701010040, No. 12375219 and No. T2422009.