CONFERENCE PRE-PRINT

DEVELOPING OPEN MACHINE LEARNING BENCHMARKS FOR TOKAMAK EVENT PREDICTION FROM MAST

P. Sharma, S. Jackson, N. Cummings, C. J. Ham, J. Hodson, A. Kirk, K. Lawal, D. Ryan, S. Pamela, and The MAST Team
United Kingdom Atomic Energy Authority (UKAEA)
Culham Science Centre, Abingdon, United Kingdom
Email: prakhar.sharma@ukaea.uk

E.d.D Zapata-Cornejo
Plasma Science and Fusion Center
Massachusetts Institute of Technology, Cambridge, United States

Abstract

Advances in fusion research require the reliable prediction and identification of critical tokamak events such as disruptions, magnetohydrodynamic (MHD) modes, confinement modes, and edge-localised modes (ELMs). Machine learning offers powerful tools to exploit large volumes of diagnostic and operational data for these tasks, yet the absence of common reference models has limited systematic evaluation and comparison across studies. This work describes progress towards the development of baseline models applied to the historical record from the MAST tokamak. The baselines address four representative tasks: disruption prediction, MHD mode segmentation from spectrogram data, confinement mode classification, and ELM spike identification. Each task illustrates specific challenges including data imbalance, multi-modal diagnostics, and the need for evaluation metrics aligned with operational priorities such as predictive accuracy and alarm-time optimisation. The baselines provide an accessible starting point for both fusion researchers and data practitioners. Initial results demonstrate the potential of these approaches to improve operational reliability and event characterisation, while highlighting methodological gaps that motivate further work. By establishing shared baselines rather than definitive benchmarks, the study provides a foundation for future community-driven benchmarking efforts and contributes toward accelerating progress in predictive modelling for fusion energy.

11. INTRODUCTION

Fusion energy is being developed as a long-term solution for clean and reliable electricity. To achieve this goal, future devices must be able to operate plasmas safely and with high reliability. A central challenge is the ability to predict and classify important plasma events. These include disruptions, magnetohydrodynamic (MHD) modes, confinement mode classification, and edge-localised modes (ELMs) spike identification. Such events influence plasma control, confinement quality, and machine protection [1].

In recent years, progress has been made in open access to fusion data. The FAIR-MAST data portal was introduced as a public release of the MAST tokamak database, built on the FAIR principles of findability, accessibility, interoperability, and re-usability [2,3]. FAIR-MAST provides public APIs, searchable metadata, and high-performance object storage that together allow efficient remote access to experimental signals. To our knowledge, this is one of the first major tokamak datasets to be openly released, with the only other large dataset of similar scale currently provided by the Large Helical Device (LHD) in Japan [4]. While FAIR-MAST makes the data accessible through APIs and metadata search, the signals are still not "AI ready" and require some further transformation, re-gridding, and consolidation for downstream use. Furthermore, the relevant metadata annotations for downstream tasks are missing and need to be curated. For this reason, we develop a processing pipeline to provide curated annotations for downstream modelling.

Machine learning has become a promising approach for prediction and classification of plasma events. It can process large volumes of diagnostic signals and identify patterns that may be difficult to capture with physics-based models alone [5,6]. Many studies have explored this potential, but progress is slowed by the lack of shared pipelines and standard evaluation procedures. Different groups often use their own datasets, most of which are not openly available, along with different preprocessing choices and metrics. This makes it difficult to compare results or build upon earlier work.

This paper presents progress towards the development of a set of baseline benchmark machine learning models trained on open data from the MAST tokamak. So far, we consider four tasks that represent key phenomena related to the operation of plasma in Tokamak devices. These are disruption prediction, MHD mode segmentation,

confinement mode classification, and ELM spike identification. Each task is associated with specific difficulties such as unbalanced data, noisy signals, or the need for application specific evaluation measures.

The aim of this work is to provide clear starting points that can be reproduced and extended by other researchers. The baselines are supported by documentation and processed datasets that allow a model to be applied directly. This creates a foundation for future shared benchmarks and encourages collaboration between the fusion and machine learning communities, while contributing to the broader goal of reliable predictive modelling for fusion energy.

12. METHODS AND DATA

This section describes the data used in this study and the four machine learning tasks considered. Each task is motivated by its relevance to tokamak operation and is linked to specific challenges in modelling and evaluation.

2.1 Disruption prediction

A disruption is the uncontrolled loss of plasma confinement. It is a critical event in a tokamak, as it can cause damage to the vessel and supporting systems. Disruptions can be identified by signatures such as magnetic instabilities, loss of plasma current, reduction in density or pressure, or changes in plasma shape as it moves towards the wall. A typical discharge can be described in four stages: plasma breakdown, current ramp-up, flat top, and termination. A disruption may occur during the termination stage [7].

For machine learning models, the key requirement is to give an alarm time. Alarm time is the lead time between the model first predicting instability and the actual disruption [8]. This warning period allows control systems to take action before confinement is lost. To be useful in operation, the model's inference must be faster than the diagnostic time window it analyses, so that predictions are delivered in real time or ahead of the event.

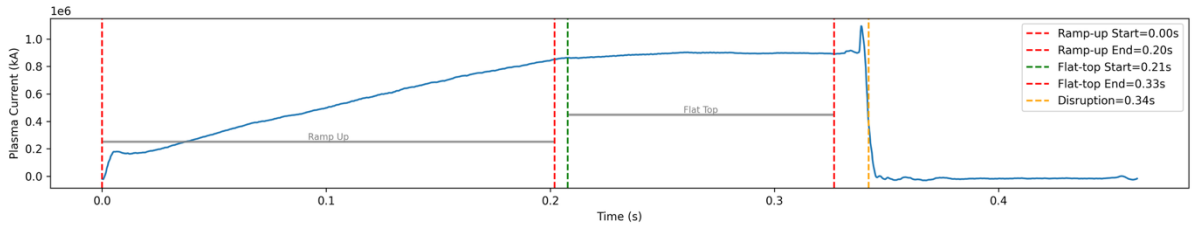

FIG. 1. Example plasma current trace from MAST shot 30109, showing disruption time (orange). Additional intervals (ramp-up, flat-top) are shown for illustration but not used in the present study.

The preliminary disruption prediction task is based on 417 MAST shots. The disruption times are not hand-labelled but are derived automatically using an algorithm for peak detection applied to the plasma current signal [9]. Plasma current trace from shot 30109 with the automatically detected disruption time is shown in Fig. 1. A set of diagnostic signals is used as model inputs: plasma current, line-averaged electron density at the plasma core, line-averaged electron temperature at the plasma core, internal inductance, radiated power from poloidal bolometer arrays, and the D-alpha signal from a tangential mid-plane plasma view. The task is formulated as a binary classification problem, with labels indicating either stable operation (0) or a disruptive state (1). Predictions are produced at each time step within the input sequence.

For this baseline, the data is transformed into a sliding window dataset. Each window contains 200 time steps, with a stride of 100 time steps between consecutive windows. This formulation provides overlapping samples and improves temporal coverage near disruption events. Since disruption windows are under-represented compared with non-disruptive ones, a weighted random sampler is applied during training to balance the classes.

A recurrent neural network with long short-term memory (LSTM) units is used as the baseline model. Several variants are explored, including normal and stacked LSTM layers, unidirectional and bidirectional configurations, and different dropout rates. Both cross-entropy and focal loss are tested as objective functions. The effect of mini-batch versus full-batch training is also evaluated. Hyperparameters such as the lead time before disruption (10 ms, 30 ms, and 60 ms) are included in the comparison.

2.2 MHD mode segmentation

Magnetohydrodynamic (MHD) instabilities are large-scale plasma deformations that grow over time and can reduce performance or trigger disruptions. They arise from the nonlinear MHD equations but are often studied through their experimental signatures [10]. In this work, the focus is not on predicting the theoretical MHD eigenmodes, but on identifying experimentally observed instabilities in diagnostic data.

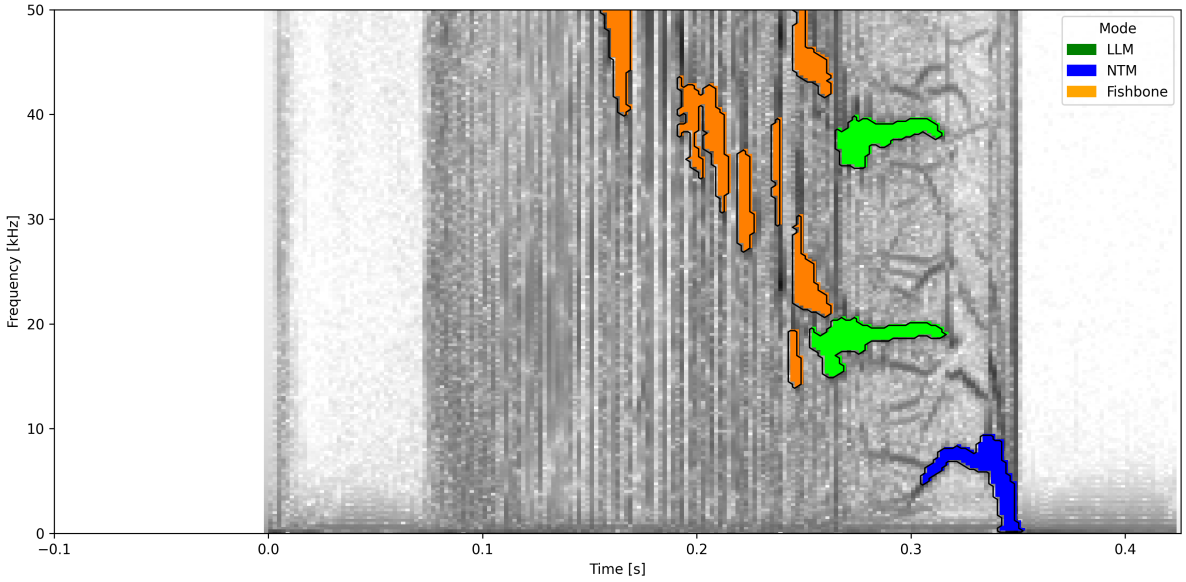

FIG. 2. Example spectrogram from MAST shot 23447, showing hand-labelled MHD activity. Fishbones are marked in orange, neoclassical tearing modes (NTM) in blue, and long-lived modes (LLM) in green.

The task is formulated as a segmentation problem on spectrograms derived from Mirnov coil measurements. Each spectrogram covers a frequency range of 0–50 kHz. Preprocessing includes logarithmic scaling of the spectral power, and additional ridge enhancement techniques are being developed. Labels are created using a custom graphical tool. The tool applies thresholding and contour detection to produce segmentation masks that approximate the visible MHD modes. While the labels are hand-drawn by a non-expert and can be uncertain in some cases, they provide a useful starting point for training.

The labelled modes include fishbones, neoclassical tearing modes (NTM), long-lived modes (LLM) and sawteeth. An illustrative spectrogram with labelled MHD activity is shown in Fig. 2. Our preliminary baseline focuses on detecting long-lived modes. The dataset consists of around 85 hand-labelled spectrograms, of which 51 contain LLM activity. The signals come from the outer midplane vertical array of Mirnov coils. The target output is a segmentation mask indicating the presence and location of an LLM in the spectrogram. As baselines, we tested both a U-Net model, applied initially to a binary semantic segmentation task (mode versus no mode), and a Mask R-CNN with a ResNet-101 backbone and feature pyramid network (FPN), implemented in Detectron2 [11] and adapted for this dataset. The U-Net performed well on the simpler mode/no-mode task, while the Mask R-CNN was used for the more refined LLM identification.

Multiple methods have been proposed for segmenting Mirnov's spectrograms. Ridge filters [12] can process unlabelled data; however, they cannot differentiate between distinct mode structures or noise without additional heuristics or labelling. The use of deep learning was first demonstrated for high-frequency activity by [13] with human-labelled data. In this work, we used later approach from [14], which showed that fine-tuning Detectron2 outperforms previous supervised methods even with few labelled shots.

2.3 Confinement mode classification

Confinement regimes in tokamaks are broadly divided into low confinement (L-mode) and high confinement (H-mode). Other regimes, such as I-mode or QH-mode, have been reported on different devices, but on MAST the dominant regimes observed are L-mode and H-mode [15]. The transition between these regimes strongly influences transport and stability, and reliable classification of confinement intervals is an important step for event tagging and downstream prediction tasks. In this study, the classification task is defined as identifying H-mode intervals within each discharge [16].

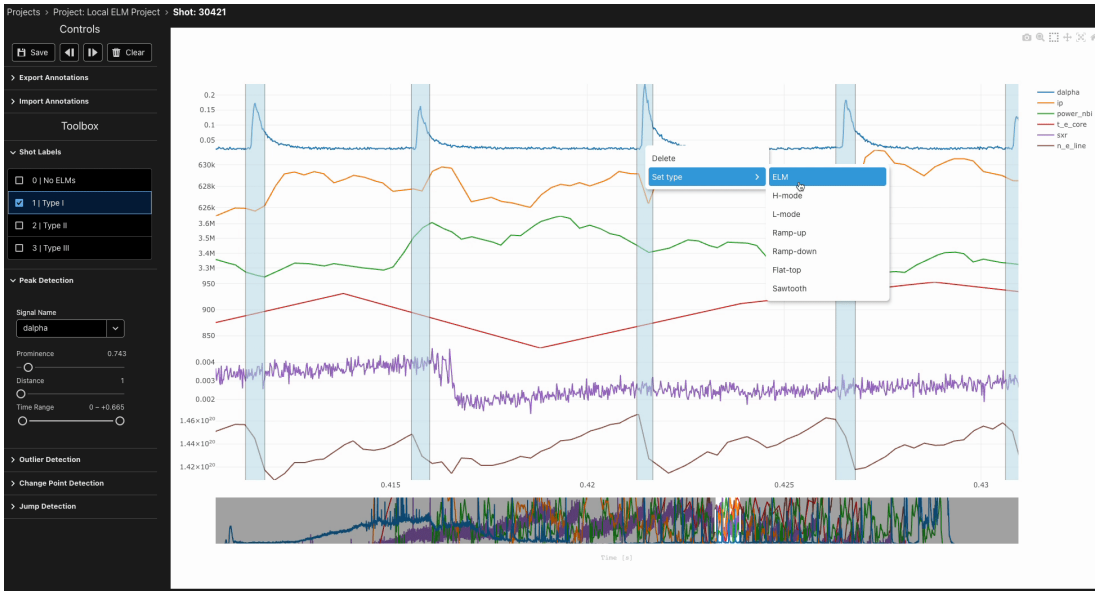

FIG. 3. Viz-annotation interface used for manual labelling of confinement mode and ELMs, illustrated here for MAST shot 30421.

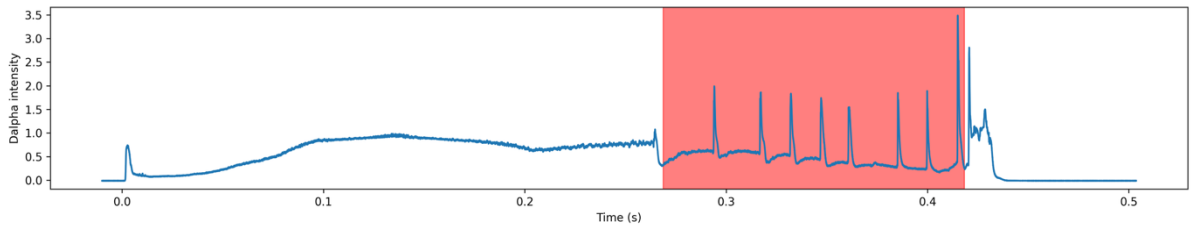

FIG. 4. Example D-alpha emission trace from MAST shot 12728, showing a labelled H-mode interval (red).

The dataset consists of 85 MAST shots with a total of 148 manually labelled H-mode intervals. Labels were created using a custom time series annotation tool (See Fig. 3), and considering D-alpha emission, interferometer density, and responsible operator comments detailing the H-mode time. Each interval is represented by start and end times, which are converted into binary masks spanning the discharge. An illustrative D-alpha trace with a labelled H-mode interval is shown in Fig. 4.

Multiple diagnostics are used as input features: D-alpha, plasma current, density gradient, line-averaged electron density, magnetic probe, and soft X-ray emission. The signals are resampled to 0.1 ms resolution, with density gradient additionally filtered using a median kernel. All channels are normalised by appropriate physical scaling factors. Input windows of 512 time steps are extracted with a step size of 512, producing a sequence-to-sequence labelling task.

A 1D U-Net model was used as the baseline architecture for this task. The model processes windowed time series from multiple diagnostics and produces a sequence of class predictions over the input window. Training was performed using a weighted sampler to mitigate the imbalance between L-mode and H-mode intervals.

2.4 ELM spike identification

Edge-localised modes (ELMs) are repetitive bursts of energy and particles that occur during H-mode. They appear as short, sharp spikes in edge diagnostics, especially D-alpha emission, and their frequency and amplitude are critical for assessing plasma–wall interaction [17]. Accurate detection of ELMs is therefore an essential complement to confinement mode classification.

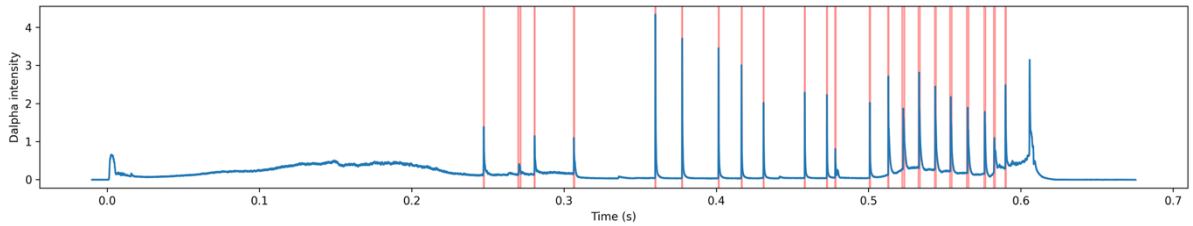

FIG. 5. Example D-alpha emission trace from MAST shot 18671, showing labelled edge-localised mode (ELM) spikes in red.

The dataset for this task consists of 101 MAST shots, where ELM spikes were labelled using a semi-automated approach: candidate spikes were first identified by a simple thresholding algorithm on the D-alpha signal, then verified and adjusted manually using a custom annotation tool (see Fig. 3). An example D-alpha trace with labelled ELMs is shown in Fig. 5. The annotations are converted into binary masks over the discharge timeline, with labels indicating the presence (1) or absence (0) of an ELM spike.

The same set of six diagnostics is used as input features: D-alpha, plasma current, density gradient, line-averaged electron density, magnetic probe, and soft X-ray emission. Signals are resampled to 0.1 ms resolution, median-filtered where appropriate, and normalised by physical scaling. Sliding windows of 512 time steps are generated with a step size of 512, producing windowed inputs aligned with the binary ELM masks.

For ELM detection, the same 1D U-Net baseline was applied. Models were trained with default hyperparameters, five-fold cross-validation, and class rebalancing to address the strong asymmetry between ELM and non-ELM intervals.

13. RESULTS AND DISCUSSIONS

Baseline performance is reported for the four representative tasks. Metrics used to measure performance are task dependent. The distributions of different metrics are summarised in Table 1 to indicate both central tendency and variability.

Table 1: Summary of baseline model performance across the four tasks on MAST data. Values are reported as mean ± standard deviation over the validation set.

Task	Confinement	ELMs	MHD modes	Disruption
Accuracy	0.883 ± 0.188	0.969 ± 0.040	0.994 ± 0.003	0.911 ± 0.083
Precision	0.822 ± 0.220	0.794 ± 0.203	0.749 ± 0.133	0.838 ± 0.070
Recall	0.834 ± 0.209	0.800 ± 0.201	0.727 ± 0.151	0.943 ± 0.063
F1-score	0.790 ± 0.247	0.781 ± 0.201	0.718 ± 0.099	0.868 ± 0.092
IoU	0.535 ± 0.397	0.364 ± 0.317	0.011 ± 0.005	0.785 ± 0.112
ROC AUC	0.90	0.85	0.81	0.515

3.1 Disruption prediction

A comprehensive parameter grid search was carried out using the weights and bias framework [18]. The search space included variation of lead time before disruption (10 ms, 30 ms, 60 ms), LSTM type (single-layer versus 2

layers stacked), bidirectional versus unidirectional recurrent units, dropout rate (0 or 0.3), and loss function (cross-entropy versus focal loss). Training was performed with a weighted random sampler to mitigate class imbalance. An illustrative case is shown in Fig. 6, where the model provides a 130 ms warning before the true disruption.

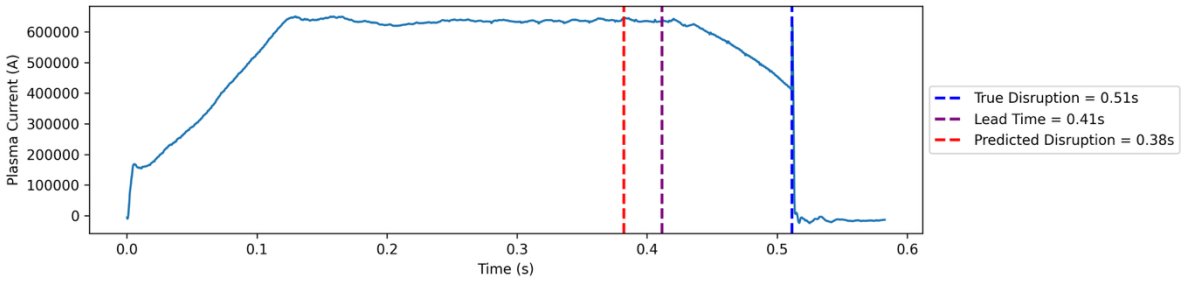

FIG. 6. Example of disruption prediction with 130 ms warning time for MAST shot 30060. Plasma current trace is shown with the predicted disruption time (red), true disruption time (blue) and the lead time (purple).

The best-performing configurations combined stacked, bidirectional LSTMs with dropout and cross-entropy loss, yielding higher recall and F1-scores compared with alternatives. Focal loss did not improve performance in this setting, and single-layer models underperformed relative to stacked ones. As expected, longer lead times degraded predictive accuracy, with 60 ms lead times showing the most pronounced drop. These findings indicate that disruption precursors can be reliably captured with recurrent architectures, but that performance is highly sensitive to both model configuration and the operationally chosen lead time.

3.2 MHD mode segmentation

The dataset for MHD activity is limited to 51 labelled shots, with long-lived modes (LLMs) making up the majority class. Other instabilities such as fishbones and neoclassical tearing modes (NTMs) occur too infrequently for meaningful baselines, so experiments were restricted to LLM segmentation.

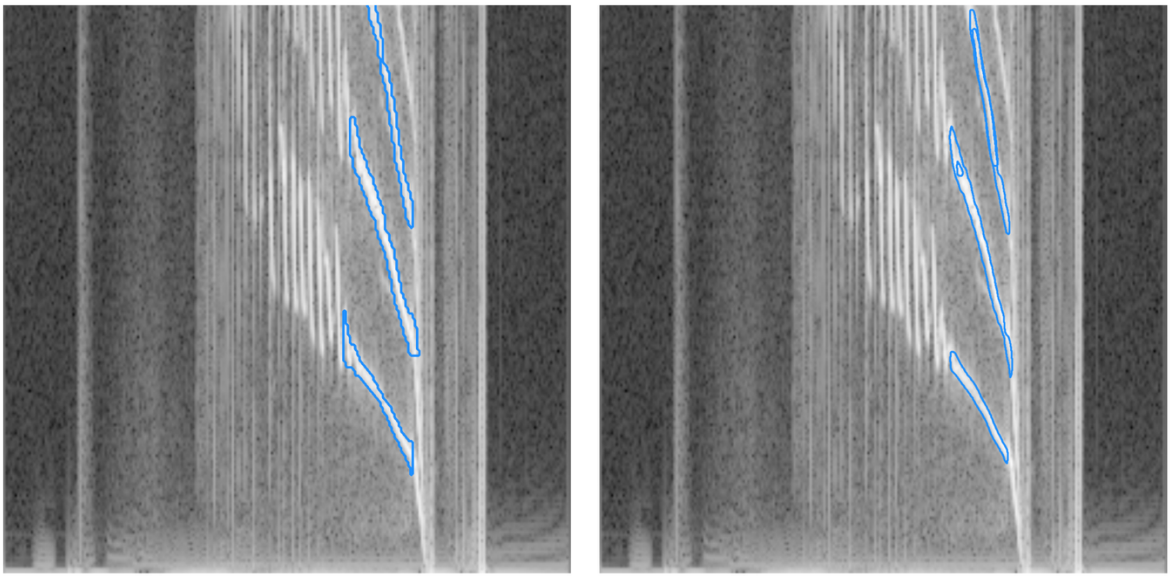

FIG. 7. Example of LLM segmentation in MAST shot 30374. Left: ground truth annotation. Right: prediction from Mask R-CNN baseline.

While accuracy and ROC AUC remain high, precision, recall, and F1-score exhibit greater variability. This reflects both the small dataset size and the extreme imbalance between background and mode pixels. Accuracy and ROC AUC are inflated by the dominance of background, while overlap-based metrics reveal that mode

boundaries are harder to capture. The IoU is especially low (0.01), emphasising the sensitivity of pixel-wise overlap to even small boundary errors.

Label quality also contributes to this variability. The annotations were created with a threshold-based graphical tool, which could not consistently capture high-frequency LLMs: lowering the threshold distorted low-frequency structures, while raising it suppressed high-frequency components. A qualitative example is shown in Fig. 7, comparing ground truth annotations and model predictions for shot 30374. The model captures the main LLM structures with reasonable fidelity, though boundaries remain approximate.

3.3 Confinement mode classification and ELM spike identification

Confinement mode classification and ELM spike identification are both based on manually labelled intervals from D-alpha emission. Confinement intervals span tens to hundreds of milliseconds, whereas ELMs appear as short bursts within those intervals. An example of model-predicted confinement intervals is shown in Fig. 8, while Fig. 9 illustrates predicted ELM spikes on a separate discharge.

The results in Table 1 show that both tasks achieve reasonably high accuracy (0.88 for confinement and 0.97 for ELMs) and ROC AUC (0.90 and 0.85, respectively), indicating that the baseline models capture the broad signal differences between labelled and unlabelled phases. However, F1-scores are lower (0.79 and 0.78), and IoU values in particular are depressed (0.54 and 0.36). This reflects the sensitivity of overlap metrics to even small temporal misalignments: for confinement, onset and termination of H-mode may be shifted by a few tens of milliseconds, while for ELMs the narrow spike widths make precise alignment especially difficult.

The number of labelled intervals is small, and performance metrics should be interpreted as indicative only. These tasks illustrate the strong dependence of performance on label quality and availability, and highlight the need for larger, systematically annotated datasets.

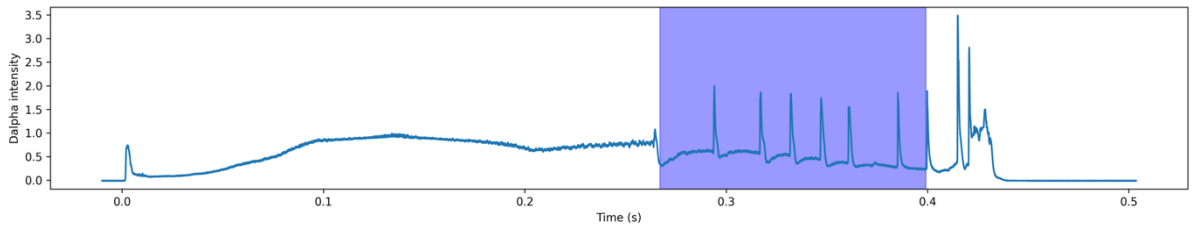

FIG. 8. Predicted confinement mode classification for MAST shot 12728. The baseline 1D U-Net model identifies H-mode intervals (shaded region) from multi-diagnostic inputs.

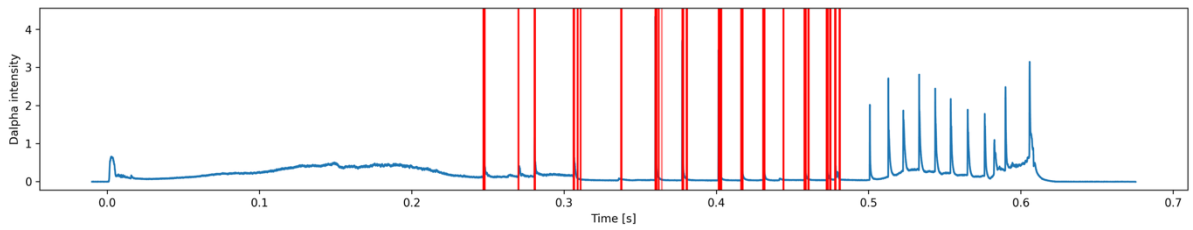

FIG. 9. Predicted ELM spike for MAST shot 18671. Short bursts in the D-alpha trace are detected (red markers).

4. CONCLUSION AND FUTURE WORK

This study presented a set of baseline models trained on FAIR-MAST data for four representative tasks: disruption prediction, MHD mode segmentation, confinement mode classification, and ELM spike identification. The baselines are intended to encourage use of FAIR-MAST data in machine learning studies and to support the creation of community benchmarks for fusion research.

A common theme across all tasks is the need for expert-curated labels. Current annotations are limited by automated detection or non-expert tools, and systematic review will be essential for reliable benchmarking. For disruption prediction, improved hand-labelling can resolve ambiguous cases where no sharp current change is visible. For MHD mode segmentation, extending annotations to NTMs, fishbones and other instabilities, and improving tools to capture multiple frequency ranges, will allow more comprehensive models. For confinement and ELMs, larger, systematically curated datasets are needed.

Beyond labels, semi-supervised and weakly supervised approaches offer a way to progressively refine datasets using the large volumes of unlabelled MAST data. Additional tasks such as locked-mode detection are natural extensions. Hosting curated datasets and models on public repositories (e.g. Hugging Face) would encourage wider use and help move towards community benchmarks for fusion, in the spirit of what large shared datasets achieved in other fields.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the MAST team for the collection of the experimental data. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3), operated by the University of Cambridge Research Computing Service. The MHD-modes task builds upon the Wavystar-labelling codes developed at MIT (Supported by ENI).

REFERENCES

[1] SNAPE, J., *Experimental studies of neoclassical tearing modes on the MAST spherical tokamak*, PhD Thesis, University of York (2012).
[2] JACKSON, S., KHAN, S., CUMMINGS, N., et al., *FAIR-MAST: A fusion device data management system*, SoftwareX (2024).
[3] JACKSON, S., KHAN, S., CUMMINGS, N., et al., *An open data service for supporting research in machine learning on tokamak data*, IEEE Trans. Plasma Sci. (2025).
[4] NAKANISHI, H., OHSUNA, M., KOJIMA, M., et al., *Data acquisition and management system of LHD*, Fusion Sci. Technol. 58 1 (2010) 445–457.
[5] ANIRUDH, R., ARCHIBALD, R., ASIF, M. S., et al., *2022 review of data-driven plasma science*, IEEE Trans. Plasma Sci. 51 (2023) 1–20.
[6] ZHU, J. X., REA, C., GRANETZ, R. S., et al., *Integrated deep learning framework for unstable event identification and disruption prediction of tokamak plasmas*, Nucl. Fusion (2023).
[7] WESSON, J. A., *Tokamak disruptions*, Plasma Phys. Control. Fusion 28 (1986) 243.
[8] SABBAGH, S. A., BERKERY, J. W., PARK, Y. S., et al., *Disruption event characterization and forecasting in tokamaks*, Phys. Plasmas 30 (2023) 032506.
[9] THORNTON, A. J., *The impact of transient mitigation schemes on the MAST edge plasma*, PhD Thesis, University of York (2011).
[10] CHAPMAN, I. T., *Resistive instabilities in tokamak plasmas*, Plasma Phys. Control. Fusion 53 (2011) 013001.
[11] WU, Y., Kirillov, A., Massa, F., Lo, W. Y., Girshick, R., *Detectron2: A PyTorch-based modular object detection library*, https://github.com/facebookresearch/detectron2 (2019).
[12] ZAPATA-CORNEJO E.D.D. et al., Plasma Phys. Control. Fusion 66, 095016 (2024).
[13] BUSTOS A. et al., Plasma Phys. Control. Fusion 63, 095001 (2021).
[14] ZAPATA-CORNEJO E.D.D., Ph.D. Thesis, Aix-Marseille University (2024).
[15] WAGNER, F., *A quarter-century of H-mode studies*, Plasma Phys. Control. Fusion 49 (2007) B1.
[16] CONNOR, J. W., WILSON, H. R., *A review of theories of the L–H transition*, Plasma Phys. Control. Fusion 42 (2000).
[17] ZOHM, H., *Edge localized modes (ELMs)*, Plasma Phys. Control. Fusion 38 (1996) 105.
[18] BIEWALD, L., Experiment Tracking with Weights and Biases (2020), www.wandb.com/