# Predicting H-to-L Mode Back-Transitions in Tokamak Plasma Dynamics

Virginia d'Adamo, Lucie Huamani-Cantrelle, Eva Quinto
*Department of Life Science Engineering, EPFL Lausanne, Switzerland*

***Abstract*—This study investigates the prediction of HL transitions in tokamak plasma confinement, which involves shifts between Low (L) mode and High (H) mode, influencing plasma stability and energy retention. The focus is on utilizing unidirectional and bidirectional Long Short-Term Memory (LSTM) models to forecast these transitions based on plasma parameters. The report outlines the process of data exploration and preprocessing, followed by the definition of the models, optimization of hyperparameters, and evaluation of their performance. The best-performing model, a Bidirectional LSTM (Bi-LSTM), achieved an F1 score of 0.7212, demonstrating its efficacy in predicting HL transitions.**

## I. Introduction

In tokamaks, plasma exists in two primary confinement modes, primarily Low (L) mode and High (H) mode. These modes are crucial for maintaining plasma stability and energy retention during fusion experiments. L mode is characterized by lower confinement, while H mode features enhanced confinement and more efficient heat retention, which are essential for fusion reactions. While the transition from L mode to H mode is a necessary and well-studied process, this study focuses on the reverse process: predicting back-transitions, where plasma shifts from H mode to L mode. These transitions are a significant challenge in fusion research and can lead to instability if not properly managed. By building predictive models leveraging machine learning techniques, the aim is to predict HL back-transitions based on a range of plasma parameters such as plasma current, electron density, and energy confinement and to provide valuable insights into the behavior of plasma in a tokamak, helping researchers improve the stability and efficiency of fusion experiments.

The remainder of this report is structured as follows:

- Section II describes the methodology (data exploration, preprocessing and the design of unidirectional and bidirectional LSTM models).
- Section III presents the results obtained from both models, including an evaluation of different configurations and preprocessing techniques.
- Section IV discusses the findings, highlighting the challenges encountered and their implications for predicting HL transitions.
- Section V concludes the report by summarizing the contributions and discussing potential future work.
- Section VI addresses ethical considerations related to the application of machine learning in plasma research.

## II. Models and Methods

### A. Data Exploration and Visualisation

The dataset initially comprised 297 distinct experiments, each with recorded transition times. An exploration of the dataset's features involved visualizing them during both HL transition and non-transition windows to identify potential anomalies. Key visualizations included electron density and temperature as functions of radius, distributions of individual features, and box plots comparing transition and non-transition windows. This analysis highlighted features that were more indicative of phase transitions, particularly emphasizing the importance of using absolute values for plasma current features, such as IP (total plasma current) and IPLA (measured plasma current), since the direction of the current (indicated by its sign) was not relevant for the analysis. Particular emphasis was placed on the feature *Halpha13*,

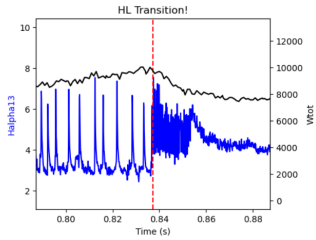



Figure 1: Time series of the *Halpha13* (blue) and *Wtot* (black) signals during a phase transition. The red vertical line indicates the time of the phase transition

which is associated with radiation emitted from the plasma (Figure 1). The signal is captured by a photodiode diagnostic system with multiple channels and filters. Channel 13, in particular, detected radiation from alpha particles in the plasma, providing valuable information about turbulence occurring at the plasma edge, known as edge-localized modes (ELMs). These ELMs are characteristic of H-mode and are crucial for identifying the confinement state of the plasma. As is evident from Figure 1 oscillations in the *Halpha13* signal were observed to be especially indicative of the phase transition. To analyze this behavior more effectively, a Fast Fourier Transform (FFT) was performed on the *Halpha13* signal.

The composition of the dataset was then analyzed. A total of 53 experiments were excluded during preprocessing due to missing key features, particularly the electron temperature and density profiles, previously identified as essential for our analysis. While necessary, these exclusions could be

addressed in future work using imputation techniques and additional computational resources to infer missing features from other experiments.

The distribution of missing values across experiments was analyzed, showing that 285 out of 297 experiments had low percentages of NaN values, indicating a favorable distribution for further analysis.

Principal Component Analysis (PCA) was performed, along with the visualizations of feature behavior and distribution during the transition and non-transition windows. All the considerations outlined above led us to select two distinct sets of features, which are hereafter referred to as features 0 and features 1, with the primary difference between them being the use respectively of IP or IPLA as the parameter for the plasma current. This distinction helped assess whether the choice of current representation influenced the capability to predict back-transitions.

### *B. Data Processing*

Preprocessing was performed using the two feature sets identified in Section II-A. Additionally, the possibility of replacing the original feature *Halpha13* with its frequency components derived from the FFT was explored. During preprocessing, missing values were handled by interpolating missing entries using a linear method. Rows that still contained NaN values after interpolation were removed. Special attention was given to the *time* column, where NaN values at the beginning of the dataset were backfilled and NaN values at the end were forward-filled to maintain continuity.

Once preprocessing was completed, the processed data was saved as parquet files for each experiment, containing all relevant features, and stored in a dedicated output folder. Data standardization was performed after preprocessing, specifically when creating time windows for the models, excluding the *time* column. Standardization was crucial to ensure all features were on a comparable scale, which is essential for subsequent analysis and model training. The data was organized into time windows as follows:

- The data was split into training, validation, and test sets, ensuring that all time windows from the same shot were allocated to the same set.
- Time windows of different sizes (0.1, 0.2 and 0.3 seconds) were created to capture patterns in the sequential data. Each window was labeled based on whether a phase transition occurred within that time frame.

### *C. Models*

The goal was to develop a model for accurately classifying time windows based on phase transitions using sequential data. Initially, a unidirectional LSTM (Uni-LSTM) model was implemented to simulate real-time predictions, relying solely on historical plasma measurements. Next, a Bidirectional LSTM (Bi-LSTM) model was introduced, based on the hypothesis that both past and future context could improve phase transition prediction. While both models share a similar structure, they differ in how they manage temporal dependencies.

The architecture includes:

- LSTM layers:
  - For the Uni-LSTM, the layers capture temporal dependencies from past time steps only.
  - For the Bi-LSTM, the layers capture dependencies from both past and future time steps.
- Dropout layers: to regularize the model and prevent overfitting by randomly deactivating a fraction of neurons during training.
- Batch normalization layers: to stabilize training by normalizing activations, helping the model converge faster and more reliably.
- Fully connected layer: The output layer consists of a single unit with a sigmoid activation function that predicts the binary class labels (phase transition: 1, no transition: 0). For the Bi-LSTM L2 regularization is applied to the weights of this layer to further prevent overfitting.

Both models are compiled with:

- Loss function: Binary crossentropy
- Optimizer: The Adam optimizer, with gradient clipping (clipvalue=1.0) to avoid exploding gradients.

To address class imbalance, class weights were computed based on the distribution of the training data, ensuring that the model did not bias the majority class. Early stopping is used to halt training based on validation performance.

### *D. Evaluation Metric*

Due to the imbalanced nature of the dataset, relying on accuracy alone was not appropriate because the model could achieve high accuracy by predominantly predicting the majority class (no phase transition). The F1 score offers a more meaningful assessment of model performance, particularly for tasks like anomaly detection and was therefore selected as the evaluation metric due. To better understand the model's learning progress, the F1 score was monitored and visualized over the training and validation epochs. The plots showed consistent improvement during training, with stable performance on the validation set, indicating that the model did not suffer from overfitting. Additionally, the confusion matrix for the predictions on the test set was displayed (see Appendix). This helped identify the classes where the model had difficulty making accurate predictions, providing insights into areas for potential improvement.

### *E. Hyperparameter Optimization*

Optuna was used for hyperparameter optimization to refine the model's performance. The key hyperparameters explored included the number of LSTM units in the layers, the dropout rate, the learning rate and the batch size. The

objective function was defined to maximize the F1 score on the validation set, ensuring that the model was optimized for a metric sensitive to class imbalance.

In addition to the hyperparameter analysis, different time window sizes were tested, starting from an initial window size of 100 ms. The influence of different window durations on the model's ability to capture relevant patterns and make accurate predictions was assessed by varying the time window. Larger windows were explored under the hypothesis that they could provide a broader context to better capture the transition process. The sampling frequency was also investigated by varying the number of time steps retained for each window. Through this exploration, the goal was to identify the optimal window size that strikes a balance between capturing sufficient temporal detail and maintaining computational efficiency, ultimately enhancing the model's ability to predict the HL transition. A different classification threshold was also evaluated. Specifically, a threshold higher than 0.5 was hypothesized to reduce the number of false positives (ensuring that predictions are made only when the model is confident), making it potentially more suitable for anomaly detection tasks.

All the explorations of time windows, thresholds and frequency were performed on the Bi-LSTM as from the first trials it already showed better performance and predictive capability.

## III. Results

All results are available in the GitHub repository; here, for brevity, only the key findings are summarized.

For the Uni-LSTM model, a 0.1-second time window and a classification threshold of 0.5 were used. All reported F1 scores are based on performance on the test set. Table I displays a variety of configurations, including tests both with and without preprocessing, as well as those utilizing different feature sets and FFT transformations applied to the *Halpha13* signal. Without preprocessing, the model achieved an F1 score of 0.4552, establishing a baseline. Introducing selected features and adjusting whether Halpha was used directly or in its FFT form generally improved the F1 score. For instance, using selected features without FFT increased the F1 to 0.5699.

| Model Configuration | Time Window (TW) | F1 Score |
|---|---|---|
| No preprocessing | 0.1 s | 0.4552 |
| With Preprocessing (Features 1) | 0.1 s | 0.5699 |
| With Preprocessing (Features 0 - FFT) | 0.1 s | 0.5145 |
| With Preprocessing (Features 0) | 0.1 s | 0.5631 |
| With Preprocessing (Features 1 - FFT) | 0.1 s | 0.4828 |

Table I: Unidirectional LSTM results for various configurations, including no preprocessing and different feature sets.

The results presented in Table II illustrate the performance of the Bi-LSTM model across various configurations, including different preprocessing methods, time window sizes, and thresholds. The model achieved its highest F1 score of 0.7193 using preprocessing with features 0, a time window of 0.2 s with 2000 steps, and a threshold of 0.5. Notably, preprocessing significantly improved the F1 score compared to using raw data, with features 0-FFT yielding an F1 score of 0.6351 for a 0.1 s time window with 1000 steps. F1 scores for combinations of features were lower than those for features 0 alone. Generally, adding FFT did not enhance the F1 score. The effect of time-frequency resolution with 2000 or 500 steps varied, showing different results depending on the preprocessing method used. Due to time constraints, higher thresholds and larger time windows have been investigated only for the best-performing preprocessing configuration.

In analyzing the confusion matrices, it was observed that with FFT and a time window of 0.2 s with 2000 steps, the model tended to favor negative predictions over positive ones. To address this imbalance, the configuration without FFT was tested with an increased threshold to reduce the number of false positives. While this adjustment significantly reduced false positives, the model struggled to correctly predict positive instances. A different time window was then explored, and with a 0.3 s window, a better balance between false positives and false negatives was observed, though the model still predicted a relatively high number of false positives.

| Preprocessing | TW | Threshold | F1 Score |
|---|---|---|---|
| No Preprocessing | 0.1 s, 1000 steps | 0.5 | 0.5923 |
| With Preprocessing (F0-FFT) | 0.1 s, 1000 steps | 0.5 | 0.6351 |
| With Preprocessing (F0-FFT) | 0.2 s, 500 steps | 0.5 | 0.6546 |
| With Preprocessing (F0-FFT) | 0.2 s, 2000 steps | 0.5 | 0.4881 |
| With Preprocessing (F0) | 0.2 s, 500 steps | 0.5 | 0.7162 |
| With Preprocessing (F0) | 0.2 s, 2000 steps | 0.5 | 0.7193 |
| With Preprocessing (F1) | 0.2 s, 500 steps | 0.5 | 0.5928 |
| With Preprocessing (F1-FFT) | 0.2 s, 500 steps | 0.5 | 0.6634 |
| With Preprocessing (F0) | 0.2 s, 2000 steps | 0.6 | 0.2674 |
| With Preprocessing (F0) | 0.3 s, 3000 steps | 0.5 | 0.7212 |

Table II: Results of different trials with varying preprocessing, time window (TW) size, threshold, and F1 scores. Note: F0 corresponds to Features 0, and F1 corresponds to Features 1.

## IV. Discussion

The Uni-LSTM model was evaluated to assess whether HL transitions could be predicted using only historical data. While this limitation impacts its predictive capacity, the Uni-LSTM achieved competitive F1 scores on the test set, especially when preprocessing steps were applied. For example, using selected features without FFT improved the F1 score to 0.5699, compared to the baseline of 0.4552 with no preprocessing. These results also demonstrate the importance of preprocessing and feature selection, even in case where bidirectional context is not available. Also, as shown in Table I, applying FFT transformations to the *Halpha13* signal often reduced the F1 score, suggesting that the frequency domain representation of the data was less informative in capturing the transition dynamics. In contrast, the use of selected features without FFT improved performance, achieving the highest F1 score of 0.5699 with

features 1. These findings align with the observations from the Bi-LSTM model, where preprocessing without FFT also yielded optimal results. However, unlike the Bi-LSTM, the Uni-LSTM was restricted to a fixed time window of 0.1 second. This smaller window likely contributed to its reduced ability to capture long-range dependencies.

On the other hand, Bi-LSTM model demonstrated promising results in predicting phase transitions, primarily due to its ability to capture both past and future dependencies in the time series data.

The results of our experiments reveal several key insights into the performance of the Bi-LSTM model under different configurations. First, the importance of preprocessing was confirmed: using features 0 alone, with a time window of 0.2 s and 2000 steps, led to the best performance, yielding the highest F1 score of 0.7193. This suggests that features 0, when combined with a sufficiently large time window, provide essential information for improving the model's predictive power.

Additionally, Bi-LSTM model's performance showed sensitivity to the time window size, with results varying significantly between 2000 and 500 steps. This underscores the importance of selecting an optimal window size that balances temporal resolution with the model's predictive performance.

An analysis of the confusion matrices revealed a notable challenge: the model tended to favor negative predictions. This issue was especially pronounced when using the 0.2 s time window with 2000 steps. Increasing the threshold helped reduce false positives but resulted in a significant drop in the number of correctly predicted positives. This trade-off highlights the critical need to carefully balance the threshold settings to ensure that the model does not sacrifice its ability to detect positive instances while trying to minimize false positives.

The increase in the time window from 0.1 s to 0.3 s was found to improve the model's ability to accurately capture phase transitions. A larger time window was shown to capture a broader range of temporal features, leading to a better understanding of the transition dynamics. Specifically, with a 0.3 s time window, a better balance between false positives and false negatives was observed, suggesting that a slightly larger time window may enhance overall model performance by providing more context for positive predictions. However, while this adjustment demonstrated a slight improvement in performance, the associated increase in computational cost was deemed impractical for routine analysis. Future work could focus on refining the data preprocessing steps and exploring the use of larger time windows with enhanced computational resources.

Furthermore, the approach used to handle class imbalance was found to be effective. Class weights were computed and incorporated into the loss function during training, ensuring that both transition and non-transition classes contributed equally to the model's learning process. This strategy was shown to mitigate the negative effects of class imbalance, enabling more effective predictions of both transitions and non-transitions.

## V. Conclusions

The Uni- and Bi-LSTM models were evaluated for their ability to predict back-transitions in tokamak plasma, each approach offering differents insights. The unidirectional LSTM model demonstrated that HL transitions can be predicted using only historical data, simulating real-time conditions. Despite this limitation, the unidirectional model achieved reasonable performance on the test set with F1 scores ranging from 0.4552 (without preprocessing) to 0.5699 (with selected features and no FFT).

The Bi-LSTM model, optimized for F1 score, provided a robust approach to predicting phase transitions in sequential data. The Bi-LSTM's ability to process information from both directions enables it to better understand the full context of the time series, thereby enhancing its performance, especially in scenarios where future behavior plays an equally important role as past behavior. By leveraging the power of bidirectional LSTMs, dropout regularization, and batch normalization, good performance was achieved even with class imbalance. Hyperparameter optimization using Optuna helped identify the best configuration for the model, and visualization of training metrics provided valuable insights into the learning process.

In general, preprocessing techniques proved to be more effective than simply selecting shots with a complete set of features.

Overall, both unidirectional and bidirectional LSTM models demonstrated their ability to predict HL transitions, with the bidirectional approach providing improved performance when future data was accessible. In conclusions, the combination of preprocessing techniques, careful threshold selection, and strategies for handling class imbalance allowed the Bi-LSTM model to achieve robust performance. Nonetheless, further exploration of model hyperparameters and additional feature engineering could improve the results further.

## VI. Digital Ethics

An Ethical Risk Assessment was performed using the Digital Ethics Canvas [1], a risk assessment grid with a series of questions that guide the analysis of software-specific ethical risks. Various risks were identified. One of particular importance for the project was in the context of autonomy: the risk of researchers not fully understanding how the ML model makes predictions or its limitations, which could lead to misinterpretations, over-reliance, and operational errors in tokamak experiments. This primarily impacts researchers and engineers, with broader implications for institutions relying on accurate plasma control. The negative impacts

include, first and foremost, operational risks. Misuse or over-reliance on the model could result in operational errors or unnecessary interventions, reducing experimental success. Moreover, misunderstanding the model's outputs could erode trust in the technology, slowing the adoption of ML-based tools in plasma research. The severity of this risk is high, as misunderstandings could lead to both costly experimental failures and setbacks in fusion research progress, eroding confidence in ML tools in the field. The likelihood is considered moderate, given the increasing knowledge of ML models among researchers worldwide.

To evaluate this risk, researchers' familiarity with ML was analyzed, and interpretability challenges in scientific ML were reviewed [2]. To address the issue, a user interface was developed with clear visualization tools to explain model predictions (e.g., showing which plasma parameters most influenced a transition), and interpretability metrics (like feature importance rankings) were integrated directly into the workflow. Nevertheless, constraints such as the inherent complexity of some methods, which may still be challenging for non-ML specialists, remain a challenge.

## ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Alessandro Pau for providing the opportunity to work in this lab and for his invaluable mentorship. Thanks are also extended to the PhD students:

- Svantner Jean-Pierre Thomas
- Cristina Venturini
- Yoeri Poels

for their critical support in understanding the data, preprocessing, and model development.

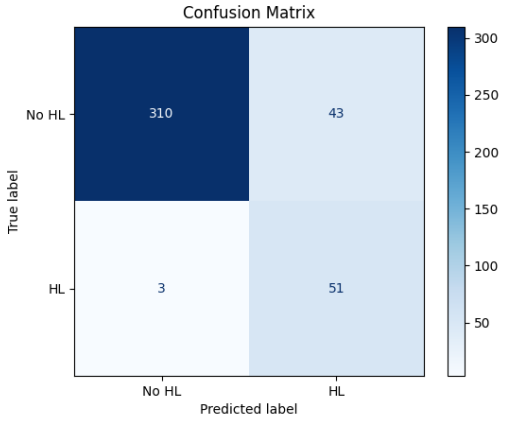
## REFERENCES

[1] E. C. for Digital Education (CEDE), "The digital ethics canvas: How to use it," n.d., accessed: 2024-12-05. [Online]. Available: https://www.epfl.ch/education/educational-initiatives/cede/training-and-support/digital-ethics/a-visual-tool-for-assessing-ethical-risks/the-digital-ethics-canvas-how-to/

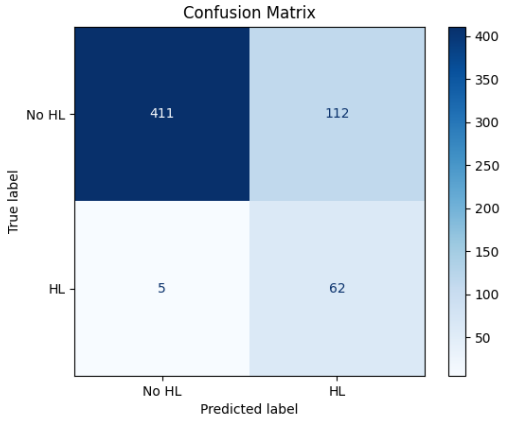
[2] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, 2019. [Online]. Available: https://www.mdpi.com/2079-9292/8/8/832

# APPENDIX

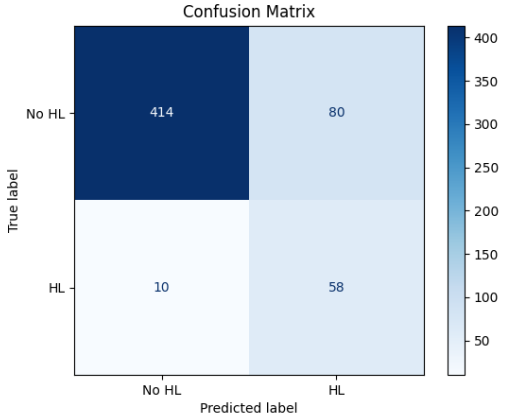
**Note:** The following confusion matrices show the classification results for different time window (TW) sizes. The number of TW tested for classification performance changes due to the variation in TW size.



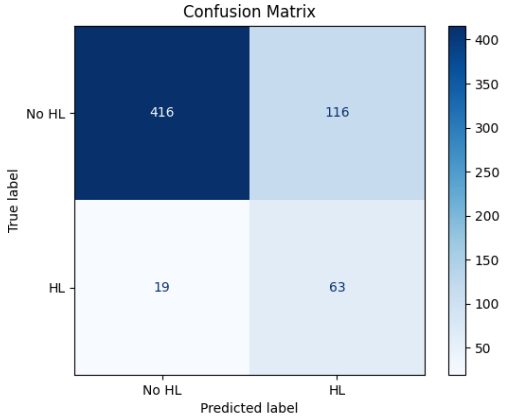

(a) Confusion Matrix for Features 0, TW=0.1s, threshold 0.5 (without FFT)



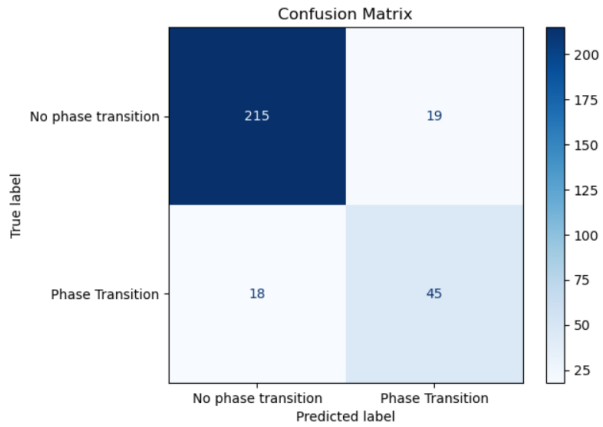

(b) Confusion Matrix for Features 0, TW=0.1s, threshold 0.5 (with FFT)



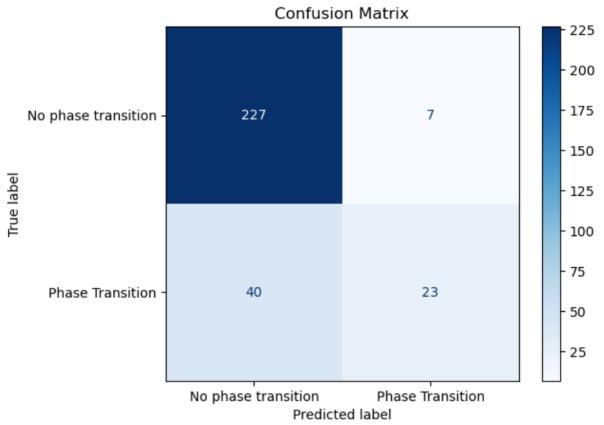

(c) Confusion Matrix for Features 1, TW=0.1s, threshold 0.5 (without FFT)



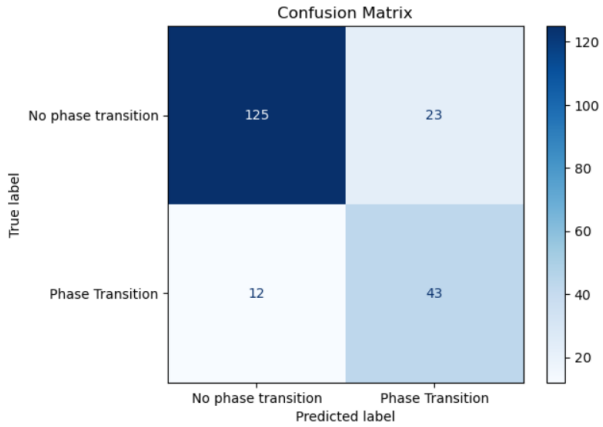

(d) Confusion Matrix for Features 1, TW=0.1s, threshold 0.5 (with FFT)



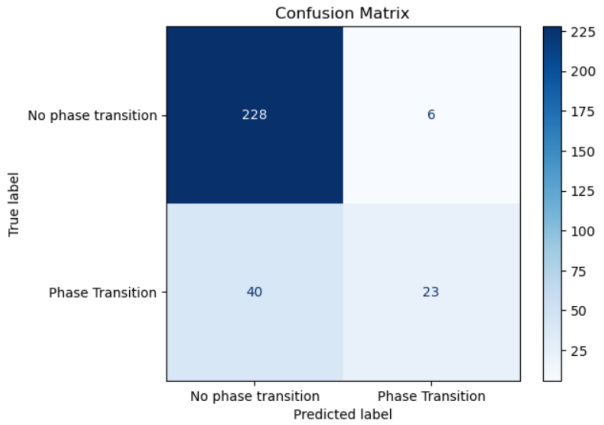

(a) Confusion Matrix for Features 0, TW=0.2s, threshold 0.5 (without FFT)




(b) Confusion Matrix for Features 0, TW=0.2s, threshold 0.5 (with FFT)




(c) Confusion Matrix for Features 0, TW=0.3s, threshold 0.5 (without FFT)




(d) Confusion Matrix for Features 0, TW=0.2s, threshold 0.6 (without FFT)

Figure 3: Confusion matrices for different configurations of Features 0, including variations in time windows (TW) and thresholds. The figures compare the impact of using FFT preprocessing (or not) and different threshold settings on model performance using Bi-LSTM.