APPENDIX: DEEPFIB: SELF-IMPUTATION FOR TIME SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

In this appendix, we first introduce the experimental settings in Section A. Then, the hyperparameter analysis for the proposed solution is presented in Section B.

A EXPERIMENTAL SETUP

We implement our method and its variants in PyTorch version 1.2.0 with CUDA 10.1, and train them on the platform with Intel(R) Core CPU i7-10700KF @ 3.80GHz and one NVIDIA RTX 2070s graphics cards. Other details are shown as follows:

A.1 DATASETS

We use 4 commonly-used real-world benchmark TS datasets, which cover different fields [1], ranging from human abnormal behavior detection, healthcare and fraud detection in finance.

2d-gesture: This dataset contains time series of X-Y coordinates of an actor's right hand. The data is extracted from an video in which the actor grabs a gun from his hip-mounted holster, moves it to the target, and returns it to the holster. The anomalous region is in the area where the actor fails to return his gun to the holster.

Power demand: This dataset contains one year of power consumption records measured by a Dutch research facility in 1997. *We believe this dataset contains some mislabeled ground truth, as shown in Fig. 1.*

ECG (Keogh et al., 2005): This dataset is often used for detecting anomalous beats from electrocardiograms readings, which comprise six 2-dimensional time series from six patients, where each time series has 3,750 to 5,400 observations.

Credit Card (Lai et al., 2021): This dataset is collected by openML, which contains transactions made by credit cards in September 2013 by European cardholders. The fraudulent transactions are labelled as outliers.

Since the training set and test set have been pre-defined, we use 10% training set for validation to allow model selection and hyperparameter tuning.

A.2 DATA PRE-PROCESSING

To mitigate the influences of data scale in different variates, we perform a data standardization on both training and testing set:

$$\widetilde{x} = \frac{x - min(X_{train})}{max(X_{train}) - min(X_{train})} \tag{1}$$

where $max(X_{train})$ and $min(X_{train})$ are the maximum value and the minimum value of the training set respectively.

[1] ECG, 2d-gesture and Power demand are from `http://www.cs.ucr.edu/eamonn/discords/`, while Credit Card is from `https://www.openml.org/d/1597`.

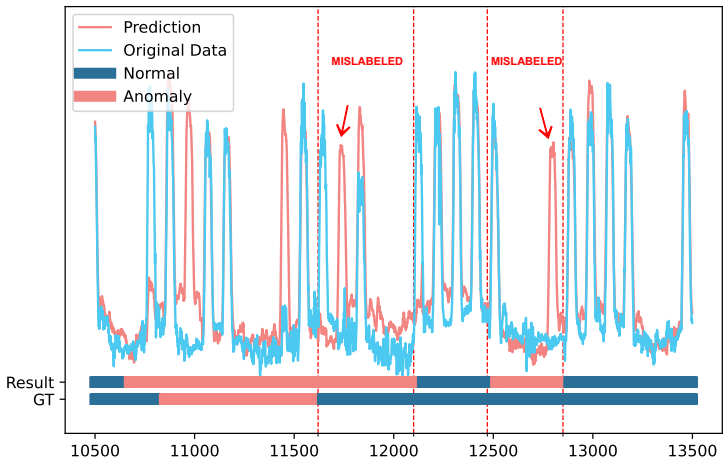


Figure 1: Two red peaks drawn in the dotted box are highly likely to be anomalies. However, they are labelled as normal data.

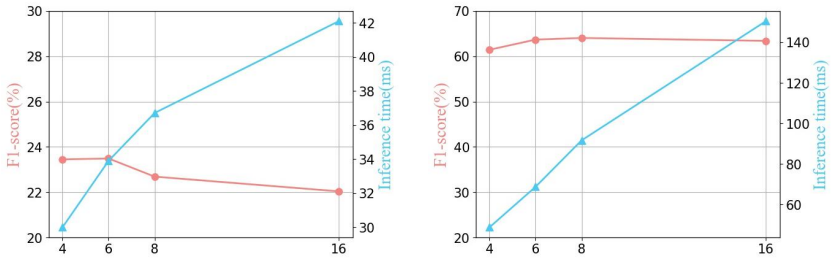
A.3 Implementation details

The models are trained using the Adam optimizer with a learning rate 1×10^{-3}. We train each model for 50 epochs at most with a batch size of 32, and the results are averaged over the five runs with different initializations. All time series from 4 datasets are partitioned into the fixed-length window (256 for ECG(E), 1472 for Power demand and 168 for others). The M in *DeepFIB-p* is set to 4 in *Credit Card*. The N in *DeepFIB-s* for ECG is set to 6 and 4 for other datasets.

In ablation study, for LSTM, the hidden state is chosen from $\{32, 64, 128, 256\}$. For TCN, in order to cover different length of the look-back window, the number of layers is chosen from $\{3, 4\}$ and the dimension of the hidden state is chosen from $\{32, 64, 128, 256\}$.

B Hyperparameter analysis

In the proposed DeepFIB framework, the number of maskings (i.e., M in *DeepFIB-p* and N in *DeepFIB-s*) is a hyperparameter. On the one hand, it indicates the number of non-overlapped samples generated in each timing window. On the other hand, a larger M or N value leads to less masking ratio in each training sample. Consequently, it could have a significant impact on AD accuracy and inference time.



(a) Credit Card (b) ECG (A)

Figure 2: The impact of masking times in DeepFIB-p (M) and DeepFIB-s (N) on *Credit Card* and *ECG(A)* datasets. The F1-scores with different M or N (X-axis) are shown in red curve and the corresponding inference time is shown in blue curve.

We conduct experiments on *Credit Card* for *DeepFIB-p* and *ECG(A)* for *DeepFIB-s*, respectively. The number of M and N are chosen from $\{4, 6, 8, 16\}$. For *Credit Card*, the outliers are densely

distributed in the time series. Therefore, in *DeepFIB-p*, a smaller M with a high masking ratio in each training sample facilitate the model learning more underlying temporal correlations to model such anomaly shapes, as shown in Fig. 2(a). For *ECG(A)*, anomalies lie in a local range, which shows less discriminative features from the normal patterns. Thus, as shown in Fig. 2(b), a larger N in *DeepFIB-s* can generate a series of training samples with shorter masking sub-sequences, making the model concentrate more on the local temporal relations to identify the delicate difference between the anomalies and normal data. On the other hand, as shown in Fig. 2, a larger value of N or M leads to a longer inference time and we need to take this issue into consideration during model development.

REFERENCES

E. Keogh, J. Lin, and A. Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *ICDM*, pp. 226–233, 2005.

Kwei-Herng Lai, D. Zha, Junjie Xu, and Yue Zhao. Revisiting time series outlier detection: Definitions and benchmarks. 2021.