DEEPFIB: SELF-IMPUTATION FOR TIME SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series (TS) anomaly detection (AD) plays an essential role in various applications, e.g., fraud detection in finance and healthcare monitoring. Due to the inherently unpredictable and highly varied nature of anomalies and the lack of anomaly labels in historical data, the AD problem is typically formulated as an unsupervised learning problem. The performance of existing solutions is often not satisfactory, especially in data-scarce scenarios. To tackle this problem, we propose a novel self-supervised learning technique for AD in time series, namely *DeepFIB*. We model the problem as a *Fill In the Blank* game by masking some elements in the TS and imputing them with the rest. Considering the two common anomaly shapes (point- or sequence-outliers) in TS data, we implement two masking strategies with many self-generated training samples. The corresponding self-imputation networks can extract more robust temporal relations than existing AD solutions and effectively facilitate identifying the two types of anomalies. For continuous outliers, we also propose an anomaly localization algorithm that dramatically reduces AD errors. Experiments on various real-world TS datasets demonstrate that DeepFIB outperforms state-of-the-art methods by a large margin, achieving up to 65.2% relative improvement in F1-score.

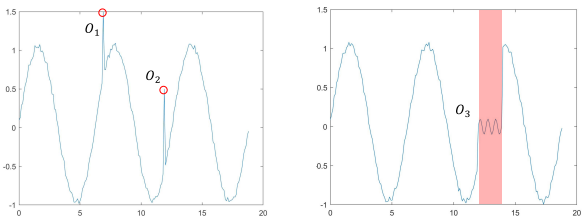
1 INTRODUCTION

Anomaly detection (AD) in time series (TS) data has numerous applications across various domains. Examples include fault and damage detection in industry (Hundman et al., 2018), intrusion detection in cybersecurity (Feng & Tian, 2021), and fraud detection in finance (Zheng et al., 2018) or healthcare (Zhou et al., 2019), to name a few.

Generally speaking, an anomaly/outlier is an observation that deviates considerably from some concept of normality (Ruff et al., 2021). The somewhat "vague" definition itself tells the challenges of the AD problem arising from the rare and unpredictable nature of anomalies. With the lack of anomaly labels in historical data, most AD approaches try to learn the expected values of time-series data in an unsupervised manner (Bl'azquez-Garc'ia et al., 2021). Various techniques use different means (e.g., distance-based methods (Angiulli & Pizzuti, 2002), predictive methods (Holt, 2004; Yu et al., 2016; Deng & Hooi, 2021) or reconstruction-based methods (Shyu et al., 2003; Malhotra et al., 2016; Zhang et al., 2019; Shen et al., 2021)) to obtain this expected value, and then compute how far it is from the actual observation to decide whether or not it is an anomaly.

While existing solutions have shown superior performance on some time series AD tasks, they are still far from satisfactory. For example, for the six ECG datasets in (Keogh et al., 2005), the average F1-score of state-of-the-art solutions (Kieu et al., 2019; Shen et al., 2021) with model ensembles are barely over 40%. Other than the TS data' complexity issues, one primary reason is that the available data is often scarce while deep learning algorithms are notoriously data-hungry.

Recently, self-supervised learning (SSL) that enlarges the training dataset without manual labels has attracted lots of attention, and it has achieved great success in representation learning in computer vision (Zhang et al., 2016; Pathak et al., 2016; Chen et al., 2020), natural language processing (Devlin et al., 2019), and graph learning (Hu et al., 2020) areas. There are also a few SSL techniques for time series analysis proposed in the literature. Most of them (Falck et al., 2020; Saeed et al., 2021; Fan et al., 2020) craft contrastive TS examples for classification tasks. (Deldari et al., 2021) also leverages contrastive learning for change point detection in time series.



(a) point-wise (b) sequence-wise

Figure 1: Anomalies in time series.

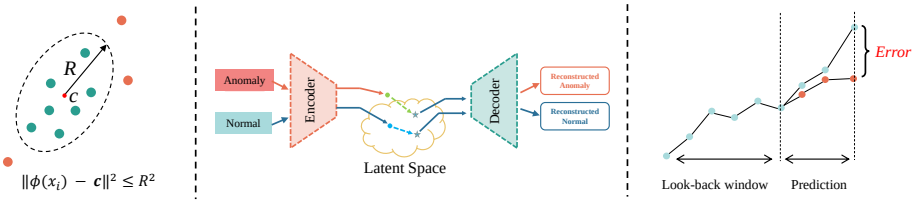
While interesting, the above SSL techniques do not apply to the AD task because detecting anomalies in time series requires fine-grained models at the element level. In this work, inspired by the context encoder for visual feature learning (Pathak et al., 2016) and the BERT model for language representation learning (Devlin et al., 2019), we propose a novel self-supervised learning technique for time series anomaly detection, namely *DeepFIB*. To be specific, we model the problem as a *Fill In the Blank* game by masking some elements in the TS and imputing them with other elements. This is achieved by revising the TS forecasting model *SCINet* (Liu et al., 2021) for the TS imputation task, in which the masked elements are regarded as missing values for imputation. Such self-imputation strategies facilitate generating a large amount of training samples for temporal relation extraction. As anomalies in time series manifest themselves as either discrete points or subsequences (see Fig. 1), correspondingly, we propose two kinds of masking strategies and use them to generate two pre-trained models. They are biased towards recovering from *point-wise* anomalies (*DeepFIB-p* model for *point outliers*) and *sequence-wise* anomalies (*DeepFIB-s* model for *continuous outliers*), respectively. To the best of our knowledge, this is the first SSL work for time series anomaly detection.

Generally speaking, AD solutions have difficulty detecting sequence-wise anomalies because it is hard to tell the real outliers from their neighboring normal elements due to their interplay. To tackle this problem, we propose a novel anomaly localization algorithm to locate the precise start and end positions of continuous outliers. As a post-processing step, we conduct a local search after determining the existence of sequence-wise anomalies within a timing window with our *DeepFIB-s* model. By doing so, the detection accuracy for continuous outliers is significantly improved.

We conduct experiments on several commonly-used time series benchmarks, and results show that DeepFIB consistently outperforms state-of-the-art solutions. In particular, the average F1-score of DeepFIB for the six ECG datasets is more than 62%, achieving nearly 50% relative improvement.

2 Related Work

In this section, we mainly discuss recent deep learning-based time series AD approaches. A comprehensive survey on the traditional techniques can be found in (Gupta et al., 2014).



(a) Density-based (b) Reconstruction-based (c) Prediction-based

Figure 2: Existing time series anomaly detection architectures.

Existing anomaly detection approaches can be broadly categorized into three types (see Fig. 2): (i) *Density-based* methods consider the normal instances compact in the latent space and identify anomalies with one-class classifiers or likelihood measurements (Su et al., 2019; Shen & Kwok, 2020; Feng & Tian, 2021). (ii) *Reconstruction-based* methods use recurrent auto-encoders (RAE) (Malhotra et al., 2016; Yoo et al., 2021; Kieu et al., 2019; Shen et al., 2021; Zhang et al., 2019) or deep generative models such as recurrent VAEs (Park et al., 2018) or GANs (Li et al., 2019; Zhou et al., 2019) for reconstruction. The reconstruction errors are used as anomaly scores. (iii) *Prediction-based* methods rely on predictive models (Bontemps et al., 2016; Deng & Hooi, 2021; Chen et al., 2021) and use the prediction errors as anomaly scores.

While the above methods have been successfully used in many real-world applications, practical AD tasks still have lots of room for improvement, especially in data-scarce scenarios. Unlike existing AD approaches, the proposed mask-and-impute method in *DeepFIB* exploits the unique property of TS data that missing values can be effectively imputed (Fang & Wang, 2020). By constructing many training samples via self-imputation, *DeepFIB* extracts robust temporal relations of TS data and improves AD accuracy dramatically. Moreover, for the more challenging sequence-wise anomalies, most prior work assumes a user-defined fixed-length for anomaly subsequences (Cook et al., 2020) or simplifies the problem by stating all the continuous outliers have been correctly detected as long as one of the points is detected (Su et al., 2019; Shen & Kwok, 2020). In *DeepFIB*, we lift these assumptions and try to locate the exact location of sequence-wise anomalies.

3 METHOD

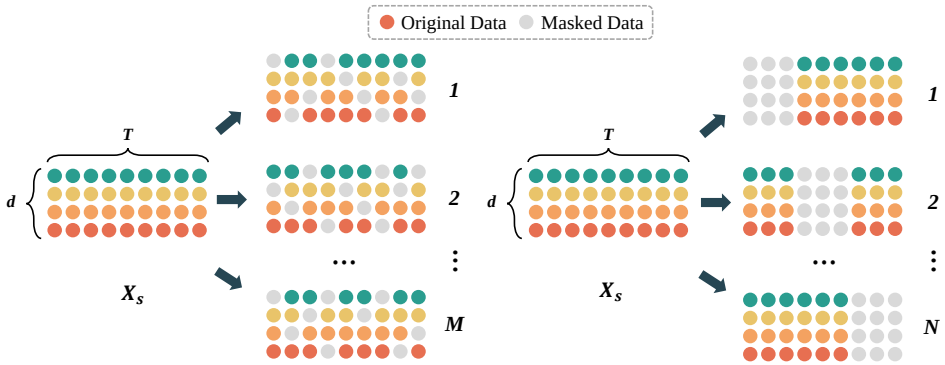
In this section, we first introduce the overall self-imputation framework in DeepFIB and then discuss the separate AD models for detecting point- and sequence-wise anomalies with different mask-and-impute strategies, namely *DeepFIB-p* and *DeepFIB-s*, respectively. Next, we describe the TS imputation method used in *DeepFIB*, based on an existing TS forecasting approach *SCINet* (Liu et al., 2021). Finally, we present our anomaly localization algorithm for continuous outliers.

3.1 SELF-IMPUTATION FOR ANOMALY DETECTION

Given a set of multivariate time series wherein $X_s = \{x_1, x_2, ..., x_{T_s}\} \epsilon \mathbb{R}^{d \times T_s}$ (T_s is the length of the s_{th} time series X_s), the objective of the AD task is to find all anomalous points $x_t \in \mathbb{R}^d$ (d is the number of variates) and anomalous subsequences $X_{t,\tau} = \{x_{t-\tau+1}, ..., x_t\}$.

The critical issue to solve the above problem is obtaining an expected value for each element in the TS, which requires a large amount of training data to learn from, especially for deep learning-based solutions. However, time-series data are often scarce, significantly restricting the effectiveness of learning-based AD solutions.

DeepFIB is a simple yet effective SSL technique to tackle the above problem. We model this problem as a *Fill In the Blank* game by randomly masking some elements in the TS and imputing them with the rest. Such self-imputation strategies generate many training samples from every time series and hence dramatically improve temporal learning capabilities.



(a) Point-wise masking (b) Sequence-wise masking

Figure 3: Self-imputation strategies of *DeepFIB-p* and *DeepFIB-s*.

In particular, we propose to train two self-imputation models (Fig. 3), biased towards point- and sequence-wise anomalies in the TS data, respectively.

- *DeepFIB-p model* targets point outliers, as shown in Fig. 3(a), in which we mask discrete elements and rely on the local temporal relations extracted from neighboring elements for reconstruction. For each time series X_s, we generate M training samples by masking it M times with randomly-selected yet *non-overlapping* $\frac{d \times T_s}{M}$ elements.

- *DeepFIB-s model* targets continuous outliers, as shown in Fig. 3(b), in which we mask continuous elements and rely on predictive models for reconstruction. For each time series X_s, we evenly divide it into N *non-overlapping* sub-sequences as $\left\{ X_{s,i}^{d\times\frac{T_s}{N}}, i \in [0, N-1] \right\}$ and generate N training samples by masking one of them each time.

During training, for each time series X_s, we obtain a set of non-overlapped imputed data with the above model and integrate them together results in a reconstructed time series $\widehat{X_s}$ (i.e., $\widehat{X_s}$-*p* for *DeepFIB-p* model and $\widehat{X_s}$-*s* for *DeepFIB-s* model). The training loss for both models are defined as the reconstruction errors between the input time series and the reconstructed one:

$$\mathcal{L} = \frac{1}{T_s} \sum_{t=1}^{T_s} \|x_t - \widehat{x_t}\| \tag{1}$$

where x_t is the original input value at time step t and the $\widehat{x_t}$ denotes the reconstructed value from the corresponding model, and $\|\cdot\|$ is the L1-norm of a vector.

During testing, to detect point outliers with the *DeepFIB-p* model, we simply use the residual error as the anomaly score, defined as $e_t = \sum_{i=0}^{d} \left| \widehat{x_t}^{\,i} - x_t^i \right|$, and when e_t is larger than a threshold value λ_p, time step t is regarded as an outlier. In contrast, for continuous outliers, we use dynamic time warping (DTW) (Sakoe & Chiba, 1978) distance metrics as our anomaly scoring mechanism, which measures the similarity between the input time series X and reconstructed sequence $\widehat{X}$. If $DTW(X, \widehat{X})$ is above a threshold value λ_s, a sequence-wise anomaly is detected.

3.2 Time Series Imputation in DeepFIB

While the time-series data imputation problem has been investigated for decades (Fang & Wang, 2020), there are still lots of rooms for improvement and various deep learning models are proposed recently (Cao et al., 2018; Liu et al., 2019; Luo et al., 2019).

SCINet (Liu et al., 2021) is an encoder-decoder architecture motivated by the unique characteristics of time series data. It incorporates a series of SCI-Blocks that conduct down-sampled convolutions and interactive learning to capture temporal features at various resolutions and effectively blend them in a hierarchical manner. Considering the highly-effective temporal relation extraction capability of SCINet when compared to other sequence models, we propose to revise it for the TS imputation task. More details about *SCINet* can be found in (Liu et al., 2021).

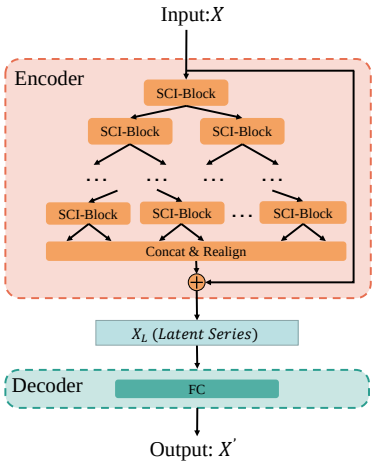


Figure 4: The structure of the SCINet.

To impute the missing elements from the two masking strategies with *DeepFIB-p* and *DeepFIB-s* models, we simply change the supervisions for the decoder part accordingly. For point imputation, we use the original input sequence as the supervision of our *DeepFIB-p* model, making it a reconstruction structure. By doing so, the model concentrates more on the local temporal relations inside the timing window for imputing discrete missing data, as shown in Fig. 5(a). As for continuous imputation, we propose to change SCINet as a bidirectional forecasting structure in our *DeepFIB-s* model, with the masked sub-sequence as supervision. As shown in Fig. 5(b), the two sub-models, namely *F-SCINet* and *B-SCINet*, are used to conduct forecasting in the forward and backward directions, respectively. By doing so, the model can aggregate the temporal features from both directions and learn a robust long-term temporal relations for imputing continuous missing data.

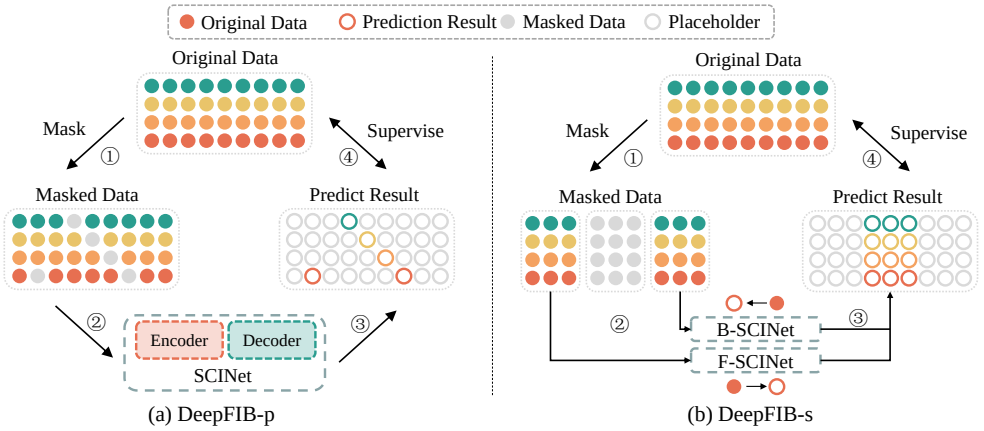


Figure 5: Time series imputation in DeepFIB.

3.3 Anomaly Localization Algorithm

During inference, we use a sliding window with stride μ to walk through the time series and find anomalies in each window. For sequence-wise anomalies, without knowing their positions a priori, we could mask some normal elements in the window and use those unmasked outliers for prediction (see Fig. 5(b)), thereby leading to mispredictions. To tackle this problem, we propose to conduct a local search for the precise locations of the sequence-wise anomalies.

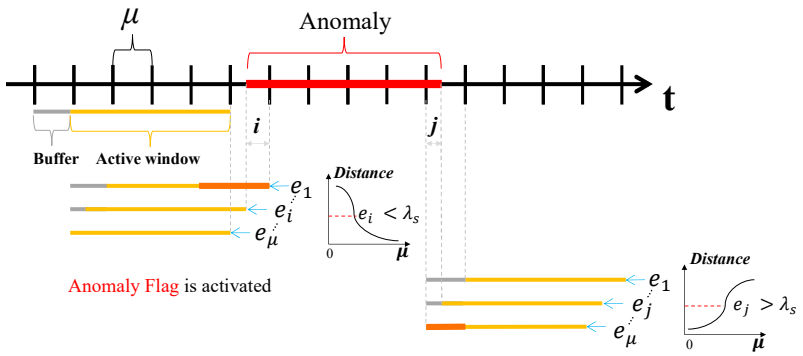


Figure 6: Anomaly localization algorithm.

As shown in Fig. 6, the *Active window* are the current input sequence to the DeepFIB-s model with length ω ($\omega > \mu$), i.e., $X_t = \{x_t, x_{t+1}, ..., x_{t+\omega-1}\}$ at time step t. When the *DTW* distance between the original time series in the *Active window* and the imputed sequence is above the threshold λ_s, a sequence-wise anomaly is detected in the current window, and the localization mechanism is triggered. As the sliding window is moving along the data stream with stride μ, if no outliers are detected in the previous window, the start position of the sequence-wise anomaly can only exist at the end of X_t in the window $\{x_{t+\omega-\mu}, ..., x_{t+\omega-1}, x_{t+\omega-1}\}$ with length μ. Consequently, by gradually shifting the *Active window* backward to include one more element in the *Buffer* window (see Fig. 6) at a time and calculating the corresponding *DTW* distances as $\{e_1, ..., e_i, ..., e_\mu\}$, we can find the maximum i with $e_i < \lambda_s$, indicating the following element after the *Active window* starting with i is the start of the anomaly subsequence. The *Anomaly flag* is then activated from this position. Similarly, to determine the ending position of the anomaly subsequence, we keep sliding the *Active windows* until we find a window with *DTW* distance smaller than λ_s, indicating that the ending position is within $\{x_{t-\mu}, ..., x_{t-2}, x_{t-1}\}$. Again, we shift the *Active window* backwardly one-by-one to include one element of the above window at a time and calculate the corresponding *DTW* distance, until we find the ending position with its *DTW* distance larger than λ_s.

4 EXPERIMENTS

In this section, we conduct extensive experiments to answer the following two questions: *Whether DeepFIB outperforms state-of-the-art AD methods* (**Q1**)? *How does each component of DeepFIB affect its performance* (**Q2**)?

Table 1: Datasets used in experiments

Datasets		#Dim	#Train	#Test	Anomaly
2d-gesture		2	8590	2420	24.63%
Power demand		1	18145	14786	11.44%
ECG	(A)chfdb_chf01_275	2	2888	1772	14.61 %
	(B)chfdb_chf13_45590	2	2439	1287	12.35 %
	(C)chfdbchf15	2	10863	3348	4.45 %
	(D)ltstdb_20221_43	2	2610	1121	11.51 %
	(E)ltstdb_20321_240	2	2011	1447	9.61 %
	(F)mitdb_100_180	2	2943	2255	8.38 %
Credit Card		29	142403	142404	0.173 %

Experiments are conducted on a number of commonly-used benchmark TS datasets, namely *2d-gesture*, *Power demand*, *ECG* and *Credit Card*, ranging from human abnormal behavior detection, power monitoring, healthcare and fraud detection in finance (see Table 1). As the anomalies in *2d-gesture*, *Power demand*, and *ECG* are mainly sequence outliers, we apply the *DeepFIB-s* model on these datasets. In contrast, the *Credit Card* dataset only contains point outliers, and hence we use *DeepFIB-p* model on it.

To make a fair comparison with existing models, we use the standard evaluation metrics on the corresponding datasets. For *2d-gesture*, *Power demand* and *Credit Card*, we use precision, recall, and F1-score following (Shen & Kwok, 2020). For *ECG* datasets, we use the AUROC (area under the ROC curve), AUPRC (area under the precision-recall curve) and F1-score, following (Shen et al., 2021). To detect anomalies, we use the maximum anomaly score in each sub-models over the validation dataset to set the threshold.

More details on experimental settings, additional experimental results and discussions (e.g., hyperparameter analysis) are presented in the supplementary materials.

4.1 **Q1**: COMPARISON WITH STATE-OF-THE-ART METHODS

Table 2: Comparison of anomaly detection performance (as %), on *2d-gesture* and *Power demand* datasets. The best results are in **bold** and the second best results are underlined.

Methods	2d-gesture			Power demand		
	precision	recall	F1-score	precision	recall	F1-score
DAGMM	25.66	80.47	38.91	34.37	41.72	37.69
EncDec-AD	24.88	**100.0**	39.85	13.98	54.20	22.22
LSTM-VAE	36.62	67.72	47.54	8.00	56.66	14.03
MADGAN	29.41	76.4	42.47	13.20	60.57	27.67
AnoGAN	57.85	46.50	51.55	20.28	44.41	28.85
BeatGAN	55.11	45.33	49.74	8.04	76.58	14.56
OmniAnomaly	27.70	*79.67*	41.11	8.55	*78.73*	15.42
MSCRED	*61.26*	59.11	60.17	*55.80*	34.32	42.50
THOC	54.78	75.00	*63.31*	**61.50**	36.34	*45.68*
DeepFIB	**93.90 ± 0.35**	60.77 ± 0.24	**73.79 ± 0.19**	52.21 ± 0.31	**99.99 ± 0.01**	**68.60 ± 0.15**

- The results of other baselines in the table are extracted from (Shen & Kwok, 2020)

2d-gesture and Power demand: The results in Table 2 show that the proposed *DeepFIB-s* achieves 16.55% and 50.18% F1-score improvements on *2d-gesture* and *Power demand*, respectively, compared with the second best methods.

For *2d-gesture*, the available training data is limited and the temporal relations contained in the data are complex (body jitter), making it difficult to obtain a discriminative representation in AD models. DAGMM (Zong et al., 2018) shows low performance since it does not consider the temporal

information of the time-series data at all. As for the AD solutions based on generative models (EncDecAD (Malhotra et al., 2016), LSTM-VAE (Park et al., 2018), MAD-GAN (Li et al., 2019), AnoGAN (Schlegl et al., 2017), BeatGAN (Zhou et al., 2019), OmniAnomaly (Su et al., 2019)), they usually require a large amount of training data, limiting their performance in data-scarce scenario. Compared to the above methods, the encoder-decoder architecture MSCRED (Zhang et al., 2019) is relatively easier to train and its AD performance is considerably higher. Moreover, the recent THOC (Shen & Kwok, 2020) work further improves AD performance by fusing the multi-scale temporal information to capture the complex temporal dynamics.

The proposed *DeepFIB-s* model outperforms all the above baseline methods since the proposed self-imputation technique allows the model to learn more robust temporal relations from much more self-generated training samples. Notably, we also observe that the *precision* of the *DeepFIB-s* dominates the other baselines. We attribute it to the anomaly localization algorithm that can locate the anomaly's precise start and end positions, significantly reducing the false positive rate.

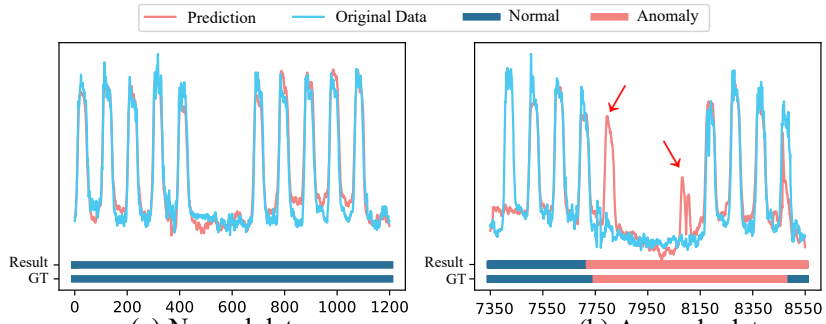


Figure 7: For Power demand dataset, (a) shows two cycles of normal data (0-1200 frame) wherein each cycle contains 5 peaks. (b) shows two cycles with anomaly with missing peaks highlighted using red arrows. The waveform of original data (light blue) is overlaid on the prediction result (light red). Lower color bars show the ground truth (GT) label and our detection result (Result).

For *Power demand*, the data contains many *contextual anomaly*[1] subsequences (see Fig. 7). It is quite challenging for existing AD approaches to learn such context information by extracting temporal features from the entire time series as a whole. In contrast, the proposed sequence-wise masking strategy facilitates learning different kinds of temporal patterns, which is much more effective in detecting such contextual anomalies. As shown in Table 2, the *recall* of our *DeepFIB-s* model almost reaches 100%, indicating all anomalies have been detected. The *precision* is not the best, and we argue that some of the false positives are in fact resulted from the poorly labeled test set (see our supplementary material).

ECG(A-F): Compared with (A),(B),(C) datasets, (D),(E),(F) are clearly noisy, which affect the performance of the anomaly detectors significantly. Nevertheless, Table 3 shows that *DeepFIB-s* achieves an average 46.3% F1-score improvement among all datasets and an impressive 65.2% improvement for ECG(F) dataset. There are mainly two reasons: (1) the data is scarce (See Table 1). Existing AD methods are unable to learn robust temporal relations under such circumstances. In contrast, the self-imputation training strategy together with the bidirectional forecasting mechanism used in our *DeepFIB-s* model can well address this issue; (2) the proposed DTW anomaly score is more effective in detecting the anomaly sequence than the previous point-wise residual scoring (see Section 4.2.1). Notably, the AUPRC of *DeepFIB* in *ECG(E)* is slightly lower than RAMED (Shen et al., 2021), and we attribute to the fact that some unlabeled sub-sequences are too similar to labeled anomalies in the raw data.

Credit Card: Due to the nature of this application, this dataset is stochastic and the temporal relation is not significant. Therefore, as shown in Table 4, traditional AD solutions without modeling the underlying temporal dependency achieve fair performance, e.g., OCSVM (Ma & Perkins, 2003), ISO

[1]Contextual anomalies are observations or sequences that deviate from the expected patterns within the time series however if taken in isolation they are within the range of values expected for that signal (Cook et al., 2020).

Table 3: Comparison of anomaly detection performance (as %), on ECG datasets.

Metrics	Methods	ECG						Average
		A	B	C	D	E	F	
AUROC	RAE	64.95	75.24	68.27	60.71	77.92	44.68	65.29
	RRN	69.50	72.07	68.49	47.05	78.81	47.87	63.97
	BeatGAN	66.51	73.14	58.69	59.33	82.98	44.19	64.14
	RAE-ensemble	68.26	77.63	70.55	64.64	83.14	39.66	67.31
	RAMED	*73.58*	*78.82*	*78.79*	*69.44*	*83.36*	*55.64*	*73.27*
	DeepFIB	**87.60 ±0.85**	**84.40 ±1.23**	**94.05 ±0.73**	**72.55 ±0.54**	**84.81 ±0.62**	**63.23 ±0.12**	**81.11**
AUPRC	RAE	51.84	40.32	31.23	15.54	24.17	7.76	28.48
	RRN	54.90	43.13	33.49	11.63	37.68	7.93	31.46
	BeatGAN	52.50	44.94	19.01	14.84	34.46	7.66	28.90
	RAE-ensemble	56.23	54.21	*49.90*	*18.47*	38.48	7.25	*37.42*
	RAMED	56.23	54.23	34.63	17.78	**45.78**	*10.59*	36.54
	DeepFIB	**85.18±0.63**	**75.48 ±0.56**	**73.47 ±0.67**	**23.14 ±0.45**	*38.27 ± 0.72*	**13.16 ± 0.23**	**51.45**
F1	RAE	52.51	49.03	32.79	25.43	33.63	15.47	34.81
	RRN	56.08	43.48	38.30	20.64	44.37	15.47	36.39
	BeatGAN	51.93	45.18	27.99	23.67	47.02	16.68	35.41
	RAE-ensemble	*56.42*	*52.40*	*58.68*	27.75	44.98	15.47	*42.62*
	RAMED	54.27	51.03	34.45	*30.87*	*52.23*	*20.63*	40.58
	DeepFIB	**80.90 ±0.63**	**78.06 ±0.82**	**78.37 ±0.13**	**44.71 ±0.19**	**58.00 ±0.26**	**34.08 ±0.73**	**62.35**

- The results of other baselines in the table are referred from (Shen et al., 2021)

Forest (Liu et al., 2008). Besides, the AR (Rousseeuw & Leroy, 1987) with a small window size (e.g., 3, 5) can also identify the local change point without considering longer temporal relations. However, the large *recall* and small *precision* values show its high false positive rates. The prediction-based method, LSTM-RNN (Bontemps et al., 2016) tries to learn a robust temporal relation from the data, which is infeasible for this dataset. In contrast, the reconstruction-based method, RAE (recurrent auto-encoder) (Malhotra et al., 2016) performs better since it can estimate the outliers based on the local contextual information. The proposed *DeepFIB-p* model outperforms all baseline methods, because it can better extract local correlations with the proposed self-imputation strategy. At the same time, compared to our results on other datasets, the relative 26.3% improvement over the second best solution (*AR*) is less impressive and the F1-score with our *DeepFIB-p* model is still less than 25%. We attribute it to both the dataset complexity and the lack of temporal relations in this dataset.

Table 4: Comparison of anomaly detection performance (as %), on Credit Card dataset.

Methods	Credit Card		
	precision	recall	F1-score
AR	11.30	**65.20**	*19.20*
ISO Forest	9.80	56.90	16.80
OCSVM	1.70	*62.00*	18.30
LSTM-RNN	0.40	11.00	0.70
RAE	**16.90**	21.52	18.89
DeepFIB-p	*16.52 ± 0.31*	46.57 ± 0.41	**24.25± 0.37**
RAE*	13.93 ± 0.12	53.36 ± 0.21	22.07± 0.36
DeepFIB-p†	16.55 ± 0.22	21.08 ± 0.12	18.50± 0.21

4.2 **Q2**: ABLATION STUDY

In this section, we first evaluate the impact of various components in our *DeepFIB-s* and *DeepFIB-p* models. Next, we replace the SCINet with other sequence models to evaluate its impact.

4.2.1 COMPONENT ANALYSIS

DeepFIB-p: To demonstrate the impact of the proposed mask-and-impute mechanism in point outlier detection. We add two baseline methods: (1) *DeepFIB-p*†, wherein we remove the self-imputation strategy; (2) *RAE**, we implement the same mask-and-impute strategy and apply it to the baseline method *RAE*. In Table 4, the performance improvement and degradation of the corresponding variants compared to *DeepFIB-p* and *RAE* clearly demonstrate the effectiveness of the proposed self-imputation strategy for point outlier detection.

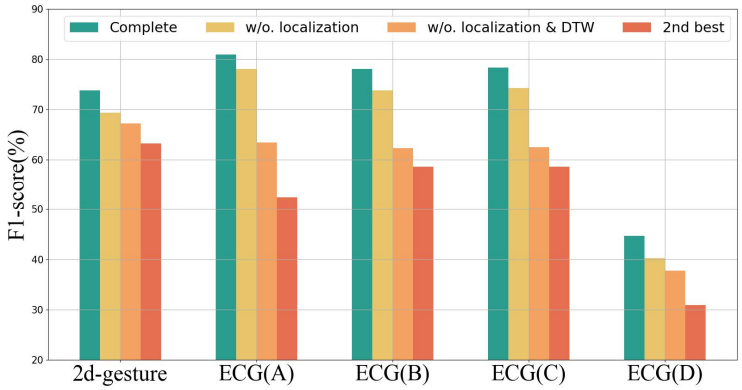


Figure 8: Ablation study of DeepFIB-s for five datasets (F1-score %). *2nd best* denotes the previous SOTA results for each dataset.

DeepFIB-s: To investigate the impact of different modules of *DeepFIB-s*, we compare two variants of the *DeepFIB-s* on five datasets. The details of the variants are described as below: For *w/o. localization*, we remove the anomaly localization algorithm from our *DeepFIB-s* model. The *w/o. localization & DTW* further removes the DTW scoring mechanism, and the anomalies are determined based on point-wise residual errors. As shown in Fig. 8, all these components are essential for achieving high anomaly detection accuracy. At the same time, the proposed self-imputation training strategy is still the main contributor to the performance of our *DeepFIB-s* model, as the results of *w/o. localization & DTW* are still much better than those of the 2nd best solution. Besides, the performance gain of the DTW anomaly scoring indicates that the point-wise outlier estimation is not suitable for evaluating sequence-wise anomalies.

4.2.2 Impact of SCINet

In our *DeepFIB* framework, we revise SCINet for time series imputation. To show its impact, we replace it with other sequence models in *DeepFIB-s*. As we can see in Table 5, compared with TCN (Bai et al., 2018) and LSTM (Hochreiter & Schmidhuber, 1997), using SCINet indeed brings significant improvements, which clearly shows its strong temporal relation extraction capability and the effectiveness of the revised architecture for TS imputation. At the same time, compared to the previous SOTA methods (*2nd best*) for the corresponding dataset, with the same mask-and-impute strategy, we can still achieve remarkable performance without using SCINet, indicating the effectiveness of the proposed self-imputation concept itself.

Table 5: The comparison of different sequence models. *2nd best* denotes the previous SOTA methods in each datasets (THOC in *2d-gesture* and RAE-ensemble in *ECG(A)*).

Methods	ECG(A)	2d-gesture
SCINet	**80.90 ± 0.63**	**73.79 ± 0.19**
TCN	69.86 ± 0.22	69.55 ± 0.28
LSTM	64.16 ± 0.21	66.83 ± 0.55
2nd best	56.42	63.31

5 Conclusion

In this paper, we propose a novel self-imputation framework *DeepFIB* for time series anomaly detection. Considering the two types of common anomalies in TS data, we implement two mask-and-impute models biased towards them, which facilitate extracting more robust temporal relations than existing AD solutions. Moreover, for sequence-wise anomalies, we propose a novel anomaly localization algorithm that dramatically improves AD detection accuracy. Experiments on various real-world TS datasets demonstrate that DeepFIB outperforms state-of-the-art AD approaches by a large margin, achieving up to more than 65% relative improvement in F1-score.

REFERENCES

F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *PKDD*, 2002.

S. Bai, J.Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *ArXiv*, abs/1803.01271, 2018.

A. Bl'azquez-Garc'ia, U. Conde, A.and Mori, and J.A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54:1 – 33, 2021.

L. Bontemps, V. L. Cao, J. McDermott, and N. A. Le-Khac. Collective anomaly detection based on long short-term memory recurrent neural networks. *International conference on future data and security engineering*, abs/1703.09752, 2016.

W. Cao, D. Wang, J. Li, H. Zhou, L. Li, and Y. Li. Brits: Bidirectional recurrent imputation for time series. *NeurIPS*, abs/1805.10572, 2018.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *PMLR*, abs/2002.05709, 2020.

Z. Chen, D. Chen, X. Zhang, Z. Yuan, and X. Cheng. Learning graph structures with transformer for multivariate time series anomaly detection in iot. *IEEE Internet of Things Journal*, abs/2104.03466, 2021.

A. A. Cook, G. Mısırlı, and Z. Fan. Anomaly detection for iot time-series data: A survey. *IEEE Internet of Things Journal*, 7:6481–6494, 2020.

S. Deldari, D. V. Smith, H. Xue, and F. D. Salim. Time series change point detection with self-supervised contrastive predictive coding. *Proceedings of the Web Conference 2021*, 2021.

A. Deng and B. Hooi. Graph neural network-based anomaly detection in multivariate time series. *ArXiv*, abs/2106.06947, 2021.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

F. Falck, S.K. Sarkar, S. Roy, and S.L. Hyland. Contrastive representation learning for electroencephalogram classification. 2020.

H. Fan, F. Zhang, and Y. Gao. Self-supervised time series representation learning by inter-intra relational reasoning. *arXiv*, abs/2011.13548, 2020.

C. Fang and C. Wang. Time Series Data Imputation: A Survey on Deep Learning Approaches. *arXiv*, 2020.

C. Feng and P. Tian. Time series anomaly detection for cyber-physical systems via neural system identification and bayesian filtering. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

M. Gupta, J. Gao, C.C. Aggarwal, and J. Han. Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 26:2250–2267, 2014.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997.

C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20:5–10, 2004.

W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V.S. Pande, and J. Leskovec. Strategies for pre-training graph neural networks. *ICLR*, 2020.

K. Hundman, V. Constantinou, Christopher Laporte, Ian Colwell, and T. Söderström. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

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

T. Kieu, B. Yang, C. Guo, and C.S. Jensen. Outlier detection for time series with recurrent autoencoder ensembles. In *IJCAI*, 2019.

D. Li, D. Chen, L. Shi, B. Jin, J. Goh, and S. Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *ICANN*, 2019.

F. Liu, K. Ting, and Z. Zhou. Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008.

M. Liu, Q. Zeng, A.and Lai, and Q Xu. Time series is a special sequence: Forecasting with sample convolution and interaction. *ArXiv*, abs/2106.09305, 2021.

Y. Liu, R. Yu, S. Zheng, E. Zhan, and Y. Yue. Naomi: Non-autoregressive multiresolution sequence imputation. In *NeurIPS*, 2019.

Y. Luo, Y. Zhang, X. Cai, and X. Yuan. E²gan: End-to-end generative adversarial network for multivariate time series imputation. In *IJCAI*, 2019.

J. Ma and S. Perkins. Time-series novelty detection using one-class support vector machines. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, 3:1741–1745 vol.3, 2003.

P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G.M. Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *ArXiv*, abs/1607.00148, 2016.

D. Park, Y. Hoshi, and C.C. Kemp. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3:1544–1551, 2018.

D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A.A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

P. Rousseeuw and A. Leroy. Robust regression and outlier detection. 1987.

L. Ruff, J. Kauffmann, R.A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T.G. Dietterich, and K. Muller. A Unifying Review of Deep and Shallow Anomaly Detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. ISSN 0018-9219.

A. Saeed, F.D. Salim, T. Ozcelebi, and J.J. Lukkien. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Internet of Things Journal*, 8:1030–1040, 2021.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:159–165, 1978.

T. Schlegl, P. Seeböck, Waldstein, S.M., U.M. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *IPMI*, 2017.

Z. Shen, L.and Li and J.T. Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. In *NeurIPS*, 2020.

Z. Shen, L.and Yu, Q. Ma, and J.T. Kwok. Time series anomaly detection with multiresolution ensemble decoding. In *AAAI*, 2021.

M. Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. 2003.

Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

Y. Yoo, U. Kim, and J. Kim. Recurrent reconstructive network for sequential anomaly detection. *IEEE Transactions on Cybernetics*, 51:1704–1715, 2021.

Q. Yu, L. Jibin, and L. Jiang. An improved arima-based traffic anomaly detection algorithm for wireless sensor networks. *International Journal of Distributed Sensor Networks*, 12, 2016.

C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, and N.a Chawla. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. *AAAI*, abs/1811.08055, 2019.

R. Zhang, P. Isola, and A.A. Efros. Colorful image colorization. In *ECCV*, 2016.

Y. Zheng, X. Zhou, W. Sheng, Y. Xue, and S. Chen. Generative adversarial network based telecom fraud detection at the receiving bank. *Neural networks : the official journal of the International Neural Network Society*, 102:78–86, 2018.

B. Zhou, S. Liu, B. Hooi, X. Cheng, and J. Ye. Beatgan: Anomalous rhythm detection using adversarially generated time series. In *IJCAI*, 2019.

B. Zong, Q. Song, M. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.