Обнаружение аномалий во временных рядах с помощью сигнатур

Отливанчик Павел

Аномалии во временных рядах

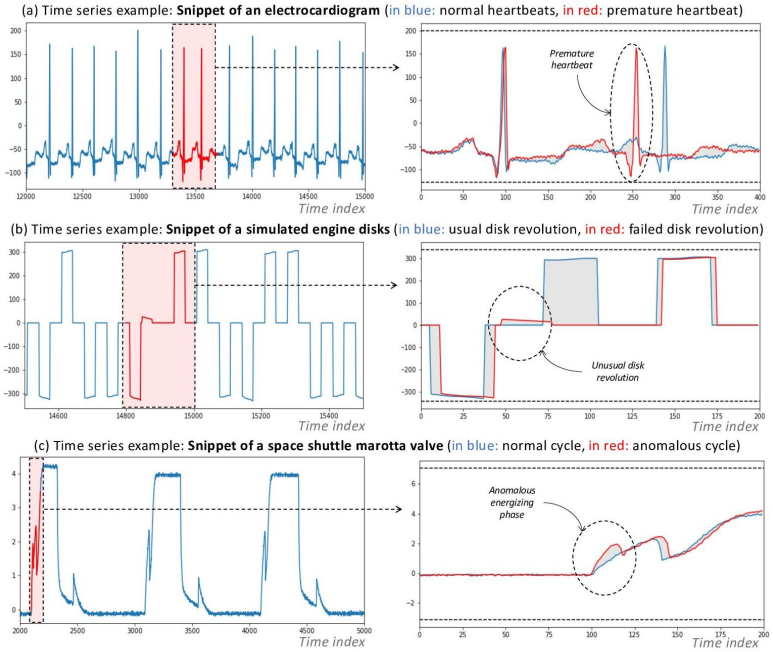


Fig. 1. Examples of different time series applications and types of anomalies.

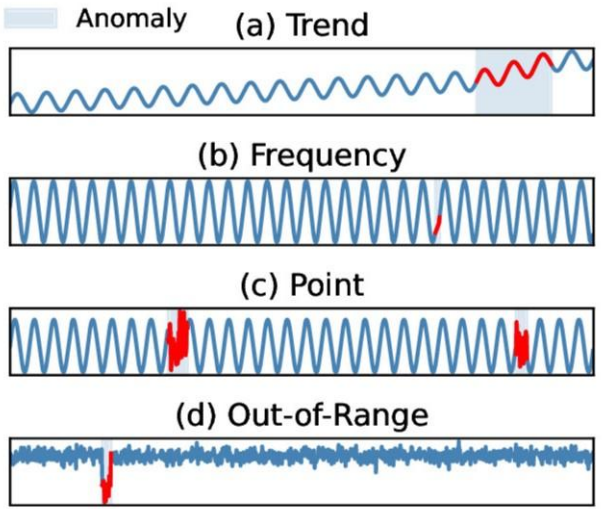


Figure 1: Example time series with different anomaly types, with anomalous regions highlighted in red.

Цели проекта

1. Изучить проблематику обнаружения аномалий.
2. Изучить актуальные подходы к решению этой задачи.
3. Изучить сигнатуры и их виды.
4. Разработать решения, основанные на сигнатурах.
5. Протестировать их и сравнить между собой и с бессигнатурными методами.

Что такое сигнатура?

Definition 1.4 (Signature). The *signature* of a path $X : [a, b] \to \mathbb{R}^d$, denoted by $S(X)_{a,b}$, is the collection (infinite sequence) of all the iterated integrals of X. Formally, $S(X)_{a,b}$ is the collection of real numbers

$$S(X)_{a,b} = (1, S(X)^1_{a,b}, \dots, S(X)^d_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, \dots)$$

where the "zeroth" term, by convention, is equal to 1, and the index in the superscript runs along the set of all *multi-indexes*

$$W = \{(i_1, \dots, i_k) \mid k \geq 0, i_1, \dots, i_k \in \{1, \dots, d\}\}. \tag{4}$$

$$S(X)^{i_1,\dots,i_k}_{a,t} = \int_{a<t_k<t} \dots \int_{a<t_1<t_2} dX^{i_1}_{t_1} \dots dX^{i_k}_{t_k}.$$

Другие виды сигнатур

Definition 5.9 (Localized Randomized Signature)**.** Let $A_1, \dots, A_d \in \mathbb{R}^{k\times k}$ and $b_1, \dots, b_d \in \mathbb{R}^k$ with entries sampled i.i.d. from a normal distribution. Let further denote σ a activation function $\sigma : \mathbb{R} \to \mathbb{R}$ that is applied componentwise. Then

$$\mathrm{d}Z_t = \sum_{i=i}^{d} \sigma(A_i Z_t + b_i)\mathrm{d}X^i(t), \quad Z_0 = z \in \mathbb{R}^k, \quad t \in [0,T] \tag{5.28}$$

is called the *Localized Randomized Signature* of X.

Definition 1.27 (Log-signature). For a path $X : [a,b] \to \mathbb{R}^d$, the log-signature of X is defined as the formal power series $\log S(X)_{a,b}$.

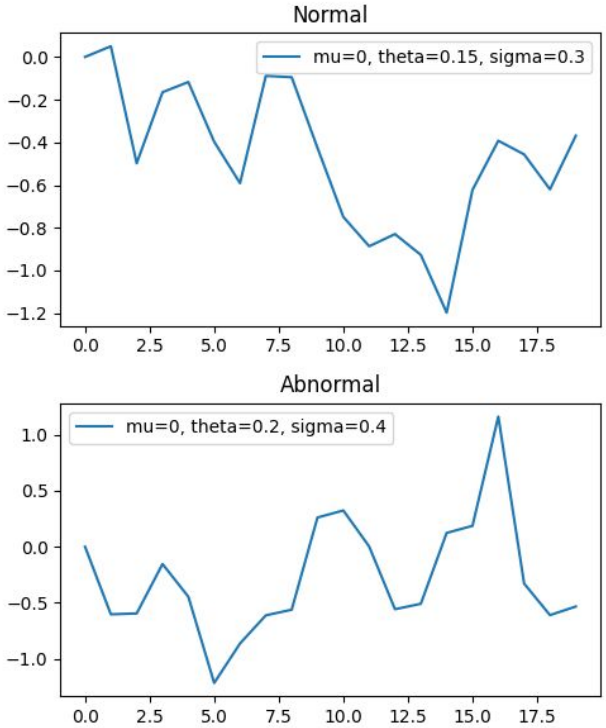
For two formal power series x and y, we define their Lie bracket by

$$[x,y] = x \otimes y - y \otimes x. \tag{24}$$

A direct computation shows that the first few terms of the log-signature are given by

$$\log S(X)_{a,b} = \sum_{i=1}^{d} S(X)^i_{a,b} e_i + \sum_{1\le i<j\le d} \frac{1}{2}\left(S(X)^{i,j}_{a,b} - S(X)^{j,i}_{a,b}\right)[e_i, e_j] + \dots. \tag{25}$$

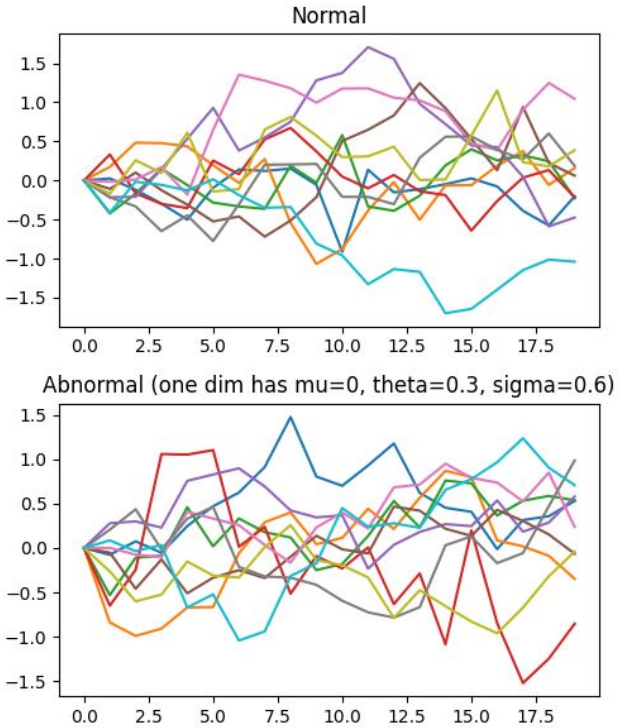
Основные датасеты



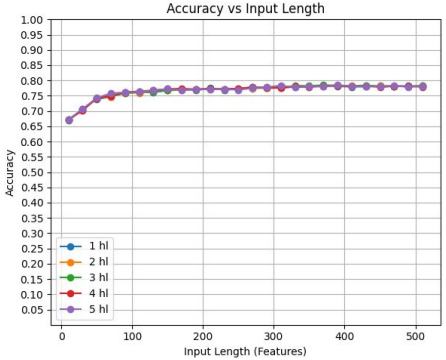
1. Одномерные процессы О-У.

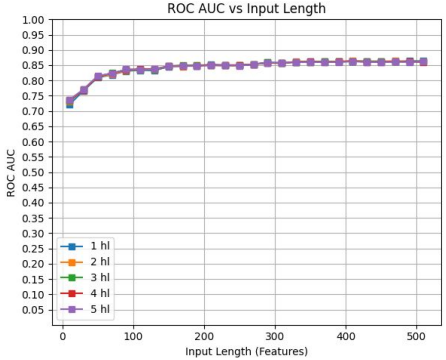
1. Десятимерные процессы О-У(1 случайное измерение аномально).

$$dx_t = \theta(\mu - x_t)\,dt + \sigma\,dW_t$$

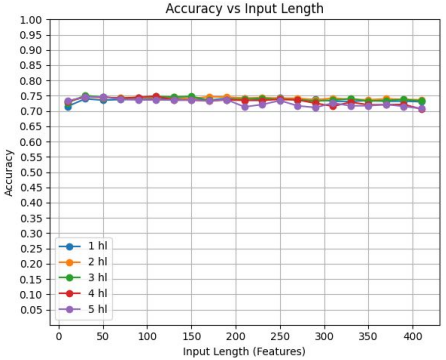


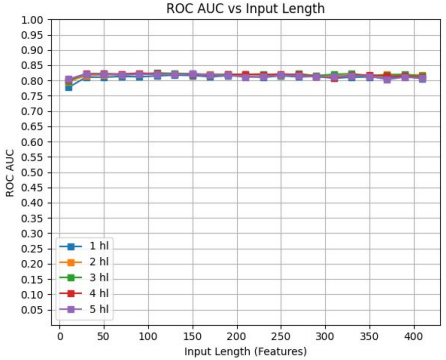
Сигнатура



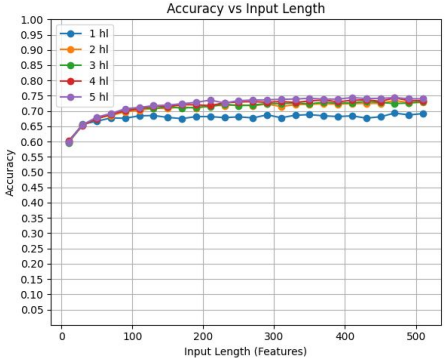


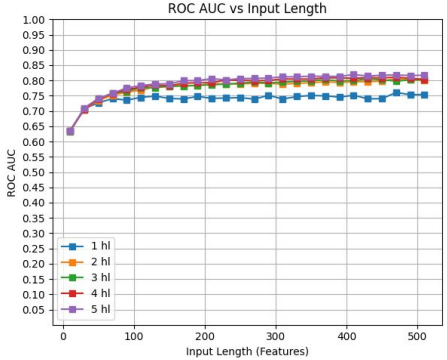
Лог-сигнатура





Рандомизированная сигнатура

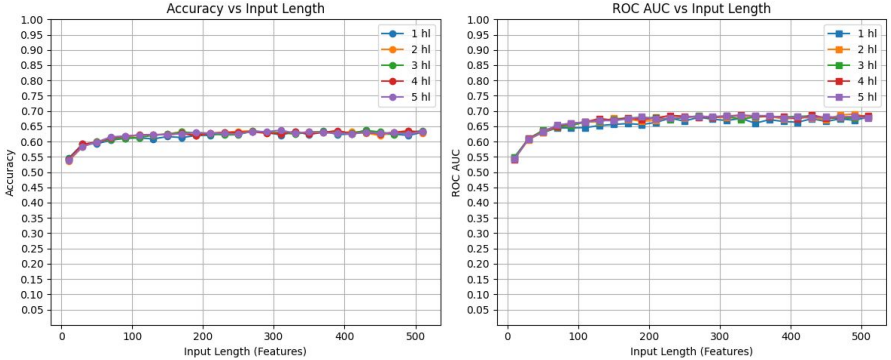




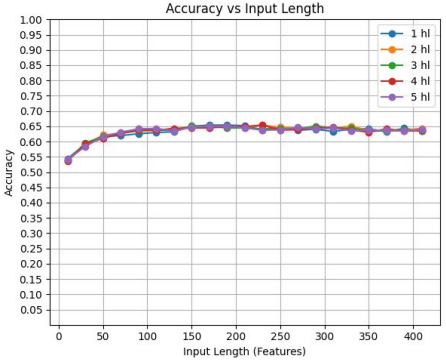
Без сигнатуры

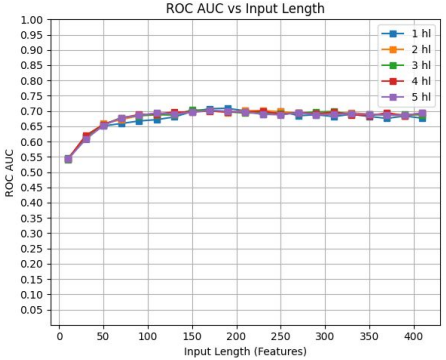
Модель	Accuracy	ROC-AUC
NN(1)	0.81	0.88
NN(2)	0.81	0.89
NN(3)	0.81	0.89
NN(4)	0.81	0.89
NN(5)	0.81	0.89

Сигнатура

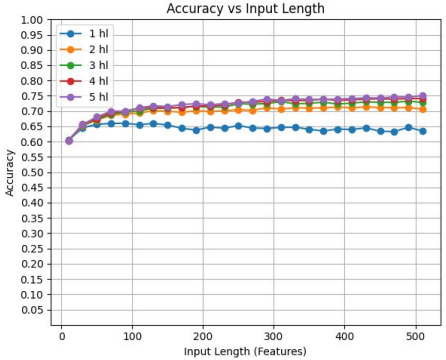


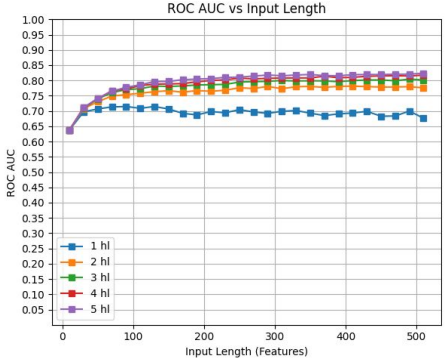
Лог-сигнатура





Рандомизированная сигнатура





Без сигнатуры

Модель	Accuracy	ROC-AUC
NN(1)	0.68	0.74
NN(2)	0.77	0.84
NN(3)	0.78	0.86
NN(4)	0.78	0.86
NN(5)	0.77	0.85

Подходы к обнаружению аномалий

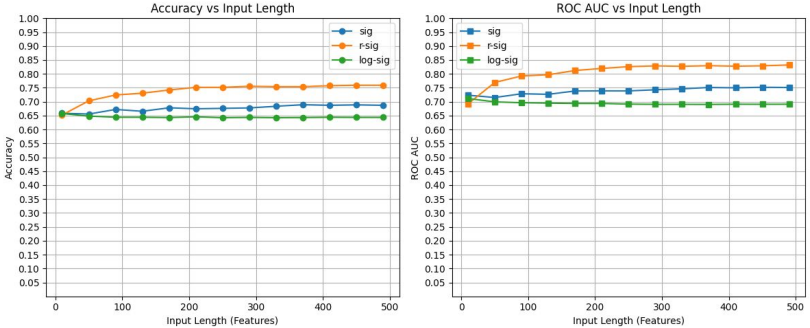
1. Брать сигнатуру по всему ряду и строить классификатор на них.
2. Использовать imputation модель(диффузионка CSDI) для генерации выборки, сравнивать сигнатуры выборки с истинной сигнатурой.
3. Брать сигнатуру на каждом шаге ряда и использовать RNN. $X_t = Sig_{[1,t]}$
4. Брать сигнатуру на каждом шаге ряда как препроцессинг для данных и затем использовать модели(TimesNet) на преобразованных данных.

1	2	1+2	3
Isolation forest	MAE	VAE+finetuning	GRU autoencoder
MAE	KDE	VAE+Mahalanobis dist	LSTM autoencoder
KDE	Mahalanobis dist		
	One Class SVM		
	HMM		
	GMM		

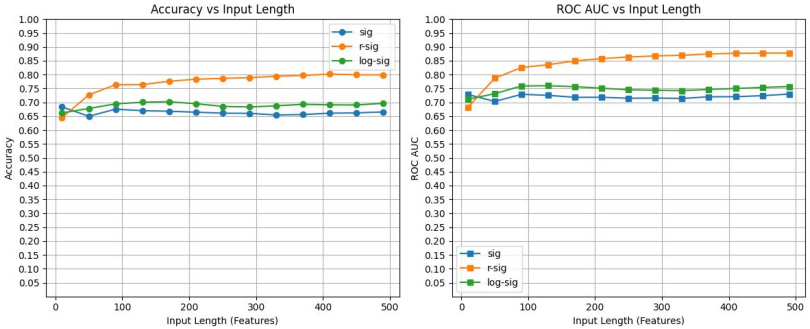
Лучшие результаты

LSTM autoencoder (Semi-supervised)

Датасет 1



Датасет 2



LSTM autoencoder без сигнатур

Accuracy	ROC-AUC
0.7	0.76

Accuracy	ROC-AUC
0.66	0.71

Без сигнатуры (Supervised)

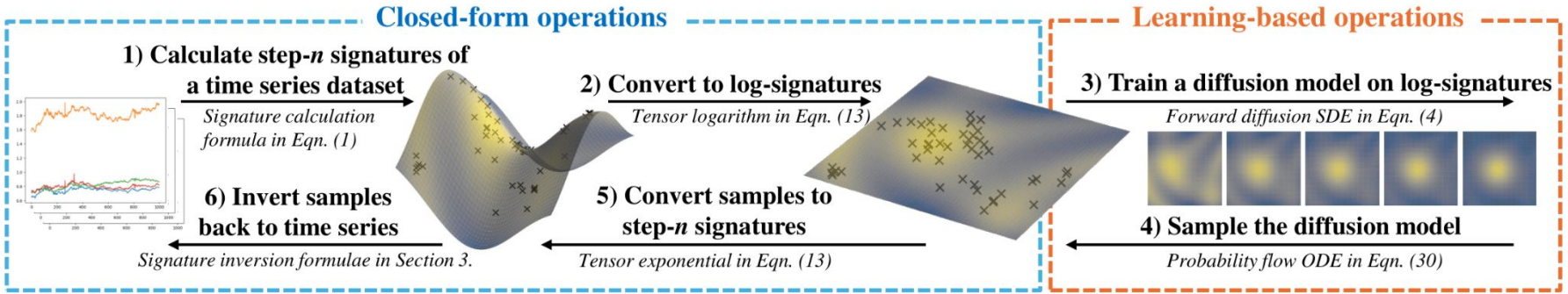
Модель	Accuracy	ROC-AUC
NN(1)	0.81	0.88
NN(2)	0.81	0.89
NN(3)	0.81	0.89
NN(4)	0.81	0.89
NN(5)	0.81	0.89

Модель	Accuracy	ROC-AUC
NN(1)	0.68	0.74
NN(2)	0.77	0.84
NN(3)	0.78	0.86
NN(4)	0.78	0.86
NN(5)	0.77	0.85

Что в мире?

Published as a conference paper at ICLR 2025

SIGDIFFUSIONS: SCORE-BASED DIFFUSION MODELS FOR TIME SERIES VIA LOG-SIGNATURE EMBEDDINGS



Планы

1. Попробовать использовать сигнатуру как предобработчик данных для зарекомендовавших себя моделей по поиску аномалий(TimesNet: temporal 2d-variation modeling for general time series analysis).
2. Исследовать базис сигнатуры в shuffle-алгебре порождаемой словами Линдона.

Theorem 1.14 (Shuffle product identity). *Consider a path* $X : [a, b] \to \mathbb{R}^d$ *and two multi-indexes* $I = (i_1, \ldots, i_k)$ *and* $J = (j_1, \ldots, j_m)$ *with* $i_1, \ldots, i_k, j_1, \ldots, j_m \in \{1, \ldots, d\}$. *Then*

$$S(X)^I_{a,b} S(X)^J_{a,b} = \sum_{K \in I \sqcup\!\sqcup J} S(X)^K_{a,b}. \tag{13}$$

Спасибо за внимание!