Автономная некоммерческая образовательная организация
высшего образования «Сколковский институт науки и
технологий»

На правах рукописи

Романенкова Евгения Дмитриевна

Методы обучения представлений для оптимальных процедур детектирования разладок

Специальность: 1.2.1. Искусственный интеллект и машинное обучение

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата физико-математических наук

Москва — 2024

Работа выполнена в Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий».

Научный руководитель: **Зайцев Алексей Алексеевич**, кандидат физико-математических наук

Защита состоится **18 июня 2025 г. в 12 часов 00 минут** на заседании диссертационного совета **1.2.1.1.**, созданного на базе Автономной некоммерческой образовательной организации высшего образования «Сколковский институт науки и технологий» (Сколтех)

по адресу: Территория Инновационного Центра «Сколково», Большой бульвар д.30, стр.1, Москва 121205

С диссертацией можно ознакомиться в библиотеке Сколтеха и на сайте организации https://dissovet.skoltech.ru/

Автореферат разослан «_____»_____2025 года.

Ученый секретарь диссертационного совета
к.ф.-м.н. Копелевич Григорий Александрович

Autonomous Non-Profit Organization for Higher Education
"Skolkovo Institute of Science and Technology"

As a manuscript

Romanenkova Evgenia Dmitrievna

Representation learning methods for optimal change point detection procedures

Speciality: 1.2.1. Artificial intelligence and machine learning

DISSERTATION ABSTRACT
of the dissertation for the Degree of
Doctor of Philosophy in Physics and Mathematics

Moscow — 2024

The work has been performed at the Autonomous Non-Profit Organization for Higher Education "Skolkovo Institute of Science and Technology".

Scientific supervisor: **Zaytsev Alexey Alexeevich**,
PhD in Physics and Mathematics

The defense will take place on **18 June 2025 at 12:00 p.m.** at the meeting of the Dissertation Council **1.2.1.1.**, based at Skolkovo Institute of Science and Technology (Skoltech)

address: Skolkovo Institute of Science and Technology, the territory of the Innovation Center "Skolkovo", Bolshoy Boulevard, 30, bld.1, Moscow 121205, Russia

The text of the dissertation is available at the Skoltech library or on the website https://dissovet.skoltech.ru/

Dissertation Abstract was sent on «_____»_____2025 .

Academic secretary of the
Dissertation Council
Candidate of Physical and
Mathematical Sciences Kopelevich Grigory Alexandrovich

General characteristics of the work

Relevance of the work. Modern industry uses complicated systems that continuously work online and are vital for the well-being of large companies and humankind in general. Sudden collapses or prolonged unavailabilities of systems lead to significant losses to business owners. Thus, it is essential to detect deviations or, in other words, *change points* in a system's behavior as fast and accurately as possible.

Typically, data for detection come from a sequential stream represented as either multivariate vectors from sensors or so-called semi-structured data such as videos and image sequences [1; 2]. We can also describe disorders, ruptures, shifts, and other changes in sequential data as point anomalies [3]. Currently, the anomaly detection area for sequential data concentrates on applying machine learning and deep learning techniques. It adopts widespread methods while considering more specific models and loss functions. Most existing methods from this area concentrate on classifying whole sequences, whether they contain anomalies or are normal [2]. However, this problem statement does not reflect industrial needs, where the reaction to an anomaly should be as close in time to its moment as possible to prevent unexpected expenses.

An alternative approach comes from the Change Point Detection (CPD) field: we aim to directly minimize the delay of disorder detection while maintaining the number of false alarms low [3]. Unlike the anomaly detection formulation, the CPD is naturally suited for various monitoring systems where the cost of waiting for reaction is high [4; 5]

Change points can also appear in normal processes, defining natural alteration of the underlying systems. Understanding the process regime segments and their shift moments is necessary to organize work properly. In this case, the objective is a precise segmentation of a sequence [1; 6].

Some methods can deal with different simple CPD problems [3]. Under strong assumptions, theoretically optimal solutions are available [1; 7]. These approaches are still popular and have

tremendous real-world applications, including heavy industrial sensor management, control of high-loaded systems, financial data analysis, health monitoring, genome sequencing, and others [4; 5]. The key particularity is the usage of a detector «as it is» while ignoring the violation of assumptions and the specificity of the considered task. For example, the paper [4] uses the standard CUSUM procedure on top of residuals from autoregressive model ARIMA for monitoring rotary machines. However, to provide theoretical optimality, CUSUM assumes independence of the input data [8], which is generally not true for ARIMA model residuals. An additional gap is that there are little or no theoretical guarantees in realistic high-dimensional cases such as controlling video streams [2].

Nowadays, authors adapt classic theory to stream data of extreme dimensionality when the vector size is much greater than the sequence length. This problem statement poses the overall difficulty of theoretical analysis: the appearance of the curse of dimensionality, considering many noise features and dealing with computational issues connected with vector size [9—11]. For example, the authors [10] suggest that the detection delay is of order $\sqrt{D}$ for traditional parametric models [8], where D is a single observation dimension. This delay is obviously prohibitively high for many real problems. Such a special real-world task can be video surveillance, where one operates with data containing several thousands of features. To overcome the above issue, recent works effort to bound the class of parametric models and concentrate on particular high-dimensional cases: multivariate Gaussian vectors [10], various graph representations [9], e.t.c. Another possible solution for CPD, such as an application of non-parametric CPD approaches [7], does not assume any specific model structure. However, it often relies on techniques that also may be limited in high-dimensional settings: density estimation [12], choice of optimal kernel [13], or the consideration of long-run dependency [11]. While additional assumptions like sparsity allow for reducing the delay, practitioners go a different road, hoping that simpler models can learn to detect changes.

A natural impulse is to use the available historical information to create a data-based change point detector via the representation

learning process [14]. To be precise, representation learning aims to obtain compact low-dimensional data embeddings by incorporating the maximum amount of information relevant to the application task at hand. In relation to sequential data, the task is challenging, as we should take into account the context, but it seems to be the only way to get high-performing models. Some modern studies use this paradigm for change point detection while still considering low-dimensional time series [12; 13; 15; 16]. Moreover, authors attempt to adapt solutions from other fields like self-supervised learning [15—17] or generative models [13] and do not incorporate the theoretical nature of the problem into a learning process.

To sum up, theoretical-based approaches related to change detection often operate with too restrictive assumptions and have a limited ability to work with complex data. The modern CPD suggests a representation learning procedures as an answer while still concentrating on simple time series. Both ways often consider general, vague solutions, ignoring the task specificity of real industrial problems. CPD-related tasks also appear in sophisticated data modalities with hundreds and thousands of features, like controlling images and video streams. Existing methods from anomaly detection, closest to the CPD area, deal with these structures through efficient neural network-based solutions with corresponding representations. Nevertheless, they rarely can point at precise change moments online, which is crucial for many applications. Therefore, the CPD area needs a theoretical-grounded representation learning scheme for data of different dimensionality and an understanding of how to involve concrete task particularities.

The aim of the work is to develop a change point detection framework suitable for data of different modalities. As for such data, we consider diverse datasets with low and moderate dimensionality from industrial applications and complex high-dimensional sequences such as video streams.

To achieve the stated aim, it is necessary to solve the following **tasks**:

1. To investigate well-established approaches for change point detection on time series of moderate dimensionality and to apply

them to real industrial problems, taking into account data specificity.

2. To develop optimal change point detection procedures for high--dimensional data and to create the corresponding software implementation.
3. To incorporate the representation learning framework for change point detection-related problems as well as for industrial sensors data and more complex video data.
4. To design a proper validation procedure for testing developed methods based on open datasets for data of low and moderate dimensionality. For testing on high-dimensional data, to collect a dataset of change points in videos with the corresponding markup.

Scientific novelty. The majority of existing research in the current change point detection area concentrates on unsophisticated numerical time series with low and moderate dimensionality. At the same time, examples of the change point detection application are wider and include complex high-dimensional structures, such as the incident registration for surveillance video. Recent approaches to processing sequential data are often based on representation learning strategies. However, there is an apparent lack of a representation-learning procedure that provides a theoretically motivated, efficient change point detection method suitable for complex data structures.

On the other hand, existing change point detection methods help to solve real industrial tasks connected with time series of low dimensionality. Nevertheless, they suffer from strong assumptions that are often violated in real-world data. Thus, incorporating industrial domain specifics into the framework of change point detection is also a crucial task.

The gaps mentioned above in the current research area stage lead to claim the following novelty of the results:

1. The work introduces a comprehensive framework for change point detection based on representation learning depicted in Figure 1. Such methods for change point detection were understudied before despite their proven effectiveness in other fields.

The proposed methods offer a flexible toolkit that allows users to efficiently work with sequential data of varying dimensions, from low to high. Moreover, the framework can be applied both in supervised and unsupervised scenarios, making it a versatile solution for different types of data analysis.

2. For the first time, the change point detection paradigm is used to enhance the traditional machine learning approaches. The suggested method ensures high-quality supervised change point detection while minimizing false alarms. The developed solution has been tested for a real industrial task, resulting in a new model for rock-type classification based on logging while drilling data.
3. A novel loss function for change point detection via neural networks is developed. The presented loss function explicitly controls detection delay and time to false alarms. Such an approach opens up new opportunities for detecting change points in video data that have not been previously researched. Moreover, the first markup for solving the CPD problem on videos for further research has been presented.
4. The dissertation contributes to the area of self-supervised learning for change point detection by developing a model for obtaining universal representations. Unlike previous studies, the presented model requires no specific modifications to address change point detection, which broadens its practical applications. For the first time, the thesis findings demonstrate a strong correlation between the qualities of similarity estimation and change point detection, highlighting the effectiveness of the proposed method. The solution has been successfully adapted to the practical industrial task of well-log analysis, which resulted in the development of a new approach for detecting similar and changing patterns in well-logs.

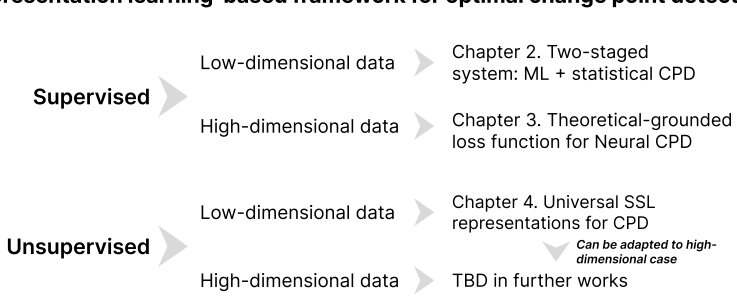

Figure 1 — A scheme of our framework for optimal change point detection based on representation learning. The proposed solution works effectively regardless of the expert's markup availability and input data dimensionality.

Scientific and practical significance. From a theoretical value point of view, the work provides a new loss function that is the lower bound for the classic rigorous solution. This also has important implications in practical applications, as it opens up new avenues for efficient solutions related to data of different dimensions and modalities. Another focus of the presented work is the data of low dimensionality from the industry and the corresponding tasks. Understanding how the change point detection methods can be applied to actual industrial problems and which data representation is necessary enforces the quality of real processes.

Research methodology. The research focuses on machine learning and deep learning techniques to address the stated objectives. More precisely, we investigate supervised and self-supervised paradigms for learning representations suitable for change point detection problems. To work accurately and efficiently, the proposed approaches rely on the statistical theory of change point detection methods and corresponding developments. Precise mathematical language allows the concise formulation of obtained results regarding their properties. Geological experts support the research, providing an understanding of oil&gas domain specificity in the

corresponding thesis chapters. The methods are implemented in Python programming language and its low-level libraries.

Propositions submitted for the defense:

1. A consistent representation learning framework for change point detection is developed. The solution effectively generates representations for multivariate time series data, as well as high-dimensional data, regardless of the availability of labels. The approach has solid theoretical properties under natural assumptions and has been validated through diverse real-world applications, including human activity recognition, rock-type detection during drilling, well-log analysis, and video surveillance.
2. In the context of low-dimensional data, a novel two-stage system for change point detection is introduced. In the first step, the machine learning classifier is trained to output change point probabilities. Then, by leveraging these probabilities as representations, the new detection statistic accurately identifies specific types of change points. Empirical and theoretical analyses confirm that this approach significantly reduces the number of false alarms compared to existing baselines. In practical scenarios, it achieved a tenfold reduction in false alarms and detection delay, which substantially improved existing methods.
3. The dissertation presents a theoretically grounded loss function for change point detection through supervised representation learning. Following the genuine criteria for the detection delay and time to false alarm, the proposed differentiable loss approximates the corresponding infinite sums with their truncated versions and novel correction terms. A presented theoretical analysis establishes two key properties of the suggested approximation: firstly, it provides a valid lower bound for the true criteria, and secondly, it is asymptotically tight. Experimental evidence complements this analysis, suggesting that deep networks trained with the proposed loss function ensure robust performance across various scenarios from low-dimensional to ultra-high-dimensional data.

4. Finally, the dissertation introduces a method for constructing representations based on self-supervised similarity learning without using expert markup. By adapting and refining existing techniques based on self-supervised loss function, this method can handle the intricacies of multivariate time series data, particularly those with complex, non-i.i.d. characteristics. Obtained embeddings have a broad range of applicability - and, in particular, allow unsupervised change point detection, completing the proposed framework.

The validity and reliability of the obtained results and conclusions are provided by comprehensive validation procedures that are in accordance with the leading works in the area. We also rely on the precise methods used to investigate the properties of change point detection methods, utilizing common assumptions in this field. All approaches are tested from different points of view, including various metrics groups, usage of several datasets (if applicable), and considering additional downstream tasks.

Approbation of the results. The results and main provisions of the work were presented at international and Russian conferences:

1. SPE Annual Technical Conference and Exhibition, Calgary, Alberta, Canada, September 2019;
2. «Information Technologies and Systems» Interdisciplinary Conference, online, November 2021;
3. 30th ACM International Conference on Multimedia (A* rank), Portugal, Lisboa, October 2022;
4. 64-th Russian Conference of MIPT, Moscow, Russia, November 2022;
5. 65-ve Russian Conference of MIPT, Moscow, Russia, April 2023;
6. Workshop «Methods of Artificial Intelligence for Industrial Tasks of Sustainable Development», Belokurikha, Russia, August 2023;
7. International Conference «Complex time series analysis: high-dimensionality, change-points, forecasting and causality», Sanya, Hainan, China, January 2024;

8. International Conference «Data Fusion» 2024, Moscow, Russia, April 2024.

Personal contribution. The content of the dissertation and the main statement submitted for defense reflect the author's personal contribution to the published works and have been received personally by the author or with her direct participation. The author's contribution includes setting research objectives, studying literature corresponding to the topic, designing and carrying out experiments, and analyzing the received results. The problem statements were formulated by A. Zaytsev, the author's scientific supervisor, and the research results were discussed with the co-authors.

Publications. The publications in peer-reviewed scientific journals also support the credibility of research results. The dissertation materials were published in 6 works, including 3 papers in Q1 scientific journals indexed by Web of Science and Scopus, 1 paper in Q2 journal indexed by Web of Science and Scopus, 1 extended abstract at international conferences, and 1 article in the proceedings of the A* conference also indexed by Scopus. In 4 publications, the author of the dissertation is the first author or has a shared contribution with a first author. In addition, 1 computer program was registered by the state.

Dissertation Summary

The **Introduction** substantiates the relevance of the work, defines the research goals and methodology, outlines the scientific novelty and theoretical and practical values of the dissertation, and formulates the main results.

The chapter 1, **Related Work**, provides an overview of the current state of research in the CPD area, highlighting challenges associated with high-dimensional data. It also covers existing solutions for the task at hand based on representation learning. Additionally, we discuss the domain-specific related works from the

oil&gas industry that are relevant to the particular tasks for application of our framework.

The proposed CPD framework presentation starts in chapter 2, **Change Point Detection via Machine Learning Classifiers**, where we suggest a new two-stage method for supervised CPD in a low-dimensional setting illustrated in Figure 2. More precisely, the initial signal is embedded into one-dimensional representations space (probabilities) via a classifier, where we apply additional detectors of changes.

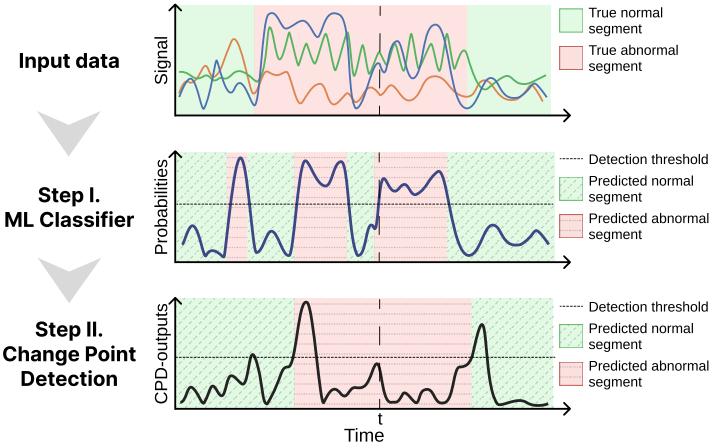

Figure 2 — A scheme of proposed two-staged approach. In the first step, the input signal is embedded into 1D representations via an ML classifier. Then, we run a change point detection method on top of the obtained probabilities.

Formally, we consider a multivariate signal of independent observations $\{\mathbf{x}_i\}_{i=1}^{t}$, $\mathbf{x}_i \in \mathbb{R}^d$ of an arbitrary length t. The signal switches between $K = 2$ regimes: a normal one, with the distribution function F_0, and an abnormal one governed by F_1. The moments of alteration between regimes $\mathcal{T} = \{\tau_j : \mathbf{x}_{\tau_j - 1} \sim F_a \text{ and } \mathbf{x}_{\tau_j} \sim F_b, a \neq b, a, b \in \{0, 1\}\}$ are change points. Our goal for CPD is to identify $\mathcal{T}$ as accurately as possible. Also, the CPD task can be formulated as a binary classification problem. For this

case, the labels are:

$$y_i = 0, \text{ if } \mathbf{x}_i \sim F_0 \text{ and } y_i = 1, \text{ if } \mathbf{x}_i \sim F_1. \tag{1}$$

Given these labels, we train a supervised model $f_{\mathbf{w}}$ on a historical dataset (Step I in Figure 2). In this context, $f_{\mathbf{w}}(\mathbf{x}_i) = p_i$, where $p_i \in [0, 1]$ is the estimated probability that $\mathbf{x}_i$ comes from F_1 at the moment i ($y_i = 1$), and $\mathbf{w}$ is a vector of the model's parameters. During Step II in Figure 2, we use these representations (probabilities) as the inputs to the statistics $S_i = S(\mathbf{l}_{\leq i})$ based on likelihoods $l_i = l(p_i)$. Such statistics include CUSUM, Shiryaev-Roberts, and Posterior probability ones [7].

The majority of traditional ML classifiers do not restrict the predicted length for a segment, so the output probabilities may oscillate near shift moments, producing a lot of false alarms and unrealistic short intervals of x_i that come from a specified F_1. To address this issue, we have designed a novel method of cutting thin layer statistics (CTL). First, based on the likelihood for $(t + 1)$ we calculate statistic S_{t+1} in the following way:

$$S_{t+1} = \begin{cases} S_t + 1, \ l_{t+1} \geq l_0, \\ 0, \ l_{t+1} < l_0, \end{cases} \tag{2}$$

where l_t is a likelihood from classifier $f_{\mathbf{w}}$, and l_0 is a threshold. Second, we compare the obtained statistics with w. If $S_t > w$, we set all the previously predicted labels $y_{t-w}, \ldots, y_t$ to the value of the label y_{t-w-1}. The only hyperparameters of the CTL are an upper bound of the layer size to drop w and threshold l_0.

Traditional statistical-based approaches limit the number of false alarms and tend to reduce the detection delay [7; 8]. In contrast, CTL offers an opposite way: limit the detection delay from below with a hyperparameter w while minimizing the number of false alarms.

Lemma 1. *Assume a model $f_{\mathbf{w}}$ provides independent likelihoods l_t for each input moment t. Using the statistic of cutting thin layer defined in Equation* (2) *decreases the models' false positive errors compared to the statistic without the aggregation from* (2).

Table 1 — The results are averaged with the median for 57 wells. TP stands for the number of True Positive errors, FP is the number of False Positive errors, and Acc N. is the specific metric for the considered applied problem. The best values are in **bold** font. Cutting thin layers does not outperform other approaches in all metrics, but it improves target Accuracy N and FP.

Used method	Acc. N ↑	Mean detect. delay ↓	TP ↑	FP ↓
Only classification model	0.5511	**0.3**	**25**	43
Cumulative sums	0.6551	4.0	8	**6**
Shiryaev-Roberts stat.	0.6400	3.7	6	7
Priory distribution stat.	0.6207	3.5	6	12
Cutting thin layers (ours)	**0.6579**	1.8	12	**6**

We evaluate our two-stage pipeline for the real industrial challenge of rock type classification during drilling, where changes happen between two states: oil-bearing zones and non-productive areas. This problem aligns with our expectations. We are dealing with noisy, low-dimensional sensor time series, where existing methods have proven inaccurate. Moreover, the potential economic benefits of a successful solution are substantial.

The approach is compared with the pure classifier model as well as other classic likelihood-based methods from the CPD area [8]. For validation, we use standard criteria for the CPD area [8], classification problem, and domain-specific metrics; see details in the full text of the dissertation. As shown in Table 1, the presented system provides only a 1.8 m mean change detection delay and has about 6 false alarms per well, which is 7 times lower than the pure classification.

In the chapter 3, **Instant Disorder Detection via a Principled Neural Network**, we present the second part of the framework for solving the CPD problem for complex semi-structured high-dimensional data. In contrast to the previous solution, our approach is suitable for data of different modalities, including images and video streams. The method detects changes in the online mode and makes a decision based on the information available

at the current moment in case of a single and multiple change points to detect. The scheme of the approach is presented in Figure 3.

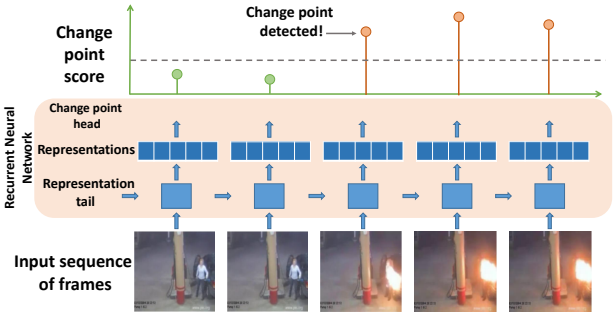

Figure 3 — A scheme of work for representation-based change point detection approach

Following traditional paradigm [8], our goal is to find a procedure that produces an estimate τ^* of the true change point θ. Such estimation should minimize the expected detection delay (DD) with the constraint that the Average Time to False alarm (Time to FA) is greater than a large value a:

$$\begin{gathered}\tilde{\mathcal{L}}(\tau) \to \min_{\tau} \text{ s.t. } c(\mathbb{E}_\theta(\tau |\ \tau < \theta) - a) = 0, \\ \text{where } \tilde{\mathcal{L}}(\tau) = \mathbb{E}_\theta(\tau - \theta)^+ - c\mathbb{E}_\theta(\tau |\ \tau < \theta).\end{gathered} \tag{3}$$

Here the expectation with index θ means that observations come from the data distribution p_θ with the true change point at time θ, $c \geq 0$ — Lagrange multiplier.

Suppose, we observe a set of sequences $D = \{(X_k, \theta_k)\}_{k=1}^N$ with similar length T, change points θ_i are in $\{1, \ldots, T, \infty\}$. Each X_i is a sequence of multivariate vectors $\{\mathbf{x}_{ij}\}_{j=1}^T$. We construct a data-driven model $f_\mathbf{w}$ that optimizes the criterion (3) for the online change point detection procedure. This model $f_\mathbf{w}$ outputs a series $\{p_t^i\}_{t=1}^T$ based only on the information available up to the moment t and $p_t^i = f_\mathbf{w}(X_i^{1:t})$. In addition, we should define the natural behavior of the "ideal" model $f_\mathbf{w}$:

Assumption 1. *The good model $f_{\mathbf{w}}$ should output $p_t^i = q$ for $t < \theta_i$ and $p_t^i = 1 - q$ for $t \geq \theta_i$, where $0 < q < 1$ and $q \to 0$.*

Assumption 2. *For $t > T$, the output probabilities of $f_{\mathbf{w}}$ follow geometric distribution with a hyperparameter r, $p_t \sim \text{Geom}(r)$ $\forall t > T$.*

To provide a representation learning process necessary to CPD on complex data streams [14], we suggest a new principled loss function $\tilde{L}^h(f_{\mathbf{w}}, D, c)$ for training a model $f_{\mathbf{w}}$:

$$\tilde{L}^h(f_{\mathbf{w}}, D, c) = \tilde{L}^h_{delay}(f_{\mathbf{w}}, D) - c\tilde{L}_{FA}(f_{\mathbf{w}}, D), \tag{4}$$

This loss consists of two terms $\tilde{L}^h_{delay}(f_{\mathbf{w}}, X_i, \theta_i)$ and $\tilde{L}_{FA}(f_{\mathbf{w}}, X_i, \theta_i)$ averaged over all sequences. Both of them are connected with principled CPD criteria. Here $\tilde{L}^h_{delay}(f_{\mathbf{w}}, X_i, \theta_i)$ approximates the detection delay:

$$\tilde{L}^h_{delay}(f_{\mathbf{w}}, X_i, \theta_i) = \sum_{t=\theta_i}^{h}(t-\theta_i)p_t^i \prod_{k=\theta_i}^{t-1}(1-p_k^i) + (h+1-\theta_i)\prod_{k=\theta_i}^{h}(1-p_k^i), \tag{5}$$

where $h \leq T$ is a hyperparameter that restricts the size of a considered segment. More precisely, we have proven that our formulation provides an asymptotically tight lower bound with a power-law convergence rate:

Lemma 2. *$\tilde{L}^h_{delay}(f_{\mathbf{w}}, D)$ is a lower bound for the expected value of the detection delay $\mathbb{E}_\theta(\tau - \theta)^+$.*

Lemma 3. *Under Assumption 1, model $f_{\mathbf{w}}$ guarantees an asymptotically tight estimation $\tilde{L}^h_{delay}(f_{\mathbf{w}}, D)$ for the expected detection delay $\mathbb{E}_\theta(\tau - \theta)^+$.*

Lemma 4. *The convergence rate for $\tilde{L}^h_{delay}(f_{\mathbf{w}}, X_i, \theta)$ is $\frac{q}{1-q}(1 - q^{h-\theta+1})$ under Assumption 1.*

Similarly, $\tilde{L}_{FA}(f_{\mathbf{w}}, X_i, \theta_i)$ approximates the expected time to false alarm looking at the interval with no changes either $[0, \theta_i]$ or $[0, T]$, if there are no change point:

$$\tilde{L}_{FA}(f_{\mathbf{w}}, X_i, \theta_i) = \sum_{t=1}^{\tilde{T}_i} t p_t^i \prod_{k=1}^{t-1}(1 - p_k^i) + \left(\tilde{T}_i + \frac{1}{r}\right) \prod_{k=1}^{\tilde{T}_i}(1 - p_k^i), \quad (6)$$

where $\tilde{T}_i = \min(\theta_i, T)$, and $0 < r \leq 1$ is a hyperparameter that controls detection at «infinity» (2). We have shown that it is an exact second term from criteria (3):

Lemma 5. *Under Assumption 2, the expected value of the time to the false alarm* $\mathbb{E}_\theta(\tau |\ \tau < \theta)$ *for model* $f_{\mathbf{w}}$ *is* $\tilde{L}_{FA}(f_{\mathbf{w}}, D)$.

All supplementary lemmas lead to the main theorem, which proves that our loss function is a tight lower bound of the principled Lagrangian function (3).

Theorem 6. *The loss function* $\tilde{L}^h(f_{\mathbf{w}}, D, c)$ *from* (4)*:*

(1) is a lower bound for a Lagrangian for $\tilde{\mathcal{L}}(\tau)$ *from criterion* (3)

(2) is differentiable with respect to p_k^i *and, thus,* $\mathbf{w}$*;*

(3) is an asymptotically tight lower bound with a power-law convergence rate.

To provide a wide-applicable representation model, we choose a Recurrent Neural Network (RNN) as $f_{\mathbf{w}}$. It processes inputs sequentially: hidden states would represent the current sequence tailored for CPD. The model is trained through minimizing the differentiable loss $\tilde{L}^h(f_{\mathbf{w}}, D, c)$ from (4). We either train the model from scratch using the proposed loss to get the pure *InDiD* method or train a classifier with binary cross-entropy loss (BCE) and then fine-tune it with our loss to get *BCE+InDiD* method. Alternatively, we consider classic CPD methods from [6] (where they are applicable), strong baseline with BCE loss [17], and state-of-the-art models KL-CPD [13] and TSCP2 [15].

We validate our approach on a diverse set of datasets, including synthetically generated sets, sensor data, image sequences, and video streams. Moreover, for the first time, we have provided a markup for CPD on video datasets that, as we hope, forces the development of the field. The considered metrics are a traditional set of criteria for the CPD [1]; see more details in the full thesis text. As shown in Table 2, experimental evidence suggests that usage of our loss function (4) leads to an overall model improvement. We observe this behavior in various scenarios, and for the most complex video data, InDiD outperforms baselines by approximately 1.5 times in terms of F1 metrics. It is important to note that the results of the classic approach are not included in Table 2, as they are limited to high-dimensional data. More detailed results can be found in the full version of the dissertation.

Table 2 — Mean performance ranks of considered methods averaged over datasets. The results are averaged by 5 datasets. The best values are highlighted with **bold** font, and the second ones are underlined.

Metric	AUC	F1	Cover
KL-CPD [13]	4.17	4.17	3.50
TSCP [15]	4.67	3.83	4.66
BCE	3	2.17	2.17
InDiD (ours)	**1.5**	**1.67**	**1.5**
BCE+InDiD (ours)	1.67	2	2

Chapter 4, **Detecting Akin and Changing Patterns Using Self-Supervised Similarity Learning**, proposes an unsupervised encoder model that produces representations suitable to a wide range of tasks, including change point detection. By adapting and refining existing ideas from the self-supervised learning (SSL) area to real-world sequential sensor data, our approach allows for flexible and effective distinguishing between similar and dissimilar subsequences, as shown in Figure 4. As a practical example of real-world sensor data, we consider well-log observations from the oil&gas domain.

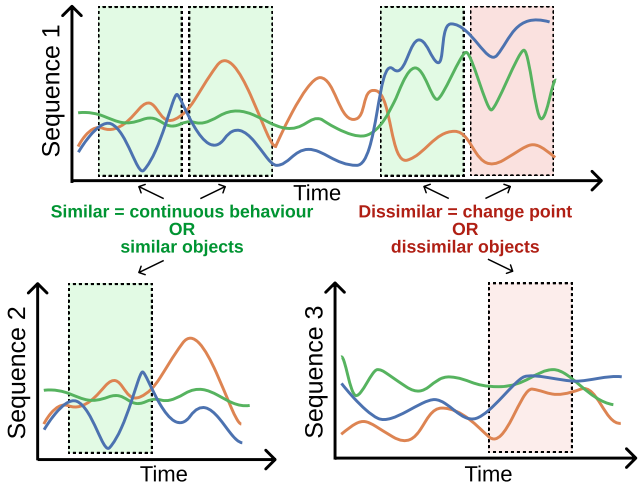

Figure 4 — An overview of the proposed approach. With the model of intervals' similarity, we can compare sensor observations from different streams and identify the regime's change points while observing a single stream.

The methodology's core, depicted in Figure 4, is strongly dependent on the similarity measure learned by the implicit model. To obtain such a measure in the unsupervised scenario, we follow the SSL paradigm and derive specified similarity labels from the inherent data structure. Thus, we evaluate four SSL methods to train a neural network encoder:

1. classification-based approach when we force a *Siamese* model to solve *classification problem* (are objects from a pair similar or not) utilizing objects' representations;
2. contrastive *Triplet* model that learns similarity directly in the representations' space and further evaluates it through common distances;
3. generative *VAE* network that does not control similarity explicitly, but, as a generative model, provides informative local embeddings [18];
4. a combination of contrastive and generative models, denoted as the *Ensemble*. This method leverages the strengths of each paradigm, thereby enhancing overall performance.

Recent findings [18] also indicate that different scales of learned patterns, local and global, can be beneficial for solving different problems. Thus, we propose two labeling procedures for pairs of intervals for training classification-based and contrastive models: a global Linking method and a more local Close Linking shown in Figure 5.

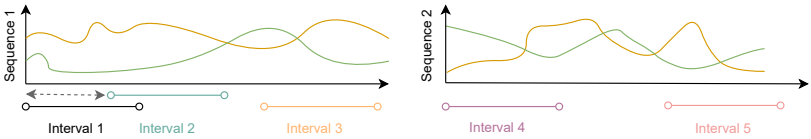

Figure 5 — Linking and Close Linking labeling example for five intervals $I_1, \ldots, I_5$ that belong to two different sequences. For **Linking**, (I_1, I_2), (I_1, I_3) are pairs of similar intervals, as they belong to the same sequence, while (I_1, I_4) and (I_1, I_5) are pairs of different intervals, as intervals belong to different objects. For **Close Linking**, all relations remain the same apart that now a pair (I_1, I_3) constitutes different intervals, as their difference in time is higher than a pre-specified threshold hyperparameter.

We utilize representations from the above SSL models to solve CPD tasks rather than working directly with the original data. This approach allows us to leverage the full expressive power of neural networks while overcoming the challenges associated with data structure complexity [15; 19]. We compare $s_{\cos}(\cdot)$, the cosine similarity, between embeddings of two consecutive intervals sampled from a stream X, with a threshold hyperparameter α. If the similarity is smaller than α, a change point is detected:

$$\hat{y}_{\text{cp}}^{i+m+l} = \mathbb{I}[s_{\cos}(\mathbf{e}_i, \mathbf{e}_{i+m+l+1}) < \alpha], \tag{7}$$

where $\mathbf{e}_k = f(X_{k:k+l})$ denotes an embedding of $X_{k:k+l}$ via the model $f(\cdot)$, l is the interval length, m is a margin between two consecutive intervals. By sliding through the entire sequence X, we obtain a set of predictions. For evaluation, we formulate a CPD problem as a classification task: if an interval $X_{i:i+m+2l}$ contains a change

in target labeling, the true CPD target $y_{\text{cp}}^{i+m+l} = 1$; otherwise, $y_{\text{cp}}^{i+m+l} = 0$ — and we want $y_{\text{cp}}^{i+m+l} \approx \hat{y}_{\text{cp}}^{i+m+l}$.

During our experiments, we evaluate a wide range of embeddings' characteristics: correct estimation of natural similarity measures, compliance with expert markup (CLASS and CLASS + LAYER) for well's closeness, and sensitivity to physical-guided changes in data. More results are in the full text. For the first experiment, we directly evaluate similarities from the model with common classification metrics as validation criteria. Moreover, we cluster the provided embeddings and compare the obtained labeling with expert markup (CLASS and CLASS + LAYER) for the well's closeness using ARI measures. Finally, we estimate sequential F1 score from CPD works [15; 19]. More details on metrics are in the full text of the thesis.

Table 3 — Comparison of the models' quality for Close Linking problem. The best values are highlighted with **bold** font, the second ones are underlined.

Models	Accuracy	ROC AUC	PR AUC
XGBoost	0.787 ± 0.042	0.802 ± 0.041	0.895 ± 0.032
Euclidean distance	0.300 ± 0.015	0.310 ± 0.034	0.266 ± 0.020
Cosine distance	0.596 ± 0.048	0.674 ± 0.045	0.530 ± 0.049
Siamese + 3 FC	0.814 ± 0.053	**0.874 ± 0.056**	0.766 ± 0.126
Siamese + Cos. dist.	0.603 ± 0.061	0.710 ± 0.070	0.592 ± 0.076
Triplet + Cos. dist	**0.926 ± 0.022**	0.783 ± 0.036	0.926 ± 0.020
VAE + Cos. dist	0.613 ± 0.111	0.738 ± 0.021	0.826 ± 0.096
Ensemble + Cos. dist	0.853 ± 0.139	0.813 ± 0.033	**0.939 ± 0.109**

Table 4 — Comparison of the clustering ARI metrics for embeddings from models trained in Close Linking setting. The best values are highlighted with **bold** font, and the second ones are underlined.

Models	CLASS	CLASS + LAYER
Feature clustering	0.340± 0.087	0.340± 0.087
Siamese + 3FC	0.416± 0.147	**0.416± 0.147**
Triplet	0.618± 0.225	0.361± 0.064
VAE	0.603± 0.122	0.348± 0.121
Ensemble	**0.660± 0.157**	0.386± 0.101

Our results from Table 3 and Table 4 suggest that there is no apparel leader between individual deep models. Both Siamese and Triplet networks provide good results, while the generative VAE approach lags behind. Nevertheless, its combination with the Triplet model provides an outstanding solution for the Close Linking task and correlates with the expert's understanding of natural well similarity.

To test the effectivity of models for complex tasks of CPD, we consider two problem statements: detection of changes in stratigraphic layers and identification of rock type shifts. Both properties provide valuable information during oilfield development. The quantitative results of CPD performance for both tasks are presented in Table 5. Figure 6 provides two qualitative examples of a model's behavior for different CPD problems.

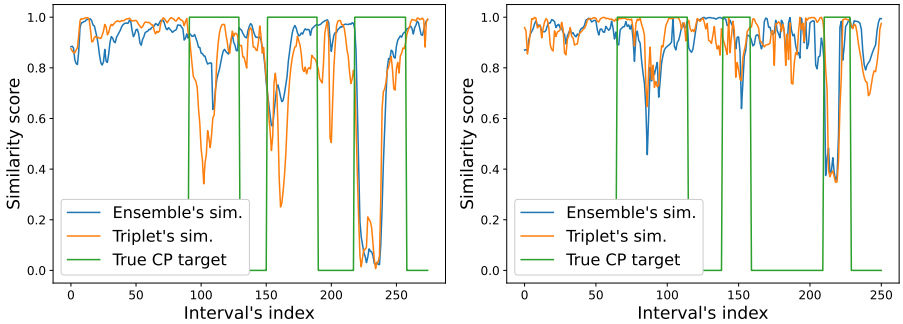

Figure 6 — Two examples of Change Point Detection via Ensemble's embeddings. The left Figure corresponds to the detection of changes in stratigraphical layers, while the right — to the change of lithotypes. The ensemble model can effectively capture pattern alternations while rapidly decreasing the similarity between embeddings and providing fewer false alarms.

We can conclude that the performance of models depends on at least two criteria: the model's type and the quality of similarity estimation. Generative models, such as VAE, are known to provide better dynamics of embeddings but may underperform in discriminative tasks [19], and our experiments confirm it. VAE has

a reasonable but moderate quality for similarity estimation while achieving a Top-2 performance for CPD. On the other hand, as expected, classification-based Siamese and contrastive Triplet networks surpass the VAE for the CPD problems while competing with each other in the case of similarity tasks. The Ensemble model integrates the strengths of two worlds, outperforming both baselines and individual models for each task. This finding aligns with our initial hypothesis: this model exhibits the best performance in discriminative similarity estimation, as shown in Tables 3,4.

Table 5 — F1 score for Change Point Detection based on embeddings from the best Ensemble model for two considered tasks. The best values are highlighted with **bold** font, the second ones are underlined.

	Str. Layers	Lithotypes
Feature-based	0.342 ± 0.004	0.548 ± 0.081
Kernel CPD [20]	0.370 ± 0.009	0.476 ± 0.053
TSCP2 [15]	0.387 ± 0.019	0.511 ± 0.075
Siamese	0.415 ± 0.021	0.580 ± 0.060
Triplet	0.419 ± 0.017	0.569 ± 0.048
VAE	0.426 ± 0.042	0.596 ± 0.073
Ensemble	**0.454** ± **0.020**	**0.630** ± **0.056**

Thus, Ensemble complements the presented framework for effective CPD in unsupervised cases. Unlike CPD-oriented models, it provides flexible embeddings that can be applied to various tasks, from similarity learning to change detections. At the same time, we see that both tasks are highly correlated: in future works, one can choose the encoder model for CPD while estimating natural data similarity like the one provided in Figure 5. Finally, we can adapt such an approach to high-dimensional cases as well.

The last chapter **Conclusion** summarizes the results and discusses future works.

Conclusion

This thesis investigates the optimal change point detection procedures for data of varying dimensionality and presents a versatile framework based on representation-learning techniques to address this challenge. Through comprehensive analysis and experiments, we assess a diverse range of data types and settings, encompassing both supervised and unsupervised cases, ranging from real-world sensor data to videos with change points. Our research is grounded in both theoretical justification and practical applications. The main results are the following:

1. The dissertation presents a consistent representation learning framework for change point detection. Throughout our work, we have accurately justified that each part of our solutions effectively solves real-world CPD problems. Altogether, they provide a flexible tool suitable for different scenarios, from simple signals with moderate dimensionality to challenging high-dimensional CPD on video data, regardless of the availability of markup.
2. For low-dimensional streams, a framework suggests a two-stage system that combines the effectiveness of classic machine learning with the CPD-oriented approaches. This part is tested on the important industrial challenge of rock type identification based on information from sensors used during directional drilling. Our procedure reduces the detection delay from 20 to 1.8 meters and the number of false positive alarms from 43 to 6 per well. This technology has been implemented by a major oil and gas company, saving billions of rubles every year.
3. Our work presents a new theoretical-grounded universal loss function for supervised change point detection via representation learning. The suggested loss function approximates the classic criteria while balancing change detection delay and time to a false alarm. We have shown that it is a differentiable tight lower bound on the conventional criteria with a power-law convergence rate. According to the findings, the neural networks trained with our loss function provide meaningful representations tailored to the specific nature of the CPD and complex

data structures. Such a model can be applied to a wide range of data types, including sensor data and videos. For example, for explosion detection in video surveillance, the F1 score for our method is 0.53 compared to baseline scores of 0.31 and 0.35

4. We have also developed a model for generating universal representations through similarity learning in an unsupervised scenario. Unlike other CPD-oriented approaches, these embeddings can be applied to a broader spectrum of tasks. For instance, our method highly correlates with expert labeling of intra- and inter-well correlations with ARI and AMI as high as 0.660 and 0.727, respectively, compared to 0.340 and 0.447 for the baseline. At the same time, our approach effectively identifies the shifting of stratigraphical layers and lithotype changes. The F1 scores for these tasks are up to 0.454 and 0.630, respectively, outperforming the most state-of-the-art CPD approach, which achieves F1 scores of 0.387 and 0.511.

The presented results provide a significant step forward in studying various representation-learning techniques for change point detection tasks and open up new avenues for further refinement.

The obvious restriction of our two-staged ML-CPD system and InDiD approach is their supervised nature. Typically, there is a lack of annotated data, and the labeling process requires additional expenses and time-consuming expert efforts. We aim to fill this gap by presenting the unsupervised method based on similarity learning. However, the InDiD loss function can also be adapted to an unsupervised setting similar to work [17], where the authors considered all possible sequence partitions on two segments.

The dissertation insights have become the foundation for various works. Several colleagues' papers have already complemented our research on a fundamental representation-learning solution for industrial sequences. Moreover, we have evaluated which model type provides desired properties for different scales of tasks on real-world event sequences with irregular grid [18]. Finally, we have studied how to incorporate CP principles in similarity learning. More particularly, we have found that Spectral Normalization [19]

ensures shifts in embedding space, and Layer Normalization spreads information about the changes facilitating their detection.

References

1. *Burg, G. J. van den.* An evaluation of change point detection algorithms / G. J. van den Burg, C. K. Williams // arXiv preprint arXiv:2003.06222. — 2020.
2. *Sultani, W.* Real-world anomaly detection in surveillance videos / W. Sultani, C. Chen, M. Shah // Proceedings of the IEEE conference on computer vision and pattern recognition. — Salt Lake City, Utah, USA : IEEE, 2018. — P. 6479—6488.
3. Sequential (quickest) change detection: Classical results and new directions / L. Xie [et al.] // IEEE Journal on Selected Areas in Information Theory. — 2021. — Vol. 2, no. 2. — P. 494—514.
4. *Wang, T.* Multi-sensors based condition monitoring of rotary machines: An approach of multidimensional time-series analysis / T. Wang, G. Lu, P. Yan // Measurement. — 2019. — Vol. 134. — P. 326—335.
5. *Artemov, A.* Ensembles of detectors for online detection of transient changes / A. Artemov, E. Burnaev // Eighth International Conference on Machine Vision. Vol. 9875. — International Society for Optics, Photonics. 2015. — 98751Z.
6. *Truong, C.* Selective review of offline change point detection methods / C. Truong, L. Oudre, N. Vayatis // Signal Processing. — 2020. — Vol. 167. — P. 107299.
7. *Tartakovsky, A.* Sequential analysis: Hypothesis testing and changepoint detection / A. Tartakovsky, I. Nikiforov, M. Basseville. — Chapman & Hall, 2014. — 603 p.
8. *Shiryaev, A. N.* Stochastic Change-Point Detection Problems / A. N. Shiryaev. — Moscow : MCME, 2017. — 393 p.

9. *Chu, L.* Sequential change-point detection for high-dimensional and non-euclidean data / L. Chu, H. Chen // IEEE Transactions on Signal Processing. — 2022. — Vol. 70. — P. 4498—4511.

10. *Chen, Y.* High-dimensional, multiscale online changepoint detection / Y. Chen, T. Wang, R. J. Samworth // Journal of the Royal Statistical Society Series B: Statistical Methodology. — 2022. — Vol. 84, no. 1. — P. 234—266.

11. Change-point detection with feature selection in high-dimensional time-series data / M. Yamada [et al.] // Twenty-Third International Joint Conference on Artificial Intelligence. — 2013.

12. *Hushchyn, M.* Online Neural Networks for Change-Point Detection / M. Hushchyn, K. Arzymatov, D. Derkach // arXiv preprint arXiv:2010.01388. — 2020.

13. Kernel change-point detection with auxiliary deep generative models / W.-C. Chang [et al.] // International Conference on Learning Representations. — Vancouver, Canada : ICLR, 2018.

14. *Bengio, Y.* Representation learning: A review and new perspectives / Y. Bengio, A. Courville, P. Vincent // IEEE transactions on pattern analysis and machine intelligence. — 2013. — Vol. 35, no. 8. — P. 1798—1828.

15. Time Series Change Point Detection with Self-Supervised Contrastive Predictive Coding / S. Deldari [et al.] // Proceedings of The Web Conference 2021. — Association for Computing Machinery, 2021.

16. A self-supervised contrastive change point detection method for industrial time series / X. Bao [et al.] // Engineering Applications of Artificial Intelligence. — 2024. — Vol. 133. — P. 108217.

17. *Puchkin, N.* A contrastive approach to online change point detection / N. Puchkin, V. Shcherbakova // International Conference on Artificial Intelligence and Statistics. — PMLR. 2023. — P. 5686—5713.

18. Universal representations for financial transactional data: embracing local, global, and external contexts / A. Bazarova [et al.] // arXiv preprint arXiv:2404.02047. — 2024.

19. *Bazarova, A.* Normalizing self-supervised learning for provably reliable Change Point Detection / A. Bazarova, E. Romanenkova, A. Zaytsev // arXiv preprint arXiv:2410.13637. — 2024.

20. A kernel two-sample test / A. Gretton [et al.] // The Journal of Machine Learning Research. — 2012. — Vol. 13, no. 1. — P. 723—773.

Author's publications on the dissertation topic

1. **Romanenkova E.**, Stepikin A., Morozov M., Zaytsev A. InDiD: Instant Disorder Detection via a Principled Neural Network, *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, p. 3152—3162, DOI: 10.1145/3503161.3548182 (A* rank conference, proceedings, Scopus);
2. **Romanenkova E.**, Rogulina A., Shakirov A., Stulov N., Zaytsev A., Ismailova L., Kovalev D., Katterbauer K., AlShehri A. Similarity learning for wells based on logging data, *Journal of Petroleum Science and Engineering*, 2022, vol. 215, p. 110690, DOI: 10.1016/j.petrol.2022.110690 (Q1 journal, WoS and Scopus);
3. **Romanenkova E.**, Zaytsev. A., Klyuchnikov N., Gruzdev A., Antipova K., Ismailova L., Burnaev E., Semenikhin A., Koryabkin V., Simon I., Koroteev D., Real-time data-driven detection of the rock-type alteration during a directional drilling, *IEEE Geoscience and Remote Sensing Letters*, 2020, vol. 17, is. 11, p. 1861—1865, DOI: 10.1109/LGRS.2019.2959845 (Q1 journal, WoS and Scopus);
4. Gurina E., Klyuchnikov N., Zaytsev A., **Romanenkova E.**, Antipova K., Simon I., Makarov V., Koroteev D., Application of machine learning to accidents detection at directional drilling, *Journal of Petroleum Science and Engineering*, 2020, vol. 184, p. 106519, DOI: 10.1016/j.petrol.2019.106519 (Q1 journal, WoS and Scopus);
5. Zholobov V.[1], **Romanenkova E.**[1], Egorov S., Gevorgyan N., Zaytsev A., Universal representations for well-logging data via ensembling of self-supervised models, *Doklady Mathematics*, 2024, vol. 110, suppl. 1, p. S126 – S136, DOI: 10.1134/S1064562424602257 (Q2 journal, WoS and Scopus);
6. Antipova K., Klyuchnikov N., Zaytsev A., Gurina E., **Romanenkova E.**, Koroteev D., Data-driven model for the drilling

[1]shared contribution

accidents prediction, *SPE Annual Technical Conference and Exhibition*, 2019, DOI: 10.2118/195888-MS;
7. Certificate of state registration of computer programs No. 2023612725(RU), Antipova K., Klyuchnikov N., Zaytsev A., **Romanenkova E.**, Nyazhemetdinov R.R., 25 Jan 2023 (in Russian).