НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2025 Том 25 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
Jule–August 2025 Vol. 25 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2025-25-4-684-693
УДК 004.93

Выявление аномалий в условиях ограниченности и неопределенности данных с использованием zero-shot и few-shot подходов

Сергей Андреевич Милантьев[1]✉, Полина Дмитриевна Михайлова[2], Игорь Александрович Бессмертный[3]

[1,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация

[1] geerkus@gmail.com✉, https://orcid.org/0000-0002-1970-5217
[2] polina.delitzsch@gmail.com, https://orcid.org/0009-0005-9731-0105
[3] bessmertny@itmo.ru, https://orcid.org/0000-0001-6711-6399

Аннотация

Введение. Выявление аномалий в условиях ограниченного объема данных представляет собой актуальную задачу в различных прикладных областях, включая медицинскую диагностику. Методы машинного обучения обычно требуют наличия образцов аномалий для их выявления, что не всегда возможно. Существующие методы выявления аномалий при малом количестве (few-shot) или полном отсутствии (zero-shot) обучающих данных об аномалиях имеют ряд ограничений. Существующее требование нормального распределения данных снижает точность распознавания аномалий. **Метод.** В представленной работе задача повышения точности и полноты выявления ранее не встречавшихся на изображениях аномалий решается путем комбинирования моделей Contrastive Language-Image Pretraining (CLIP) и доменно-ориентированного трансформера BERT Pre-Training of Image Transformers (BeiT). Модели CLIP и BeiT позволяют одновременно решать задачи бинарной сегментации и классификации аномалий. Более точное выявление аномалий достигается использованием взвешенных эмбеддингов от каждого модуля. Одновременно автоматизируется генерация текстовых представлений на основе Large Language Model, что существенно улучшает обобщающую способность модели. **Основные результаты.** Оценка эффективности разработанных моделей выполнена на тестовой выборке Benchmarks for Medical Anomaly Detection). Для домена кожных новообразований тестовая выборка сформирована из датасетов ISIC-18, ISIC-19, SD-198 и 7-point criteria database. Разработанный метод продемонстрировал в среднем увеличение метрики ROC AUC (при классификации, на уровне image-level) на 10,95 %, а метрики ROC AUC (при сегментации, на уровне pixel-level) — на 0,66 % по сравнению с известными решениями. **Обсуждение.** Проведенные эксперименты показали высокую эффективность предложенного подхода на задачах классификации и сегментации аномалий, метод продемонстрировал превосходящие результаты по средним значениям метрик. Анализ инференса показал, что использование вариационного автоэнкодера в составе CLIP+BeiT для генерации центроидов способствует более стабильной работе модели в few-shot подходе. Практическая значимость предложенного метода заключается в его адаптивности и устойчивости к изменяющимся распределениям данных, что делает его перспективным решением для автоматизированного анализа аномалий в медицинской диагностике, промышленном контроле и других областях, где может наблюдаться высокая неопределенность данных.

Ключевые слова

детекция аномалий, zero-shot, few-shot, трансформер, мультизадачное обучение

Благодарности

Работа поддержана Минобрнауки Российской Федерации, тема № FFZM-2025-0005.

Ссылка для цитирования: Милантьев С.А., Михайлова П.Д., Бессмертный И.А. Выявление аномалий в условиях ограниченности и неопределенности данных с использованием zero-shot и few-shot подходов // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 4. С. 684–693. doi: 10.17586/2226-1494-2025-25-4-684-693

Anomaly detection under data scarcity and uncertainty using zero-shot and few-shot approaches

Sergey A. Milantev[1]✉, Polina D. Mikhailova[2], Igor A. Bessmertny[3]

[1,3] ITMO University, Saint Petersburg, 197101, Russian Federation

[2] Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation

[1] geerkus@gmail.com✉, https://orcid.org/0000-0002-1970-5217

[2] polina.delitzsch@gmail.com, https://orcid.org/0009-0005-9731-0105

[3] bessmertny@itmo.ru, https://orcid.org/0000-0001-6711-6399

Abstract

Anomaly detection under conditions of limited data volume represents a pressing challenge across numerous applied domains, including medical diagnostics. Machine learning methods typically rely on the availability of annotated anomalous samples for training, which is often impractical. Existing anomaly detection techniques designed for few-shot or zero-shot scenarios suffer from various limitations. In particular, the common assumption of normally distributed data reduces the accuracy of anomaly classification. In this study, the task of improving the accuracy and completeness of anomaly detection in previously unseen images by leveraging a combination of the Contrastive Language-Image Pretraining (CLIP) and the domain-specific transformer BERT Pre-Training of Image Transformers (BeiT) models. The integration of CLIP and BeiT models enables simultaneous binary segmentation and anomaly classification. Enhanced anomaly detection is achieved through the use of weighted embeddings from each module. Additionally, the automated generation of textual representations based on a Large Language Model significantly enhances the generalization capacity of the system. The performance of the proposed models was evaluated on the Benchmarks for Medical Anomaly Detection test set. For the dermatological domain, a test set was constructed from ISIC-18, ISIC-19, SD-198, and 7-point criteria database. The proposed method demonstrated an average improvement in the ROC-AUC metric by 10.95 % at the image-level and by 0.66 % at the pixel-level compared to existing state-of-the-art solutions. Experimental results confirm the high effectiveness of the proposed approach in anomaly classification and segmentation tasks, showing superior average metric values. Inference analysis revealed that the incorporation of a variational autoencoder within the CLIP+BeiT architecture for centroid generation enhances the model stability in few-shot scenarios. The practical significance of the proposed method lies in its adaptability and robustness to changing data distributions, making it a promising solution for automated anomaly analysis in medical diagnostics, industrial monitoring, and other domains characterized by high data uncertainty.

Keywords

anomaly detection, zero-shot, few-shot, transformer, multi-task learning

Acknowledgements

The work was supported by the Ministry of Education and Science of the Russian Federation, topic No. FZZM-2025-0005.

For citation: Milantev S.A., Mikhailova P.D., Bessmertny I.A. Anomaly detection under data scarcity and uncertainty using zero-shot and few-shot approaches. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 4, pp. 684–693 (in Russian). doi: 10.17586/2226-1494-2025-25-4-684-693

Введение

Разработка универсального метода для выявления аномалий и нетипичных закономерностей представляет собой ключевое направление развития в различных предметных областях. Под универсальностью метода понимается способность архитектурного решения, лежащего в его основе, обеспечивать обобщение на ранее не наблюдавшиеся поддоменные области, что отражает общий уровень его генерализационной эффективности. В медицине подобные методы позволяют анализировать мультимодальные данные в контексте различных поддоменов, включая те, с которыми модель ранее не сталкивалась (zero-shot подход). Данный подход способствует расширению области применения одной модели, ускоряет процессы принятия решений и минимизирует влияние человеческого фактора.

В настоящей работе выполнен анализ современных научных разработок в области мультидоменного обнаружения аномалий, протестированных на датасетах BMAD: Benchmarks for Medical Anomaly Detection [1]. Рассмотрены следующие state-of-the-art подходы, продемонстрировавшие высокие показатели метрик на бенчмарке BMAD:

— UTRAD [2] анализирует эмбеддинги автоэнкодером на основе трансформерной архитектуры, предполагая, что реконструкция содержит более информативные признаки, чем исходное изображение;
— MKD [3] и RD4AD [4] используют дистилляцию знаний для оценки различий между моделями «ученик-учитель», предобученными на ImageNet. При этом RD4AD модифицирует классическую схему, применяя архитектуру «энкодер-учитель — декодер-ученик»;
— PatchCore [5], CFA [6] и CFLOW [7] формируют few-shot эмбеддинги для нормальных изображений и анализируют их косинусное расстояние, предполагая нормальность распределения этих данных. CFLOW использует двухпотоковый анализ аномалий на глобальном и локальном уровнях;
— VAND-APRIL-GAN [8] является усовершенствованной версией WinCLIP, комбинируя zero-shot и few-shot подходы. В данной модели CLIP разбивается на четыре этапа, после чего эмбеддинги сравниваются по косинусному сходству, формируя две независимые оценки: zero-shot score и few-shot score.

Несмотря на высокие значения метрик на сложных датасетах, большинство рассмотренных методов

(за исключением VAND-APRIL-GAN) предполагают нормальное распределение данных, что ограничивает их универсальность и снижает эффективность при анализе аномалий с семантической точки зрения. В то же время VAND-APRIL-GAN полностью полагается на Contrastive Language-Image Pretraining (CLIP), который изначально обучался на общих задачах, что может снижать его адаптивность к специализированным доменам.

Современные методы либо предполагают априорную нормальность распределения, либо плохо масштабируются на поддомены с выраженной семантической неоднородностью. Настоящее исследование направлено на устранение этого противоречия путем объединения zero-shot и few-shot методов с механизмом доменной адаптации и мультизадачного обучения. Таким образом, формулируется новая задача: разработка универсальной архитектуры, устойчивой к междоменным сдвигам при отсутствии аномальных примеров в обучении.

В работе предложена модель, сочетающая zero-shot и few-shot подходы в условиях ограниченности и неопределенности данных. Для повышения ее адаптивности CLIP корректируется с использованием доменно-ориентированной трансформерной модели. В рамках исследования распределения оценивались через эмпирическое распределение признаков в скрытых слоях моделей CLIP и BERT Pre-Training of Image Transformers (BeiT), а также с использованием расстояний между эмбеддингами в косинусном пространстве. Модель оптимизируется сразу на двух задачах: выявление областей интереса (бинарных масок) и классификация аномалий. Итоговый результат формируется на основе взвешенных эмбеддингов от каждого модуля. Дополнительно усовершенствована система текстового представления данных: вместо ручного формирования текстовых представлений применяется автоматизированный подход с использованием Large Language Model (LLM), что существенно повышает обобщающую способность модели.

Под предлагаемым методом понимается архитектурное решение, сочетающее мультимодальные модели с доменной адаптацией. Новизна заключается в интеграции CLIP и BeiT по скрытым состояниям с регулируемыми весами, что ранее не применялось в задачах мультизадачной детекции аномалий в условиях отсутствия аномальных примеров. Также впервые используется механизм сравнения выходов модели с эталонами, сформированными через Variational Autoencoder (VAE), на этапе инференса в качестве few-shot примеров.

Методы и подходы

Предобработка данных. В качестве предметной области для разработки универсальной модели выявления аномалий был выбран медицинский домен. В текущей работе под неопределенностью и ограниченностью данных понимается отсутствие образцов аномальных объектов в тренировочной выборке или априорных знаний о распределении аномалий. Таким образом, модели обучаются семантически понимать нормальные образцы, что позволяет во время валидации и тестов выявлять аномалии. В проведенных экспериментах были использованы наборы данных, представленные на рис. 1.

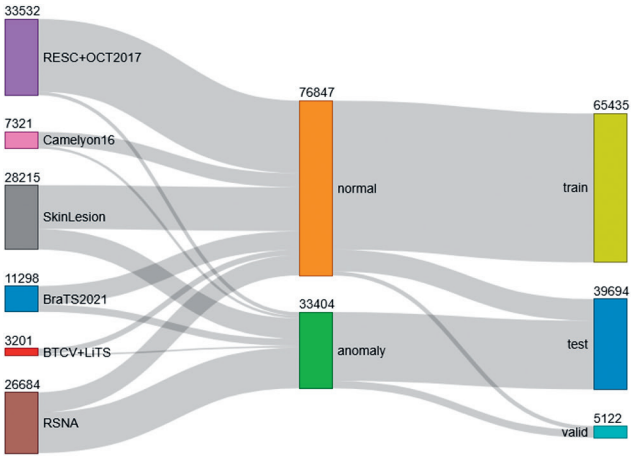

Рис. 1. Используемые наборы данных.

Цифровые обозначения — количество образцов в наборе данных, где normal — изображения нормального класса; anomaly — изображения класса аномалии; train — обучающая выборка; test — тестовая выборка; valid — валидационная выборка; OCT[1] — Optical Coherence Tomography Dataset; RESC [9] — Retinal Edema Segmentation Challenge Dataset; RSNA [10] — Radiological Society of North America Pneumonia Detection Dataset

Fig. 1. The datasets used in the study. Numerical labels indicate the number of samples.

Digits — number of samples in the data set; normal — normal class images; anomaly — anomaly class images; train — testing sample; test — test sample; valid — validation sample; OCT[1] — Optical Coherence Tomography Dataset; RESC [9] — Retinal Edema Segmentation Challenge Dataset; RSNA [10] — Radiological Society of North America Pneumonia Detection Dataset

[1] Kermany D. Zhang K., Goldbaum M. Labeled Optical Coherence Tomography (OCT) for Classification, Mendeley Data v2, doi: 10.17632/rscbjbr9sj.

Перед использованием в модели все изображения проходили предварительную проверку на дубликаты с использованием MD5-хэшей и перцептивных хэшей. Далее выполнялся этап предобработки, включающий операции CenterCrop и Resize (до 336 × 336 пикселов для $CLIP_{vis}$ и 384 × 384 пикселов для BeiT), а также нормализацию данных.

Текстовые представления для каждого объекта формировались автоматически с помощью модели LLAMA3 в соответствии с заранее заданным промптом.

Архитектурное решение. Для решения задачи обнаружения аномалий в условиях ограниченности и неопределенности данных предлагается использовать комбинацию zero-shot подхода с добавлением промежуточных состояний из трансформера, специализированного на конкретном поддомене. Применена интеграция моделей CLIP и BeiT, объединенных посредством линейных слоев по каждому четвертому скрытому состоянию. Этот подход позволяет обогащать внутренние эмбеддинги CLIP за счет адаптивного смешивания с доменно-ориентированными представлениями BeiT, что особенно важно, поскольку анализ аномалий выходит за рамки исходных задач CLIP.

На рис. 2 представлена архитектура разработанной модели.

Zero-shot подход основан на предварительно обученной мультимодальной модели CLIP (*openai/clip-vit-large-patch14-336*). В предложенной схеме CLIP разделен на две компоненты:

— $CLIP_{vis}$ — модуль для обработки входных изображений;

— $CLIP_{text}$ — модуль для обработки текстовых представлений.

Основное внимание в архитектуре уделяется обработке визуальных признаков. Каждое входное изображение проходит через модуль Processor, где выполняется нормализация и приведение к требуемому размеру. Для $CLIP_{vis}$ изображение масштабируется до 336 × 336 пикселов, после чего разбивается на фрагменты 14 × 14. В свою очередь, для BeiT (*microsoft/beit-large-patch16-384*) входное изображение трансформируется в размер 384 × 384 пикселов и разбивается на фрагменты 16 × 16.

Объединение представлений осуществляется следующим образом: к эмбеддингам encoder-слоев $CLIP_{vis}$ на уровнях {4, 8, 12, 16, 20, 24} добавляются соответствующие скрытые состояния BeiT. Для каждого слоя MLP_i применяется поэлементная сумма с весовыми коэффициентами, регулирующими вклад каждой модели:

$$MLP_{inputs\_i} = \beta_i C_i + \gamma_i B_i,$$

где β_i и γ_i — весовые коэффициенты эмбеддингов $CLIP_{vis}$ и BeiT; C_i и B_i — эмбеддинги $CLIP_{vis}$ и BeiT. В выполненном эксперименте приняты значения $\beta_i = 0{,}4$ и $\gamma_i = 0{,}6$ для всех слоев. Значения параметров подбирались эмпирически для разных доменов с помощью метода Grad-CAM, использующийся для визуализации и интерпретируемости предсказания слоев, и показателей метрик в процессе обучения. Было установлено, что увеличение γ_i (вклад BeiT) позволяет повысить чувствительность модели к редким структурам внутри домена.

Поэлементная сумма эмбеддингов размерности [577, 1024] передается в residual-соединение (MLP_i), что позволяет скорректировать внимание модели CLIP на задаче выявления аномалий. Подробная архитектура Multi-Layer Perceptron (MLP) модуля представлена на рис. 3.

Объединенные эмбеддинги C_i из $encoder_i$ и B_i из $hidden\_state_i$ параллельно подаются в линейные слои, выполняющие задачи классификации и сегментации. Конструктивно эти слои идентичны, но оптимизированы под различные задачи. Архитектурно они состоят из двух последовательных линейных преобразований с размерностями [$in = 1024$, $out = 768$] и [$in = 768$, $out = 1024$].

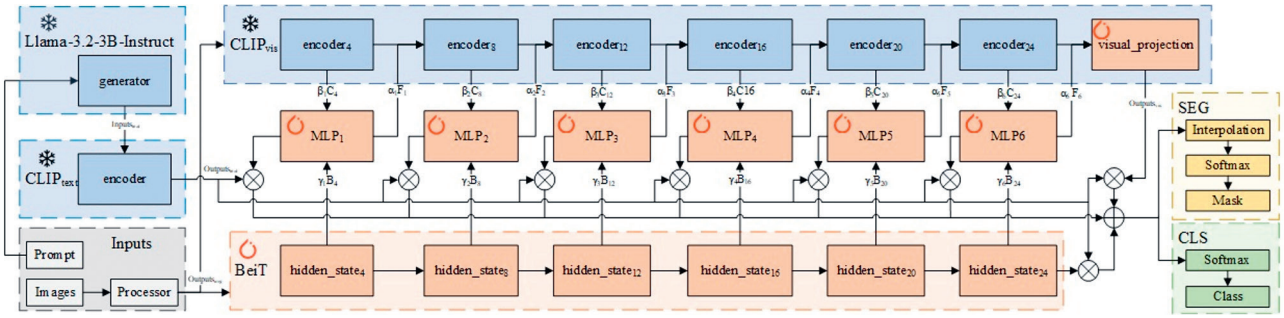

Рис. 2. Архитектура модели CLIP + BeiT.

Оранжевый цвет — все обучаемые модули; синий — замороженные параметры; серый — входящие данные; желтый — модуль сегментации; зеленый — модуль классификации; β_i и γ_i — весовые коэффициенты эмбеддингов $CLIP_{vis}$ и BeiT; C_i и B_i — эмбеддинги $CLIP_{vis}$ и BeiT; $inputs_{text}$ — сгенерированный текст; $outputs_{text}$ — текстовое представление; $outputs_{vis}$ — визуальное представление; SEG — модуль результирующего сегментатора; CLS — модуль результирующего классификатора

Fig. 2. CLIP + BeiT model architecture.

Colors: orange — all training modules; blue — freezed parameters; gray — input data; yellow — segmentation module; green — classification module; β_i and γ_i — embedding weighting factors of $CLIP_{vis}$ and BeiT; C_i и B_i — embeddings of $CLIP_{vis}$ and BeiT; $inputs_{text}$ — generated text; $outputs_{text}$ — text representation; $outputs_{vis}$ — visual representation; SEG — resulting segmenter module; CLS — resulting classificator module

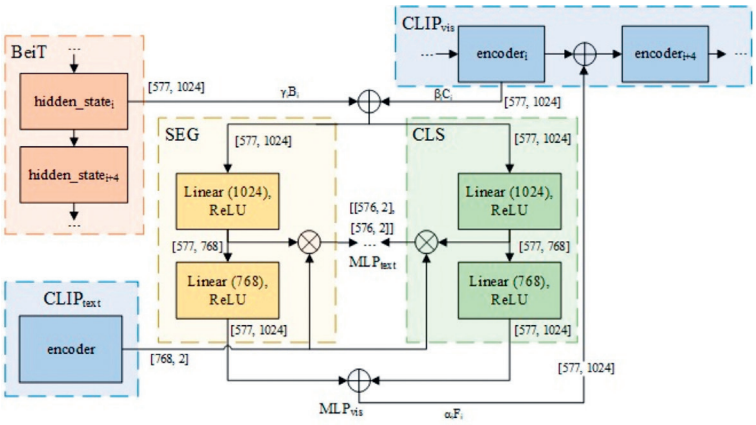

Рис. 3. Детальный вид модуля MLP.

Оранжевый цвет — модуль BeIT; синий — модуль CLIP; желтый — модуль сегментации; зеленый — модуль классификации

Fig. 3. Detailed view of the MLP module.

Colors: orange — BeIT module; blue — CLIP module; yellow — segmentation module; green — classification module

После первого линейного слоя эмбеддинги размерности [577, 768] умножаются на текстовые представления из $CLIP_{text}$ размерности [768, 2]. Полученные произведения обозначаются как MLP_{text_i} и передаются для вычисления итогового эмбеддинга по всей модели. Параллельно эмбеддинги первого линейного слоя поступают во второй линейный слой, формируя на выходе представление размерности [577, 1024].

Далее эмбеддинги, полученные с классификационного и сегментационного слоев, поэлементно суммируются. Итоговый результат, взвешенный коэффициентом α_i, вычисляется по формуле:

$$MLP_{vis\_i} = \alpha_i F_i + (1 - \alpha_i) C_i,$$

где α_i управляет вкладом дополнительных эмбеддингов. В данном эксперименте установлено значение $\alpha_i = 0{,}5$ для всех слоев ($\alpha_1 = \ldots = \alpha_6 = 0{,}5$), что обеспечивает эквивалентное влияние различных модулей.

Таким образом, MLP_i формирует три выходных эмбеддинга:

— результат классификационного уровня;
— результат сегментационного уровня;
— произведение эмбеддинга с текстовым представлением.

Одновременное решение задач бинарной классификации и сегментации в рамках единого MLP-модуля реализует парадигму мультизадачного обучения, позволяя модели учитывать пространственную структуру и семантическую информацию при формировании предсказаний.

Экспериментально подтверждено, что введение классификационного и сегментационного слоев способствует улучшению обобщающей способности модели, ускоряя ее сходимость и значительно повышая метрики на задачах классификации и сегментации.

На заключительном этапе MLP_6, а также выходные эмбеддинги $CLIP_{vis}(encoder_{24})$ и $BeiT(hidden\_state_{24})$ передаются в слой *visual_projection* (*in = 1024*, *out = 768*) из $CLIP_{vis}$, веса которого подлежат обучению.

Нормализованный выход данного слоя $outputs_{vis}$ ([577, 768]) умножается на текстовые эмбеддинги $CLIP_{text}$ ([768, 2]). Аналогично, нормализованный выход $outputs_{BeiT}$ из модуля BeiT также перемножается с текстовыми эмбеддингами $CLIP_{text}$.

Финальный эмбеддинг формируется как поэлементная сумма:

$$outputs_{final} = outputs_{vis} + outputs_{BeiT} + MLP_{text\_i}.$$

Итоговое представление *outputs_final* размерности [577, 768] передается в сегментационный и классификационный модули для постобработки. В классификационном модуле применяется softmax(dim = –1), а итоговый класс вычисляется усреднением значений по всем патчам с последующей нормализацией в диапазон $result_{df} \in [0, 1]$. В сегментационном модуле для $result_{seg}$ выполняются аналогичные операции, но с дополнительной интерполяцией, позволяющей сформировать бинарную маску.

Для формирования текстовых входных данных используется модель *LLAMA-3.2-3B-Instruct*. В рамках заданного промпта для каждого класса LLM генерирует по 10 текстовых шаблонов, описывающих нормальные и аномальные изображения.

Обучение и оптимизация модели. Проведены эксперименты по интеграции промежуточных состояний модели BeiT в zero-shot модель CLIP. Модель BeiT обучалась на тех же тренировочных данных, что и основная модель, но решала задачу бинарной классификации в пределах конкретного домена. В целях обеспечения модульности и итеративного подхода обучение BeiT осуществлялось отдельно от основного пайплайна.

Для оптимизации BeiT использовался *AdamW* с параметром *learning_rate* = $5 \cdot 10^{-5}$. В качестве функции потерь применялась *FocalLoss*, а для обеспечения стабильной начальной сходимости использовался *warmup scheduler*, где *warmup_steps* составлял 5 % от общего числа шагов обучения.

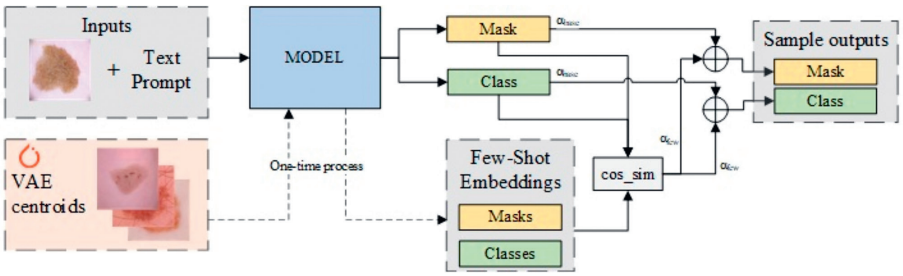

Рис. 4. Механизм инференса модели.

Розовый цвет — модуль VAE; синий — модуль модели CLIP + BeiT; серый — необучаемые модули (модуль предобработки входных данных, база данных для few-shot примеров и выходной модуль); желтый — модули сегментационной части модели; зеленый — модули классификационной части модели; α_{base} и α_{few} — весовые коэффициенты эмбеддингов модели CLIP + BeiT и few-shot эмбеддингов

Fig. 4. Model inference mechanism.

Colors: pink — VAE module; blue — module of CLIP + BeiT model; gray — untrainable modules (input data preprocessing module, database for few-shot examples and output module); yellow — modules of the segmentation part of the model; green — modules of the classification part of the model; α_{base} and α_{few} — model embedding weighting coefficients of CLIP + BeiT and few-shot embeddings

Для обучения MLP-модуля использовалась комбинированная функция потерь:

$$Loss_{MLP} = Loss_{clf} + Loss_{seg} = BCEWithLogitsLoss + \\ + DiceLoss + FocalLoss,$$

где *BCEWithLogitsLoss* отвечает за классификацию, *DiceLoss* и *FocalLoss* — за сегментацию. Оптимизация MLP-модуля выполнялась совместно с последним слоем $CLIP_{vis}$, что позволяло корректировать веса модели для решения обеих задач в рамках одной итерации обучения.

Few-shot подход. Во время инференса модели применяется few-shot подход, в котором полученные маска и класс входного сэмпла сравниваются с предвычисленными few-shot центроидами. Для каждого класса (0 — normal, 1 — anomaly) было сформировано по 16 центроидов, и сравнение проводилось с использованием коэффициента Отиаи.

Для генерации центроидов по каждому классу обучен VAE с использованием оптимизатора *Adam*(*learning_rate* = 10^{-4}) и функции потерь:

$$Loss_{VAE} = KLD_{loss} + Recon_{loss}.$$

Дополнительно применен механизм *clipping* для предотвращения взрыва градиентов. В результате обучения VAE были получены по 16 центроидов для каждого домена.

На рис. 4 представлена схема инференса модели. VAE-центроиды предобрабатываются и загружаются в память (или хранятся в сериализованном формате) единожды перед запуском инференса модели.

Во время инференса косинусное сходство между входным сэмплом и few-shot центроидами суммируется по всем патчам с последующей нормализацией. Итоговое значение представляет собой матрицу коэффициентов сходства, которая используется при агрегировании с маской и классом текущего сэмпла. Взвешивание результата осуществляется с коэффициентом $\alpha_{few} = 0{,}15$.

Результаты

Рассмотрим основные результаты, полученные в ходе экспериментов на бенчмарке BMAD [1]. Оценка эффективности разработанных моделей проводилась на тестовой выборке BMAD, а для домена кожных новообразований тестовая выборка была сформирована из датасетов ISIC-18, ISIC-19, SD-198 и 7-point criteria database [12–16].

В процессе обучения модели оценивались по метрикам F1-score, Precision, Recall, ROC AUC, Dice и IoU при различных значениях *threshold*. Для корректного сравнения с альтернативными подходами, представленными в бенчмарке BMAD, основными метриками были выбраны ROC AUC на уровне изображения (image-level) и ROC AUC на уровне пикселов (pixel-level).

В табл. 1 приведены полученные значения метрик в результате проведенного эксперимента. В столбце ISIC+ представлены результаты на тестовой подвыборке, включающей данные из SD-198, 7-point criteria database и ISIC-19.

Разработанный метод демонстрирует наилучшие результаты среди представленных моделей (в ISIC+ модель уступила предыдущей нашей модели по метрике F1-score). Метод *vanilla* CLIP + LLM демонстрирует значительно более низкие метрики по сравнению с методами, использующие модули MLP + BeiT. Особенно заметное снижение качества наблюдается на датасете ISIC+, что указывает на слабую способность модели CLIP к генерализации в условиях более сильного доменного сдвига.

В табл. 2 приведены полученные значения метрик в рамках проведенного эксперимента.

Разработанный метод показывает одни из лучших значений метрик и во всех доменах входит в топ-3 лучших показателей, при этом в среднем обеспечил прирост метрики ROC AUC (image-level) на 10,95 % и ROC AUC (pixel-level) на 0,66 % по сравнению с ранее предложенными методами. Таким образом, предложенный метод демонстрирует наивысшую эффективность как в задаче классификации, так и в задаче сегмента-

Таблица 1. Показатели метрик на классификационной задаче
Table 1. Metrics scores on classification task

Методы	Модели							
	OCT2017		RSNA [10]		Camelyon16 [11]		ISIC + [12–16]	
	Метрики							
	F1-score	ROC AUC	F1-score	ROC AUC	F1-score	ROC AUC	F1-score	ROC AUC
UTRAD [2]	—	0,9678 ± 0,0056	—	0,7564 ± 0,0124	—	0,6996 ± 0,0464	—	—
MKD [3]	—	0,9674 ± 0,0026	—	0,8201 ± 0,0012	—	0,7754 ± 0,0027	—	—
RD4AD [4]	—	0,9730 ± 0,0079	—	0,6763 ± 0,0111	—	0,6681 ± 0,0071	—	—
PatchCore [5]	—	0,9857 ± 0,0003	—	0,7614 ± 0,0067	—	0,6934 ± 0,0021	—	—
CFA [6]	—	0,7947 ± 0,0056	—	0,6683 ± 0,0023	—	0,6564 ± 0,001	—	—
CFLOW [7]	—	0,8535 ± 0,0211	—	0,7153 ± 0,0149	—	0,5566 ± 0,0197	—	—
VAND-APRIL-GAN [8]	—	0,9941	—	0,7743	—	0,7611	—	—
Ours (SkinNetV2)	—	—	—	—	—	—	0,7903	0,8633
Zero-shot method (vanilla CLIP + LLM)	0,5609	0,9712	0,4002	0,8120	0,6115	0,7474	0,6085	0,7427
BeiT	0,9419	0,9937	0,7627	0,8284	0,6347	0,7538	0,7334	0,8291
$(\text{BeiT} + \text{CLIP} + \text{LLM})_4$	0,8320	0,9829	0,5089	0,8007	0,6050	0,7090	0,6936	0,8151
$(\text{BeiT} + \text{CLIP} + \text{LLM})_8$	0,9367	0,9821	0,7885	0,8148	0,5423	0,7047	0,6058	0,7859
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{12}$	0,9416	0,9806	0,7696	0,8135	0,6304	0,7328	0,7362	0,834
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{16}$	0,8431	0,9719	0,5674	0,8127	0,6330	0,7419	0,7194	0,7495
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{20}$	0,9323	0,9839	0,8126	0,7959	0,5715	0,7602	0,7594	0,8566
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{24}$	0,9485	0,9908	0,8036	0,8171	0,6278	0,7156	0,7271	0,8502
BeiT + Zero/Few-shot	0,9648	0,9948	0,8510	0,8377	0,6943	0,7789	0,7829	0,8529

Примечание. Серым цветом выделены лучшие показатели на соответствующих задачах.

Таблица 2. Показатели метрик на сегментационной задаче

Table 2. Metrics scores on segmentation task

Методы	BraTS2021 [17]		BTCV[1] + LiTs [18]		RESC [9]		ISIC + [12–16]	
	roc_auc	*roc_auc*	*roc_auc*	*roc_auc*	*roc_auc*	*roc_auc*	*roc_auc*	*roc_auc*
	image-level	pixel-level	image-level	pixel-level	image-level	pixel-level	image-level	pixel-level
UTRAD [2]	0,8292 ± 0,0232	0,9261 ± 0,0067	0,5581 ± 0,0566	0,8788 ± 0,0132	0,8939 ± 0,0192	0,9454 ± 0,0124	—	—
MKD [3]	0,8147 ± 0,0036	0,8944 ± 0,0024	0,6072 ± 0,0119	0,9606 ± 0,0027	0,8900 ± 0,0025	0,8674 ± 0,0065	—	—
RD4AD [4]	0,8945 ± 0,0091	0,9645 ± 0,0017	0,6038 ± 0,0117	0,9601 ± 0,019	0,8777 ± 0,0087	0,9618 ± 0,0015	—	—
PatchCore [5]	0,9165 ± 0,0036	0,9697 ± 0,0004	0,6028 ± 0,0076	0,9643 ± 0,0019	0,9155 ± 0,001	0,9648 ± 0,001	—	—
CFA [6]	0,8438 ± 0,0087	0,9633 ± 0,0014	0,6200 ± 0,0108	0,9724 ± 0,0014	0,6990 ± 0,0026	0,9110 ± 0,0087	—	—
CFLOW [7]	0,7482 ± 0,0532	0,9376 ± 0,0067	0,5080 ± 0,0447	0,9241 ± 0,0116	0,7495 ± 0,0581	0,9378 ± 0,0057	—	—
VAND-APRIL-GAN [8]	0,8918	0,9467	0,5305	0,9624	0,9470	0,9798		
Ours (SkinNetV2)	—	—	—	—	—	—	0,8633	0,84311
Zero-shot method (vanilla CLIP + LLM)	0,6375	0,9024	0,5385	0,9498	0,7595	0,9387	0,7427	0,8137
BeiT	0,9071	0,8939	0,8283	0,9710	0,8836	0,9703	0,8291	0,9280
$(\text{BeiT} + \text{CLIP} + \text{LLM})_4$	0,9143	0,9692	0,6846	0,9215	0,9163	0,9660	0,8151	0,9096
$(\text{BeiT} + \text{CLIP} + \text{LLM})_8$	0,7238	0,9392	0,5385	0,9498	0,8877	0,9640	0,7859	0,8837
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{12}$	0,8995	0,9690	0,8710	0,9535	0,9133	0,9703	0,8340	0,9243
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{16}$	0,7054	0,9316	0,5532	0,9440	0,8826	0,9617	0,7495	0,8743
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{20}$	0,9334	0,9755	0,7334	0,9676	0,9089	0,9746	0,8566	0,9457
$(\text{BeiT} + \text{CLIP} + \text{LLM})_{24}$	0,8982	0,9720	0,7227	0,9701	0,9045	0,9756	0,8502	0,9372
BeiT + Zero/Few-shot	0,9477	0,9817	0,9083	0,9723	0,9195	0,9745	0,8529	0,9481

Примечание. Серым цветом выделены лучшие показатели на соответствующих задачах.

[1] Landman B., et al. MICCAI Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge. 2015.

ции медицинских изображений, превосходя методы из бенчмарка BMAD.

Заключение

В работе представлен универсальный метод выявления аномалий, основанный на zero-shot/few-shot подходах, что позволяет эффективно работать в условиях ограниченности и неопределенности данных в различных доменах.

Предложенный метод интегрирует модель CLIP, которая адаптируется с помощью скрытых состояний доменно-ориентированного трансформера BeiT. При этом модель одновременно решает две задачи: сегментацию (RoI → выделение бинарной маски) и бинарную классификацию аномалий, используя взвешенные эмбеддинги от каждого модуля для формирования окончательного результата.

Дополнительно оптимизирован процесс генерации текстовых представлений, что позволяет автоматизировать их создание на основе LLM. Это значительно повышает обобщающую способность модели, превосходя подход с ручной формулировкой текстовых описаний, особенно на специфичных доменах кожных новообразований.

Разработанный метод продемонстрировал высокие результаты на различных наборах данных, включая те, где наблюдаются значительные различия в распределении классов. Он показал лучшие средние значения метрик по сравнению с существующими методами из BMAD и традиционным zero-shot подходом.

Таким образом, дальнейшее развитие zero-shot и few-shot методологий представляется перспективным направлением для автоматизированного анализа аномалий, обеспечивая высокую точность, адаптивность и устойчивость даже в условиях ограниченного количества данных и высокой вариативности распределений.

Литература

1. Bao J., Sun H., Deng H., He Y., Zhang Z., Li X. BMAD: Benchmarks for Medical Anomaly Detection // arXiv. 2023. arXiv:2306.11876. https://doi.org/10.48550/arXiv.2306.11876
2. Chen L., You Z., Zhang N., Xi J., Le X. UTRAD: Anomaly detection and localization with U-Transformer // Neural Networks. 2022. V. 147. P. 53–62. https://doi.org/10.1016/j.neunet.2021.12.008
3. Salehi M., Sadjadi N., Baselizadeh S., Rohban M.H., Rabiee H.R. Multiresolution knowledge distillation for anomaly detection // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021. P. 14897–14907. https://doi.org/10.1109/CVPR46437.2021.01466
4. Deng H., Li X. Anomaly detection via reverse distillation from one-class embedding // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 9727–9736. https://doi.org/10.1109/CVPR52688.2022.00951
5. Roth K., Pemula L., Zepeda J., Schölkopf B., Brox T., Gehler P. Towards total recall in industrial anomaly detection // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 14298–14308. https://doi.org/10.1109/CVPR52688.2022.01392
6. Lee S., Lee S., Song B. CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization // IEEE Access. 2022. V. 10. P. 78446–78454. https://doi.org/10.1109/ACCESS.2022.3193699
7. Gudovskiy D., Ishizaka S., Kozuka K. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022. P. 1819–1828. https://doi.org/10.1109/WACV51458.2022.00188
8. Chen X., Han Y., Zhang J. APRIL-GAN: a Zero-/Few-shot anomaly classification and segmentation method // arXiv. 2023. arXiv:2305.17382v3. https://doi.org/10.48550/arXiv.2305.17382
9. Hu J., Chen Y., Yi Z. Automated segmentation of macular edema in OCT using deep neural networks // Medical Image Analysis. 2019. V. 55. P. 216–227. https://doi.org/10.1016/j.media.2019.05.002
10. Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 3462–3471. https://doi.org/10.1109/CVPR.2017.369
11. Bejnordi B., Veta M., van Diest P.J., van Ginneken B., Karssemeijer N., Litjens G., van der Laak J.A.W.M. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer // JAMA Journal of the American Medical Association. 2017. V. 318. N 22. P. 2199–2210. https://doi.org/10.1001/jama.2017.14585

References

1. Bao J., Sun H., Deng H., He Y., Zhang Z., Li X. BMAD: Benchmarks for Medical Anomaly Detection. *arXiv*, 2023, arXiv:2306.11876. https://doi.org/10.48550/arXiv.2306.11876
2. Chen L., You Z., Zhang N., Xi J., Le X. UTRAD: Anomaly detection and localization with U-Transformer. *Neural Networks*, 2022, vol. 147, pp. 53–62. https://doi.org/10.1016/j.neunet.2021.12.008
3. Salehi M., Sadjadi N., Baselizadeh S., Rohban M.H., Rabiee H.R. Multiresolution knowledge distillation for anomaly detection. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14897–14907. https://doi.org/10.1109/CVPR46437.2021.01466
4. Deng H., Li X. Anomaly detection via reverse distillation from one-class embedding. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9727–9736. https://doi.org/10.1109/CVPR52688.2022.00951
5. Roth K., Pemula L., Zepeda J., Schölkopf B., Brox T., Gehler P. Towards total recall in industrial anomaly detection. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14298–14308. https://doi.org/10.1109/CVPR52688.2022.01392
6. Lee S., Lee S., Song B. CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 2022, vol. 10, pp. 78446–78454. https://doi.org/10.1109/ACCESS.2022.3193699
7. Gudovskiy D., Ishizaka S., Kozuka K. CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1819–1828. https://doi.org/10.1109/WACV51458.2022.00188
8. Chen X., Han Y., Zhang J. APRIL-GAN: a Zero-/Few-shot anomaly classification and segmentation method. *arXiv*, 2023, arXiv:2305.17382v3. https://doi.org/10.48550/arXiv.2305.17382
9. Hu J., Chen Y., Yi Z. Automated segmentation of macular edema in OCT using deep neural networks. *Medical Image Analysis*, 2019, vol. 55, pp. 216–227. https://doi.org/10.1016/j.media.2019.05.002
10. Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471. https://doi.org/10.1109/CVPR.2017.369
11. Bejnordi B., Veta M., van Diest P.J., van Ginneken B., Karssemeijer N., Litjens G., van der Laak J.A.W.M. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA Journal of the American Medical Association*, 2017, vol. 318, no. 22, pp. 2199–2210. https://doi.org/10.1001/jama.2017.14585

12. Tschandl P., Rosendahl C., Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions // Scientific Data. 2018. V. 5. P. 180161. https://doi.org/10.1038/sdata.2018.161
13. Codella N.C.F., Gutman D., Celebi M.E., Helba B., Marchetti M.A., Dusza S.W., Kalloo A., Liopyris K., Mishra N., Kittler H., Halpern A. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC) // Proc. of the IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). 2018. P. 168–172. https://doi.org/10.1109/ISBI.2018.8363547
14. Combalia M., Codella N.C.F., Rotemberg V., Helba B., Vilaplana V., Reiter O., Carrera C., Barreiro A., Halpern A.C., Puig S., Malvehy J. BCN20000: Dermoscopic lesions in the wild // arXiv. 2019. arXiv:1908.02288. https://doi.org/10.48550/arXiv.1908.02288
15. Sun X., Yang J., Sun M., Wang K. A benchmark for automatic visual classification of clinical skin disease images // Lecture Notes in Computer Science. 2016. V. 9910. P. 206–222. https://doi.org/10.1007/978-3-319-46466-4_13
16. Kawahara J., Daneshvar S., Argenziano G., Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets // IEEE Journal of Biomedical and Health Informatics. 2019. V. 23. N 2. P.538–546. https://doi.org/10.1109/JBHI.2018.2824327
17. Baid U., Ghodasara S., Mohan S., Bilello M., Calabrese E., Colak E., et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification // arXiv. 2021. arXiv:2107.02314. https://doi.org/10.48550/arXiv.2107.02314
18. Bilic P, Christ P., Li H.B., Vorontsov E., Ben-Cohen A., Kaissis G., et al. The Liver Tumor Segmentation benchmark (LiTS) // arXiv. 2019. arXiv:190.04056. https://doi.org/10.48550/arXiv.1901.04056

12. Tschandl P., Rosendahl C., Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018, vol. 5, pp. 180161. https://doi.org/10.1038/sdata.2018.161
13. Codella N.C.F., Gutman D., Celebi M.E., Helba B., Marchetti M.A., Dusza S.W., Kalloo A., Liopyris K., Mishra N., Kittler H., Halpern A. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). *Proc. of the IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 168–172. https://doi.org/10.1109/ISBI.2018.8363547
14. Combalia M., Codella N.C.F., Rotemberg V., Helba B., Vilaplana V., Reiter O., Carrera C., Barreiro A., Halpern A.C., Puig S., Malvehy J. BCN20000: Dermoscopic lesions in the wild. *arXiv*, 2019, arXiv:1908.02288. https://doi.org/10.48550/arXiv.1908.02288
15. Sun X., Yang J., Sun M., Wang K. A benchmark for automatic visual classification of clinical skin disease images. *Lecture Notes in Computer Science*, 2016, vol. 9910, pp. 206–222. https://doi.org/10.1007/978-3-319-46466-4_13
16. Kawahara J., Daneshvar S., Argenziano G., Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 2019, vol. 23, no. 2, pp.538–546. https://doi.org/10.1109/JBHI.2018.2824327
17. Baid U., Ghodasara S., Mohan S., Bilello M., Calabrese E., Colak E., et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv*, 2021, arXiv:2107.02314. https://doi.org/10.48550/arXiv.2107.02314
18. Bilic P, Christ P., Li H.B., Vorontsov E., Ben-Cohen A., Kaissis G., et al. The Liver Tumor Segmentation benchmark (LiTS). *arXiv*, 2019, arXiv:190.04056. https://doi.org/10.48550/arXiv.1901.04056

Авторы

Милантьев Сергей Андреевич — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57225127274, https://orcid.org/0000-0002-1970-5217, geerkus@gmail.com

Михайлова Полина Дмитриевна — магистр, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197022, Российская Федерация, https://orcid.org/0009-0005-9731-0105, polina.delitzsch@gmail.com

Бессмертный Игорь Александрович — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 36661767800, https://orcid.org/0000-0001-6711-6399, bessmertny@itmo.ru

Authors

Sergey A. Milantev — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57225127274, https://orcid.org/0000-0002-1970-5217, geerkus@gmail.com

Polina D. Mikhailova — Magister, Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, 197022, Russian Federation, https://orcid.org/0009-0005-9731-0105, polina.delitzsch@gmail.com

Igor A. Bessmertny — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 36661767800, https://orcid.org/0000-0001-6711-6399, bessmertny@itmo.ru

Статья поступила в редакцию 03.04.2025
Одобрена после рецензирования 28.05.2025
Принята к печати 17.07.2025

Received 03.04.2025
Approved after reviewing 28.05.2025
Accepted 17.07.2025