Авторлар туралы мәліметтер

Индира Байназарова – PhD докторант, Автоматтандыру және басқару кафедрасы, М. Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан; e-mail: B.k.indira@mail.ru. ORCID: https://orcid.org/0009-0009-1553-7090.

Жанат Умарова* – PhD, қауымдастырылған профессор, Ақпараттық жүйелер және модельдеу кафедрасы, М. Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан; e-mail: Zhanat-u@mail.ru. ORCID: https://orcid.org/0000-0002-0257-4417.

Юсмади Джусох – қауымдастырылған профессор, Компьютерлік ғылымдар және ақпараттық технологиялар факультеті, Путра Малайзия университеті, Малайзия; e-mail: yusmadi@upm.edu.my. ORCID: https://orcid.org/0000-0002-7767-5001.

Жалғасбек Изтаев – педагогика ғылымдарының кандидаты, қауымдастырылған профессор, Ақпараттық жүйелер және модельдеу кафедрасы, М. Әуезов атындағы Оңтүстік Қазақстан университеті, Шымкент, Қазақстан; e-mail: zhalgasbek71@mail.ru. ORCID: https://orcid.org/0000-0002-3210-2963.

Сәуле Ботаева – техника ғылымдарының кандидаты, IT және цифрлық технологиялар кафедрасы, Тәшенев университеті, Шымкент, Қазақстан; e-mail: saule_bb@mail.ru. ORCID: https://orcid.org/0000-0002-4995-7730.

Received 01.06.2026
Revised 18.08.2026
Accepted 19.08.2026

https://doi.org/10.53360/2788-7995-2026-3(23)-22

МРНТИ: 28.23.15

Г.К. Абишева, Б.Ш. Разахова, Т.А. Айдынов*, М.К. Сапарбек
Евразийский национальный университет имени Л. Н.Гумилева,
10000, Республика Казахстан, г. Астана, ул. Сатпаева, 2
*e-mail: tolegen.ch@gmail.com

ИССЛЕДОВАНИЕ ПУБЛИЧНЫХ ДАТАСЕТОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЖЕЛЕЗНОДОРОЖНЫХ ПРИЛОЖЕНИЙ

Аннотация: *Развитие технологий искусственного интеллекта (ИИ) в железнодорожной отрасли напрямую зависит от наличия качественных и репрезентативных публичных датасетов. В статье представлен систематический анализ открытых наборов данных, применяемых для задач обнаружения дефектов рельсов, диагностики контактной сети, мониторинга подвижного состава и обеспечения безопасности железнодорожной инфраструктуры. Анализ выполнен в соответствии с методологией PRISMA: поиск проводился в базах IEEE Xplore, Scopus и Web of Science, а также в репозиториях Kaggle, GitHub и Zenodo. Проанализировано 30 ключевых публикаций и 13 открытых датасетов, которые классифицированы по типу сенсоров (RGB, LiDAR, вибрационные, мультисенсорные), задачам машинного обучения (детекция, сегментация, классификация), объёму данных и формату аннотаций. Установлено, что большинство наборов данных основаны на RGB-изображениях, тогда как доля мультимодальных датасетов постепенно увеличивается. Выявлены ключевые ограничения существующих наборов данных -малый объём выборок, дисбаланс классов, ограниченная географическая вариативность и отсутствие унифицированных benchmark-протоколов, и определены направления дальнейших исследований. Полученные результаты подчёркивают необходимость создания крупных мультимодальных наборов данных и разработки единых стандартов аннотации и оценки моделей ИИ.*

Ключевые слова. *железная дорога; искусственный интеллект; датасет; компьютерное зрение; дефект рельса; сегментация; YOLO; U-Net.*

Введение. Железнодорожный транспорт является одной из ключевых составляющих современной транспортной инфраструктуры и играет важную роль в обеспечении устойчивого развития экономики. Повышение интенсивности перевозок и увеличение скорости движения поездов предъявляют высокие требования к надёжности и безопасности железнодорожной инфраструктуры. В связи с этим особое значение приобретает своевременное обнаружение

©Г.К. Абишева, Б.Ш. Разахова, Т.А. Айдынов, М.К. Сапарбек, 2026

дефектов элементов железнодорожного пути, таких как рельсы, шпалы, крепления и балластный слой. Традиционные методы инспекции железнодорожных путей, основанные на визуальном осмотре и специализированных измерительных вагонах, требуют значительных временных и трудовых ресурсов и не всегда позволяют оперативно выявлять потенциально опасные дефекты. Согласно международным исследованиям в области транспортной безопасности, значительная часть инцидентов на железных дорогах связана с дефектами инфраструктуры, выявленными с задержкой [1].

В настоящее время инспекция путей преимущественно осуществляется традиционными методами: визуальным осмотром, ультразвуковой и магнитной дефектоскопией, а также с использованием специализированных вагонов-путеизмерителей. Несмотря на высокую точность, данные методы характеризуются значительными эксплуатационными затратами и ограниченной масштабируемостью в рамках всей сети железных дорог [2]. В последние годы развитие методов искусственного интеллекта (ИИ) и глубокого обучения позволило перейти к автоматизированным системам мониторинга, основанным на компьютерном зрении и анализе сенсорных данных.

В последние годы мониторинг путевой инфраструктуры всё чаще осуществляется с применением нейронных сетей и методов компьютерного зрения. Вместо ручного анализа видеозаписей системы обучаются выполнять данную задачу автоматически: камеры, установленные на дронах или инспекционных поездах, передают видеопоток, а алгоритмы, такие как YOLO и U-Net, обнаруживают трещины и иные повреждения в режиме, близком к реальному времени. В условиях зашумлённых изображений подобные глубокие нейронные сети демонстрируют существенно более высокую эффективность по сравнению с традиционными методами [3].

Результаты исследований последних лет свидетельствуют о том, что применение ИИ в диагностике железнодорожной инфраструктуры перешло из экспериментальной стадии в разряд практической необходимости. Нейросетевые модели способны обнаруживать мелкие дефекты с точностью, превышающей возможности человека-оператора, а также снижать вероятность ошибок, связанных с человеческим фактором.

Основные направления, где ИИ демонстрирует наилучшие результаты:

- поиск повреждений на рельсах [4];
- контроль отсутствия посторонних предметов на путях [5];
- контроль состояния балласта и шпал [6];
- проверка контактных сетей [7];
- анализ вибраций для прогнозирования неисправностей подвижного состава [8].

Одними из первых применение сверточных нейронных сетей (CNN) для дефектоскопии рельсов продемонстрировали Faghih-Roohi et al. [9]. Впоследствии Zhang et al. [10] расширили данный подход за счёт использования более мощных архитектур, таких как ResNet [11] и U-Net [12], что позволило существенно повысить точность сегментации дефектов. Тем не менее ключевой проблемой на сегодняшний день остаётся не столько развитие самих алгоритмов, сколько поиск качественных обучающих данных и разработка эффективных методов обработки больших объёмов сенсорной информации.

В задачах контроля путевой инфраструктуры широко применяются алгоритмы детекции объектов, такие как YOLO [13] и Faster R-CNN [14]. Если ранние версии данных архитектур требовали выбора между скоростью и точностью, то современные модификации YOLO позволяют обрабатывать видеопоток в реальном времени без существенной потери качества. Для задач, требующих точного выделения границ элементов инфраструктуры, чаще применяются архитектуры семантической сегментации, такие как DeepLabV3+ [15], либо трансформерные модели, например Vision Transformer [16].

Вместе с тем ключевым ограничивающим фактором для любого алгоритма остаётся качество и объём обучающих данных: без достаточного количества размеченных изображений эффективность нейросетевых моделей существенно снижается. Поиск подходящих датасетов затруднён узкоспециализированным характером разметки, требуемой для задач железнодорожной отрасли. Данная проблема постепенно решается – в частности, для задач семантической сегментации широкое распространение получил датасет RailSem19 [17], характеризующийся детальной проработкой элементов железнодорожной сцены.

Помимо специализированных датасетов для отдельных задач, исследователи всё чаще используют мультисенсорные наборы данных, такие как OSDaR23 [18], объединяющие данные камер и лидаров, установленных на подвижном составе. Комбинирование различных типов сенсоров имеет существенное значение для разработки систем автономного управления, поскольку позволяет алгоритмам получать не только плоскостное, но и объёмное представление об окружающей среде. Появление подобных открытых датасетов способствует развитию отрасли, предоставляя разработчикам возможность тестировать модели на реальных, а не синтетических данных.

Тем не менее ситуация с доступностью данных в отрасли остаётся неоднозначной. Pappaterra et al. [19] провели систематический анализ и установили, что количество открытых железнодорожных датасетов по-прежнему критически недостаточно. Несмотря на широкий охват рассмотренных источников - от данных на переездах до сенсорных данных подвижного состава, - прогресс в данной области существенно ограничен дефицитом качественных общедоступных массивов данных по сравнению, например, с автомобильным сектором.

Несмотря на стремительный рост числа публикаций после 2018 года [20-22], ключевой проблемой остаётся ограниченность и фрагментарность доступных обучающих данных. В отличие от универсальных компьютерных датасетов, таких как ImageNet [23] и COCO [24], железнодорожные наборы данных характеризуются:

- малым объёмом выборок;
- дисбалансом классов;
- ограниченной географической и климатической вариативностью;
- отсутствием мультимодальности;
- отсутствием унифицированных benchmark-протоколов (эталонных протоколов сравнительного тестирования).

Большинство исследований сосредоточено на разработке алгоритмов, в то время как систематизированный анализ самих датасетов остаётся недостаточно представленным в научной литературе. Отсутствие комплексного обзора публичных железнодорожных датасетов затрудняет воспроизводимость экспериментов, объективное сравнение моделей и формирование стандартизированных метрик оценки.

В связи с этим возникает необходимость в проведении систематического анализа существующих публичных датасетов ИИ для железнодорожных приложений с использованием формализованной методологии отбора и оценки публикаций.

Научная новизна настоящей работы состоит не в самом факте проведения систематического обзора, выполненного ранее Pappaterra et al. [19] для публикаций до 2021 года, а в существенном обновлении и переосмыслении ландшафта открытых железнодорожных датасетов за период 2021-2026 годов. В отличие от предшествующего обзора, в данном исследовании: (1) систематизированы недавно появившиеся мультимодальные (камера + LiDAR + радар), трёхмерные (облака точек LiDAR) и синтетические либо AR-дополненные наборы данных (OSDaR23, OSDaR-AR, WHU-Railway3D, SynRailObs, SynDRA, RailGoerl24), отсутствовавшие на момент публикации [19]; (2) предложена трёхуровневая классификация датасетов одновременно по типу сенсоров, задачам машинного обучения и уровню аннотации; (3) выявлен и количественно охарактеризован сдвиг от одномодальных RGB-наборов к мультимодальным и синтетическим данным, а также установлено отсутствие унифицированных benchmark-протоколов как ключевого барьера воспроизводимости. Тем самым работа не дублирует, а развивает и актуализирует результаты [19].

Таким образом, актуальной задачей является анализ существующих открытых наборов данных, используемых в железнодорожных исследованиях, а также оценка их применимости для разработки интеллектуальных систем мониторинга и диагностики железнодорожной инфраструктуры.

Методы исследования. Для обеспечения прозрачности и воспроизводимости процедуры отбора исследований в настоящей работе используется подход, рекомендованный международным руководством PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [25].

Железнодорожная инфраструктура является критически важной системой транспортной безопасности. Современные методы ИИ активно внедряются для:

- обнаружения трещин рельсов;
- детекции препятствий;
- анализа состояния балласта;
- мониторинга контактной сети;
- предиктивной диагностики подвижного состава.

Ключевым фактором развития является наличие открытых, стандартизированных и масштабируемых датасетов. Цель исследования - систематически проанализировать существующие публичные датасеты ИИ для железнодорожных приложений.

2.1 Источники поиска

Поиск осуществлялся в следующих научных базах данных и репозиториях: IEEE Xplore, Scopus, Web of Science, Google Scholar, Kaggle, GitHub, Zenodo.

Ключевые поисковые запросы: “railway defect dataset”, “rail crack public dataset”, “railway computer vision dataset”, “rail track inspection dataset”.

2.2 Критерии включения

Включались датасеты, которые: 1) публично доступны; 2) используются для задач ИИ; 3) имеют аннотацию; 4) относятся к железнодорожной инфраструктуре. Сводные критерии включения и исключения приведены в таблице 1.

Таблица 1 – Критерии включения и исключения

Критерии включения	Критерии исключения
Публикация относится к железнодорожной инфраструктуре	Проприетарные или закрытые датасеты
Используются методы искусственного интеллекта	Отсутствие описания структуры данных
Описан публично доступный датасет	Статьи без эмпирической проверки
Указаны характеристики данных (объём, формат, аннотация)	Дублирующие публикации
Доступна полнотекстовая версия статьи	Технические отчёты без научного рецензирования

2.3 Стратегия поиска

Поиск публикаций осуществлялся в период январь–март 2026 года в следующих научных базах данных: IEEE Xplore, Scopus, Web of Science Core Collection, ScienceDirect, SpringerLink, Google Scholar. Для поиска публичных датасетов дополнительно использовались: Kaggle, GitHub, Zenodo, Mendeley Data, IEEE DataPort. Данные источники кратко описаны в таблице 2.

Таблица 2 – Источники поиска

Репозиторий	Краткое описание	URL
Kaggle	Платформа для задач машинного обучения и размещения публичных датасетов, поддерживаемая Google	https://www.kaggle.com
IEEE DataPort	Репозиторий научных датасетов, поддерживаемый IEEE, включает рецензируемые наборы данных	https://ieee-dataport.org
Zenodo	Европейский открытый репозиторий научных публикаций и датасетов (поддержка CERN)	https://zenodo.org
Mendeley Data	Репозиторий исследовательских данных, связанный с Elsevier	https://data.mendeley.com
Figshare	Платформа для публикации научных данных и материалов исследований	https://figshare.com
Google Dataset Search	Поисковая система Google для обнаружения открытых датасетов	https://datasetsearch.research.google.com
Data.gov	Портал открытых данных правительства США	https://www.data.gov
Data.gov.uk	Портал открытых данных правительства Великобритании	https://data.gov.uk
European Data Portal	Портал открытых данных стран Европейского союза	https://data.europa.eu
OpenRail Data (UK)	Платформа открытых железнодорожных данных Великобритании (расписания, движение, API)	https://opendata.nationalrail.co.uk

Поисковая строка формировалась с использованием булевых операторов AND/OR:

(“railway” OR “rail track” OR “rail infrastructure”) AND (“dataset” OR “public dataset”) AND (“deep learning” OR “artificial intelligence” OR “computer vision”).

Дополнительно применялись уточняющие запросы: “rail crack dataset”, “rail defect segmentation dataset”, “railway intrusion dataset”, “railway LiDAR dataset”. Период публикаций: 2010–2026 гг. Язык публикаций: английский.

2.4 Процедура отбора исследований

Процедура отбора выполнялась в четыре этапа:

- Идентификация - первичный поиск по ключевым словам (n = 412).
- Удаление дубликатов - автоматическая фильтрация и ручная проверка (n = 326).
- Скрининг аннотаций - исключение нерелевантных работ (n = 118).
- Полнотекстовая оценка - анализ методологии и доступности данных (n = 54).

В итоговый обзор включено 30 научных публикаций.

Все включённые датасеты были распределены по четырём основным категориям:

1. Дефекты рельсов (поверхностные трещины, сколы, износ).
2. Обнаружение препятствий и вторжений.
3. LiDAR и 3D-сегментация инфраструктуры.
4. Вибрационная и акустическая диагностика.

Дополнительно проводилась классификация по уровню аннотации: Image-level labels (аннотации на уровне изображения), Bounding box (ограничивающая рамка), Pixel-wise segmentation (попиксельная сегментация), 3D semantic annotation (трёхмерная семантическая аннотация).

На основании описанной методологии в следующем разделе представлен детальный анализ выявленных публичных датасетов, их количественных характеристик и сравнительная классификация по типу задач и сенсоров.

Результаты исследований. В результате систематического поиска и отбора публикаций было выявлено 13 публичных датасетов, каждый из которых проверен по первоисточнику, применяемых для задач искусственного интеллекта в железнодорожной отрасли. Данные наборы используются для задач обнаружения дефектов рельсов, мониторинга инфраструктуры, обнаружения препятствий и анализа сенсорных данных. Полный перечень выявленных датасетов и их характеристики представлены в таблице 3.

Таблица 3 – Публичные датасеты для железнодорожных приложений, выявленные в ходе систематического обзора

Датасет	Год	Тип данных	Основная задача	Объём данных	Источник
1	*2*	*3*	*4*	*5*	*6*
RailSem19	2019	RGB-изображения	семантическая сегментация ж/д сцены	~8500 изображений	Zendel et al. [17]
OSDaR23	2023	RGB + LiDAR + радар	мультимодальная сегментация/ детекция	45 последовательностей, ~204k кадров	Tagiew et al. [18]
Rail-5K	2021	RGB-изображения	детекция дефектов рельсов	5000 размеченных + 3000 неразмеченных	Zhang et al. [26]
SynRailObs	2025	синтетические RGB	детекция препятствий	~1000 изображений	Guo, Rambach [27]
OSDaR-AR	2026	RGB + LiDAR (AR-дополнение)	детекция препятствий, аугментация данных	18 дополненных последовательностей (2 камеры + LiDAR)	Nesti et al. [28]
RSDDs (Type I/II)	2014	RGB (Ч/Б)	Детекция /сегментация дефектов рельсов	195 изображений (67 + 128)	Niu et al. [29]
WHU-Railway3D	2024	LiDAR (облако точек)	семантическая сегментация инфраструктуры	~30 км, 4.6 млрд точек, 11 классов	Qiu et al. [30]
FRSign	2020	RGB-изображения	детекция/ распознавание сигналов	>100 000 изображений	Harb et al. [31]
GERALD	2023	RGB-изображения	детекция сигналов (Германия)	5000 изображений	Leibner et al. [32]

1	2	3	4	5	6
RailGoerl24	2025	RGB (Full HD) + LiDAR	детекция людей на путях	12 205 кадров, 33 556 аннотаций	Tagiew et al. [33]
SynDRA	2025	синтетические RGB	семантическая сегментация ж/д сцены	RGB-последовательности + карты сегментации	D'Amico et al. [34]
Railway Track Fault Detection	2022	RGB-изображения	классификация дефектов рельсов/креплений	383 изображения	Eunus et al. [35]
Augmented Ballast Dataset	2024	RGB-изображения (реальные + синтетические)	сегментация/оценка состояния балласта	многомерный агрегированный набор	Ding et al. [36]

3.1 Классификация датасетов по типу сенсоров

Для более детального анализа выявленных наборов данных была проведена их классификация по типу используемых сенсоров. Тип сенсорных данных является одним из ключевых факторов, определяющих возможности применения датасета в задачах искусственного интеллекта, поскольку он влияет на выбор архитектуры моделей машинного обучения, методы обработки данных и точность алгоритмов обнаружения и сегментации объектов.

В современных системах мониторинга железнодорожной инфраструктуры используются различные типы сенсоров, включая оптические камеры, лазерные сканеры (LiDAR), а также вибрационные и акустические датчики. RGB-камеры обеспечивают высокую детализацию изображений и широко применяются для задач компьютерного зрения, в то время как LiDAR позволяет получать трёхмерную информацию о геометрии инфраструктуры.

В рамках настоящего исследования все выявленные публичные датасеты были распределены по типу используемых сенсоров. Результаты классификации представлены в таблице 4.

Таблица 4 – Классификация датасетов по типу сенсоров

Тип сенсора	Количество датасетов
RGB-изображения	9
LiDAR	1
Мультисенсорные данные (RGB + LiDAR/радар)	3

Анализ показывает, что большинство датасетов основаны на RGB-изображениях, поскольку камеры являются наиболее доступным и относительно недорогим источником данных для мониторинга железнодорожной инфраструктуры. Это объясняется высокой доступностью камерных систем и простотой сбора данных: RGB-камеры широко применяются в системах автоматической инспекции путей, поскольку позволяют фиксировать визуальные дефекты рельсов, состояние шпал, балластного слоя и других элементов инфраструктуры.

Видеопоток остаётся основным источником данных для задач, связанных с безопасностью – мониторинга переездов и поиска препятствий на путях, поскольку позволяет анализировать динамику сцены в момент возникновения угрозы.

Датасеты на основе LiDAR встречаются заметно реже, однако именно они рассматриваются как основной источник данных для построения точных 3D-моделей инфраструктуры: в отличие от камер, лидары обеспечивают точную пространственную геометрию, что критически важно при переходе от простого распознавания объектов к их картографированию и решению задач автономного управления.

Мультимодальные датасеты, объединяющие данные камер, LiDAR и инерциальных датчиков, представляют собой наиболее перспективное направление развития железнодорожных систем искусственного интеллекта, поскольку позволяют повысить устойчивость алгоритмов и точность обнаружения объектов и дефектов в сложных условиях эксплуатации.

3.2 Анализ датасетов по задачам искусственного интеллекта

Анализ выявленных наборов данных показал, что большинство железнодорожных датасетов разрабатываются для задач компьютерного зрения, связанных с обнаружением

дефектов инфраструктуры и мониторингом безопасности движения. Основными задачами искусственного интеллекта, для которых используются данные наборы данных, являются обнаружение объектов, семантическая сегментация, классификация и мультимодальный анализ.

Наиболее распространённой задачей является детекция объектов, применяемая для обнаружения посторонних предметов на путях, сигналов и людей в опасной зоне. Для решения таких задач широко используются алгоритмы глубокого обучения, включая архитектуры YOLO, Faster R-CNN и SSD.

Второй по распространённости задачей является семантическая сегментация, позволяющая проводить детальный анализ железнодорожной сцены на уровне пикселей или точек облака. Для данных задач используются архитектуры U-Net, DeepLabV3+ и специализированные сети для облаков точек.

Задачи классификации применяются преимущественно для оценки состояния железнодорожных элементов – например, определения степени износа рельсов или наличия дефекта на изображении. Мультимодальный анализ, объединяющий данные нескольких типов сенсоров, представлен пока небольшим числом датасетов, однако их доля постепенно растёт (табл. 5).

Таблица 5 – Классификация датасетов по задачам машинного обучения

Задача ИИ	Количество датасетов	Пример
Детекция объектов/дефектов	6	SynRailObs, RailGoerl24, Rail-5K
Сегментация	3	RailSem19, WHU-Railway3D, SynDRA
Классификация	2	Railway Track Fault Detection, Augmented Ballast Dataset
Мультимодальный анализ	2	OSDaR23, OSDaR-AR

Анализ выявленных датасетов показывает несколько ключевых тенденций. Во-первых, большинство доступных наборов данных ориентированы на задачи компьютерного зрения, такие как обнаружение дефектов рельсов и сегментация железнодорожных объектов, что объясняется высокой эффективностью методов глубокого обучения в обработке изображений. Во-вторых, наблюдается постепенный рост числа мультимодальных датасетов, включающих данные LiDAR и видео, что позволяет разрабатывать более устойчивые модели для автономных систем мониторинга. В-третьих, большинство существующих датасетов имеют относительно небольшой объём данных по сравнению с крупными компьютерными датасетами общего назначения, такими как ImageNet или COCO, что ограничивает возможности обучения глубоких нейронных сетей и снижает обобщающую способность моделей.

3.3 Ограничения существующих датасетов

Проведённый анализ позволил выявить несколько ключевых ограничений существующих публичных железнодорожных датасетов:

– Большинство наборов данных имеют ограниченный объём, что затрудняет обучение глубоких нейронных сетей и снижает точность моделей при работе с новыми данными; многие датасеты ориентированы только на один тип сенсорных данных, чаще всего RGB-изображения, в то время как современные системы мониторинга требуют мультимодальных данных.

– Отсутствие стандартизированных протоколов оценки и унифицированных benchmark-наборов данных усложняет объективное сравнение различных алгоритмов и моделей искусственного интеллекта.

– Часть датасетов содержит ограниченную или неполную аннотацию, что снижает возможности их использования для задач семантической сегментации или трёхмерного анализа инфраструктуры.

На основе полученных результатов в следующем разделе проводится обсуждение ключевых ограничений существующих железнодорожных датасетов и формулируются направления дальнейших исследований.

Обсуждение научных результатов. Проведённый систематический анализ публичных датасетов показал, что развитие методов искусственного интеллекта в железнодорожной отрасли тесно связано с доступностью качественных обучающих данных.

Несмотря на активный рост публикаций в области применения глубокого обучения для диагностики железнодорожной инфраструктуры, количество открытых и хорошо структурированных датасетов остаётся относительно ограниченным.

Одной из ключевых тенденций является доминирование датасетов, основанных на RGB-изображениях. Это связано с тем, что камеры являются наиболее доступным и широко используемым источником данных для мониторинга железнодорожных путей. Такие датасеты позволяют эффективно решать задачи обнаружения дефектов рельсов, сегментации инфраструктурных элементов и выявления препятствий на путях. Однако использование исключительно визуальных данных может ограничивать возможности анализа, особенно в условиях плохой освещённости, неблагоприятных погодных условий или при наличии скрытых дефектов.

Другой важной тенденцией является постепенное появление мультимодальных датасетов, включающих данные LiDAR, тепловизионных камер и других сенсорных систем. Такие наборы данных позволяют значительно расширить возможности анализа инфраструктуры и повышают надёжность алгоритмов искусственного интеллекта. Мультимодальный подход становится особенно важным для задач автономного управления железнодорожным транспортом и интеллектуальных систем поддержки машиниста.

Тем не менее проведённый анализ выявил ряд существенных ограничений существующих датасетов. Во-первых, многие наборы данных имеют относительно небольшой объём, что затрудняет обучение глубоких нейронных сетей и снижает способность моделей к обобщению. Во-вторых, значительная часть датасетов охватывает ограниченные географические регионы, что может влиять на универсальность разработанных алгоритмов. Кроме того, в большинстве случаев отсутствуют стандартизированные benchmark-наборы данных и унифицированные методики оценки моделей, что усложняет объективное сравнение алгоритмов и снижает воспроизводимость исследований.

Таким образом, дальнейшее развитие интеллектуальных систем мониторинга железнодорожной инфраструктуры требует создания крупных и разнообразных датасетов, включающих мультимодальные сенсорные данные, а также разработки единых стандартов аннотации и оценки моделей искусственного интеллекта.

Заключение. В данной работе представлен систематический обзор публичных датасетов искусственного интеллекта, применяемых для задач железнодорожных приложений. Обзор выполнен с использованием методологии PRISMA и включает анализ научных публикаций и открытых репозиториев данных.

В результате проведённого исследования было выявлено 13 публичных датасетов, используемых для задач обнаружения дефектов рельсов, мониторинга инфраструктуры, обнаружения препятствий и диагностики подвижного состава. Датасеты были классифицированы по типу сенсоров, типу аннотаций и задачам машинного обучения.

Анализ показал, что большинство существующих наборов данных основаны на RGB-изображениях и ориентированы на задачи компьютерного зрения, такие как детекция объектов и семантическая сегментация. При этом количество мультимодальных датасетов, включающих данные LiDAR и других сенсоров, постепенно увеличивается, что отражает современную тенденцию развития интеллектуальных транспортных систем.

В ходе исследования также были выявлены ключевые ограничения существующих датасетов, включая небольшой объём данных, дисбаланс классов, ограниченную географическую вариативность и отсутствие стандартизированных benchmark-протоколов.

Полученные результаты показывают необходимость разработки новых масштабируемых и мультимодальных датасетов для железнодорожных приложений. Создание таких наборов данных позволит повысить точность моделей искусственного интеллекта и ускорить внедрение интеллектуальных систем мониторинга и диагностики железнодорожной инфраструктуры.

В дальнейшем исследования могут быть направлены на формирование открытых международных benchmark-датасетов, объединяющих различные типы сенсорных данных, а также на разработку методов автоматизированной разметки данных для повышения эффективности подготовки обучающих выборок.

Список литературы

1. European Union Agency for Railways. Railway Safety Performance in the European Union 2022. – Luxembourg: Publications Office of the European Union, 2022.
2. Papaelias M. A review on non-destructive evaluation of rails / M. Papaelias, C. Roberts, C. Davis // Insight -Non-Destructive Testing and Condition Monitoring. – 2008. – Vol. 50, № 7. – P. 367-373.
3. Goodfellow I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge: MIT Press, 2016. – 775 p.
4. Rail surface defect detection method based on YOLOv3 deep learning networks / S. Yanan et al // 2018 Chinese Automation Congress (CAC). – IEEE, 2018. – P. 1563-1568. https://doi.org/10.1109/CAC.2018.8623082.
5. Ning S. Research on the method of foreign object detection for railway tracks based on deep learning / S. Ning, F. Ding, B. Chen // Sensors. – 2024. – Vol. 24, № 14. – Article 4483. https://doi.org/10.3390/s24144483.
6. Deep learning approach for automated railroad ballast condition evaluation / J. Luo et al // Proceedings of the 5th International Conference on Transportation Geotechnics (ICTG 2024). Lecture Notes in Civil Engineering. – Springer, 2025. – Vol. 402. – P. 49-57. https://doi.org/10.1007/978-981-97-8213-0_7.
7. A robust pantograph-catenary interaction condition monitoring method based on deep convolutional network / D. Zhang et al // IEEE Transactions on Instrumentation and Measurement. – 2020. – Vol. 69, № 5. – P. 1920-1929. https://doi.org/10.1109/TIM.2019.2920721.
8. Applications of machine learning to machine fault diagnosis: A review and roadmap / Y. Lei et al // Mechanical Systems and Signal Processing. – 2020. – Vol. 138. – Article 106587. https://doi.org/10.1016/j.ymssp.2019.106587.
9. Deep convolutional neural networks for detection of rail surface defects / S. Faghih-Roohi et al // Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN). – IEEE, 2016. – P. 2584-2589. https://doi.org/10.1109/IJCNN.2016.7727522.
10. MCnet: Multiple context information segmentation network of no-service rail surface defects / D. Zhang et al // IEEE Transactions on Instrumentation and Measurement. – 2021. – Vol. 70. – Article 5004909. https://doi.org/10.1109/TIM.2020.3040890.
11. Deep residual learning for image recognition / K. He et al // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – P. 770-778.
12. Ronneberger O. U-Net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer, T. Brox // Medical Image Computing and Computer-Assisted Intervention. – Cham: Springer, 2015. – P. 234-241.
13. You Only Look Once: Unified, real-time object detection / J. Redmon et al // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – P. 779-788.
14. Faster R-CNN: Towards real-time object detection with region proposal networks / S. Ren et al // Advances in Neural Information Processing Systems. – 2015. – Vol. 28.
15. Encoder-decoder with atrous separable convolution for semantic image segmentation / L.-C. Chen et al // Proceedings of the European Conference on Computer Vision. – 2018. – P. 801-818.
16. An image is worth 16x16 words: Transformers for image recognition at scale / A. Dosovitskiy et al // International Conference on Learning Representations. – 2021.
17. RailSem19: A Dataset for Semantic Rail Scene Understanding / O. Zendel et al // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – 2019.
18. OSDaR23: Open sensor data for rail 2023 / R. Tagiew et al // 2023 8th International Conference on Robotics and Automation Engineering. – IEEE, 2023. – P. 270-276.
19. A systematic review of artificial intelligence public datasets for railway applications / M.J. Pappaterra et al // Infrastructures. – 2021. – Vol. 6, № 9. – Article 136.
20. A survey on audio-video based defect detection through deep learning in railway maintenance / L. De Donato et al // IEEE Access. – 2022. – Vol. 10. – P. 65376-65400. https://doi.org/10.1109/ACCESS.2022.3183102.
21. Liu S. A review of applications of visual inspection technology based on image processing in the railway industry / S. Liu, Q. Wang, Y. Luo // Transportation Safety and Environment. – 2019. – Vol. 1, № 3. – P. 185-204. https://doi.org/10.1093/tse/tdz007.

22. Artificial intelligence in railway transport: Taxonomy, regulations and applications / N. Bešinović et al // IEEE Transactions on Intelligent Transportation Systems. – 2022. – Vol. 23, № 9. – P. 14011-14024. https://doi.org/10.1109/TITS.2021.3131637.
23. ImageNet: A large-scale hierarchical image database / J. Deng et al // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2009. – P. 248-255.
24. Microsoft COCO: Common objects in context / T.-Y. Lin et al // Proceedings of the European Conference on Computer Vision. – Cham: Springer, 2014. – P. 740-755.
25. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews / M. Page et al // BMJ. – 2021. – Vol. 372. – Article n71.
26. Rail-5k: A real-world dataset for rail surface defects detection / Z. Zhang et al // arXiv preprint. – 2021. – arXiv:2106.14366.
27. Guo Q. SynRailObs: A synthetic dataset for obstacle detection in railway scenarios / Q. Guo, J. Rambach // arXiv preprint. – 2025. -arXiv:2505.10784.
28. OSDaR-AR: Enhancing Railway Perception Datasets via Multi-modal Augmented Reality / F. Nesti et al // arXiv preprint. – 2026. – arXiv:2602.22920.
29. Unsupervised saliency detection of rail surface defects using stereoscopic images / M. Niu et al // IEEE Transactions on Industrial Informatics. – 2021. – Vol. 17, № 3. – P. 2271-2281.
30. WHU-Railway3D: A diverse dataset and benchmark for railway point cloud semantic segmentation / B. Qiu et al // IEEE Transactions on Intelligent Transportation Systems. – 2024. – Vol. 25. – P. 20900-20916.
31. FRSign: A large-scale traffic light dataset for autonomous trains / J. Harb et al // arXiv preprint. – 2020.
32. Leibner P. GERALD: A novel dataset for the detection of German mainline railway signals / P. Leibner, F. Hampel, C. Schindler // Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit. – 2023.
33. RailGoerl24: Görlitz Rail Test Center CV Dataset 2024 / R. Tagiew et al // arXiv preprint. – 2025. – arXiv:2504.00204.
34. SynDRA: Synthetic dataset for railway applications / G. D'Amico et al // 2025 IEEE/CVF Winter Conference on Applications of Computer Vision. – IEEE, 2025. – P. 3437-3446.
35. Eunus S.I. Railway Track Fault Detection [Electronic resource] / S.I. Eunus, S. Hossain, A.E.M. Ridwan // Kaggle Dataset. – 2022. https://www.kaggle.com/datasets/salmaneunus/railway-track-fault-detection (accessed: 09.07.2026).
36. Augmented dataset for vision-based analysis of railroad ballast via multi-dimensional data synthesis / K. Ding et al // Algorithms. – 2024. – Vol. 17, № 8. – Article 367.

Финансирование. *Данное исследование финансируется Комитетом науки Министерства науки и высшего образования Республики Казахстан (грант № AP25796503).*

G.K. Abisheva, B.Sh. Razakhova, T.A. Aidynov*, M.K. Saparbek
L.N. Gumilyov Eurasian National University,
10000 Kazakhstan, Astana city, 2 Satbaev st.
*e-mail: tolegen.ch@gmail.com

RESEARCH OF PUBLIC ARTIFICIAL INTELLIGENCE DATASETS FOR RAILWAY APPLICATIONS

The development of artificial intelligence (AI) technologies in the railway industry directly depends on the availability of high-quality and representative public datasets. This article provides a systematic analysis of open datasets used for the tasks of rail defect detection, contact network diagnostics, rolling stock monitoring and railway infrastructure safety. The analysis was conducted following the PRISMA methodology: the search covered IEEE Xplore, Scopus and Web of Science, as well as the Kaggle, GitHub and Zenodo repositories. In total, 30 key publications and 13 open datasets were analysed and classified by sensor type (RGB, LiDAR, vibration, multi-sensor), machine learning task (detection, segmentation, classification), data volume and annotation format. The analysis shows that most datasets are based on RGB images, while the share of multimodal datasets is gradually increasing. Key limitations of existing datasets were identified -small sample size, class imbalance, limited geographical variability and the absence of unified benchmark protocols and directions for further research were defined. The results highlight the need to create large multimodal datasets and to develop unified standards for the annotation and evaluation of AI models.

Key words: *railway dataset, artificial intelligence, computer vision, rail defect detection, PRISMA, deep learning.*

Г.К. Абишева, Б.Ш. Разахова, Т.А. Айдынов*, М.К. Сапарбек
Л.Н. Гумилев атындағы Еуразия Ұлттық Университеті,
10000 Қазақстан, Астана қ., Сәтбаев к-сі, 2
*e-mail: tolegen.ch@gmail.com

ТЕМІРЖОЛ ҚОСЫМШАЛАРЫНА АРНАЛҒАН ҚОҒАМДЫҚ ЖАСАНДЫ ИНТЕЛЛЕКТ ДЕРЕКТЕРІН ЗЕРТТЕУ

Теміржол саласында жасанды интеллект (ЖИ) технологияларының дамуы сапалы әрі репрезентативті ашық деректер жиынтықтарының болуына тікелей байланысты. Мақалада рельс ақауларын анықтау, контакт желісін диагностикалау, жылжымалы құрам мониторингі және теміржол инфрақұрылымының қауіпсіздігін қамтамасыз ету міндеттері үшін қолданылатын ашық деректер жиынтықтарына жүйелі талдау жасалған. Шолу PRISMA әдіснамасы негізінде орындалды: іздеу IEEE Xplore, Scopus, Web of Science дереккөздерінде, сондай-ақ Kaggle, GitHub және Zenodo репозиторийлерінде жүргізілді. Барлығы 30 негізгі басылым және 13 ашық датасет талданып, олар сенсор түрі (RGB, LiDAR, вибрациялық, мультисенсорлық), машиналық оқыту міндеті (детекция, сегментация, классификация), деректер көлемі және аннотация форматы бойынша жіктелді. Талдау көрсеткендей, датасеттердің басым бөлігі RGB-кескіндерге негізделген, ал мультимодальды датасеттердің үлесі біртіндеп артып келеді. Қолданыстағы датасеттердің негізгі шектеулері -іріктеме көлемінің аздығы, кластар теңгерімсіздігі, географиялық вариативтіліктің шектеулілігі және бірыңғай benchmark-хаттамалардың болмауы анықталып, әрі қарайғы зерттеу бағыттары ұсынылды. Алынған нәтижелер ірі мультимодальды деректер жиынтықтарын құру және ЖИ модельдерін аннотациялау мен бағалаудың бірыңғай стандарттарын әзірлеу қажеттілігін көрсетеді.

Түйін сөздер: *теміржол, жасанды интеллект, датасет, компьютерлік көру, рельс ақауы, сегментация, YOLO, U-Net.*

Авторлар туралы мәліметтер

Гульсипат Кайыргаликызы Абишева – жасанды интеллект технологиялары кафедрасының докторанты, Л.Н. Гумилев Атындағы Еуразия Ұлттық Университеті, Астана қ., Қазақстан; e-mail: gulsipat.abisheva@gmail.com. ORCID: https://orcid.org/0009-0003-6692-197X.

Бибигул Шамшановна Разахова – т.ғ.к., жасанды интеллект технология кафедрасының басқарушысы, Л.Н. Гумилев Атындағы Еуразия Ұлттық Университеті, Астана қ., Қазақстан; e-mail: razakhova_bsh@enu.kz. ORCID: https://orcid.org/0000-0002-8152-8661.

Төлеген Айдынұлы Айдынов* – хат-хабар авторы, ақпараттық қауіпсіздік кафедрасының докторанты, Л.Н. Гумилев Атындағы Еуразия Ұлттық Университеті, Астана қ., Қазақстан; e-mail: tolegen.ch@gmail.com. ORCID: http://orcid.org/0009-0005-3327-9719.

Минира Қайырғалиқызы Сапарбек – жасанды интеллект технологиялары кафедрасының магистранты, Л.Н. Гумилев Атындағы Еуразия Ұлттық Университеті, Астана қ., Қазақстан; e-mail: sapminira@gmail.com. ORCID: https://orcid.org/0009-0005-0334-0910.

Сведения об авторах

Гульсипат Кайыргаликызы Абишева – докторант кафедры искусственного интеллекта Евразийского национального университета им. Л.Н. Гумилева, Астана, Казахстан; e-mail: gulsipat.abisheva@gmail.com. ORCID: https://orcid.org/0009-0003-6692-197X.

Бибигул Шамшановна Разахова – к.т.н., заведующий кафедрой технологий искусственного интеллекта Л.Н. Гумилева, Астана, Казахстан; e-mail: razakhova_bsh@enu.kz. ORCID: https://orcid.org/0000-0002-8152-8661.

Төлеген Айдынұлы Айдынов* – корреспондент автор, докторант кафедры информационной безопасности Евразийского национального университета им. Л.Н. Гумилева, Астана, Казахстан; e-mail: tolegen.ch@gmail.com. ORCID: http://orcid.org/0009-0005-3327-9719.

Минира Кайыргаликызы Сапарбек – магистрант кафедры искусственного интеллекта Евразийского национального университета им. Л.Н. Гумилева, Астана, Казахстан; e-mail: sapminira@gmail.com. ORCID: https://orcid.org/0009-0005-0334-0910.

Information about the authors

Gulsipat Abisheva – doctoral student of the Department of artificial intelligence, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: gulsipat.abisheva@gmail.com. ORCID: https://orcid.org/0009-0003-6692-197X.

Bibigul Razakhova – c.t.s., Head of the Department of Artificial Intelligence Technologies, L.N. Gumilev Eurasian National University, Astana, Kazakhstan; e-mail: razakhova_bsh@enu.kz. ORCID: https://orcid.org/0000-0002-8152-8661.

Tolegen Aidynov* – doctoral student of the Department of Information Security, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: tolegen.ch@gmail.com. ORCID: http://orcid.org/0009-0005-3327-9719.

Minira Saparbek – master of the Department of artificial intelligence, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: sapminira@gmail.com. ORCID: https://orcid.org/0009-0005-0334-0910.

Поступила в редакцию 19.07.2026
Поступила после доработки 29.08.2026
Принята к публикации 01.09.2026

https://doi.org/10.53360/2788-7995-2026-3(23)-23

CC BY 4.0
Check for updates

IRSTI: 28.23.02

K.E. Iklassova[1]*, L.N. Abdullah[2], M.Zh. Bazarova[3]
[1]M. Kozybayev North Kazakhstan University,
150000, Republic of Kazakhstan, Petropavllovsk, Pushkina Str, 86
[2]Putra University of Malaysia, Malaysia, Serang, Selangor, 43400 Seri Kembangan
[3]Sarsen Amanzholov East Kazakhstan University,
070002, Republic of Kazakhstan, Ust-Kamenogorsk, 30th Gvardeiskoy Divisii Str, 34
*e-mail: keiklasova@ku.edu.kz

ASSESSMENT OF INSTITUTIONAL DATA HETEROGENEITY FOR ARTIFICIAL INTELLIGENCE SYSTEMS

Abstract: *This study quantifies institutional heterogeneity in digital engagement and academic outcomes using anonymized data from Kazakhstani schools, comprising records from two EdTech platforms (Twig-bilim.kz, BilimLand.kz) and entry diagnostic test scores. Applying one-way ANOVA/intraclass correlation (ICC) decomposition and random-intercept mixed-effects modeling, we find that school membership explains 8.8-9.8% of variance in digital platform engagement and 30.0-30.7% of variance in entry test scores, with close numerical agreement between the ANOVA-based variance shares and the mixed-effects ICC estimates. These estimates indicate non-negligible school-level heterogeneity, particularly for entry diagnostic-test scores. The findings indicate that pooled, school-agnostic machine learning models risk systematically miscalibrated predictions for students in schools that deviate from the national average, and argue for school-aware calibration in the design of AI-based early-warning and adaptive-learning systems intended to support flexible and equitable access to education. The results are discussed in relation to Kazakhstan's ongoing school digitalization initiatives and to the broader literature on algorithmic fairness in education.*

Key words: *artificial intelligence, learning analytics, institutional heterogeneity, intraclass correlation, multilevel modeling, algorithmic fairness, educational data mining.*

Introduction

The development of artificial intelligence (AI) in education – from early warning systems to adaptive personalized-learning platforms – is increasingly regarded as an important means of improving the flexibility and accessibility of education. Recent systematic reviews indicate that generative AI, intelligent tutoring systems, and learning analytics are increasingly positioned as technologies capable of transforming personalized learning from a static, uniform model into an adaptive, data-driven process that supports dynamic learning pathways [1].

©K.E. Iklassova, L.N. Abdullah, M.Zh. Bazarova, 2026