

Двадцать третья международная конференция  
«СОВРЕМЕННЫЕ ПРОБЛЕМЫ ДИСТАНЦИОННОГО  
ЗОНДИРОВАНИЯ ЗЕМЛИ ИЗ КОСМОСА»,  
10 - 14 ноября 2025 г., Москва, Россия



СЕКЦИЯ

Технологии и методы использования спутниковых данных в системах мониторинга

# **Большие и малые геодатасеты машинного обучения открытого доступа для линейных объектов инфраструктуры**

Подольская Е. С., к. т. н., ведущий научный сотрудник  
Лаборатории мониторинга лесных экосистем,  
Центр по проблемам экологии и продуктивности лесов им. А. С. Исаева Российской академии наук  
(ЦЭПЛ РАН), Москва, Россия

# Датасеты машинного обучения

Набор данных (dataset) — совокупность данных, в том числе соответствующих им метаданных, организованных по определенным правилам и принципам описания.

**ГОСТ Р 59898-2021 «Оценка качества систем искусственного интеллекта. Общие положения»**

<https://docs.cntd.ru/document/1200181913>

## Open Datasets and Tools: An overview for Hugging Face

Community Article

Published August 27, 2025

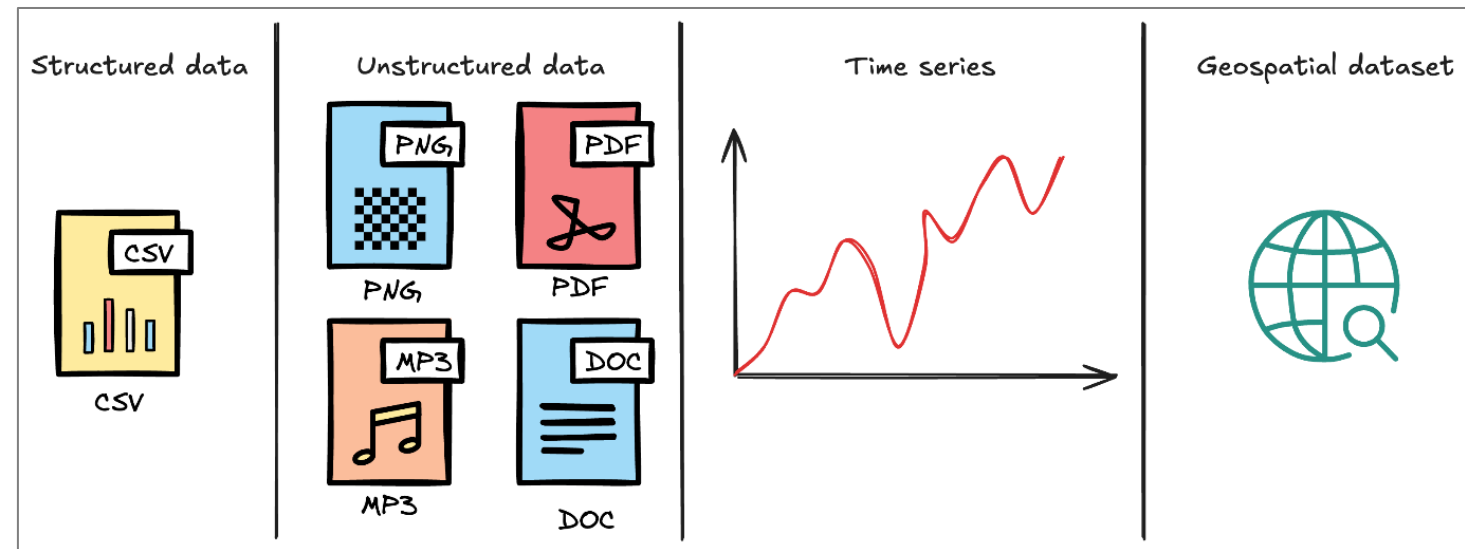


Avi Chawla

avi81

Follow

### 4 типа датасетов



<https://cdn-uploads.huggingface.co/production/uploads/62ff8c02a80e701897ac0331/3h3Hu3Iyy9dhAZr04Gg7y.png>

- Наряду с технологиями (развитием и доступностью платформ и инструментов, которые позволяют использовать искусственный интеллект в производственных процессах), датасеты продолжают быть важной частью проектов по машинному обучению.
- Количество и качество записей об объектах в датасете определенно влияет на качество результатов проектов машинного обучения, **НО** создание объемного качественного датасета является време- и трудозатратным процессом.
- На современном уровне развития технологий машинного обучения большой размер датасета перестает быть определяющим фактором результативного проекта по машинному обучению. ***В работе представлен анализ актуальных представлений о больших и малых наборах данных с выделением особенностей геодатасетов по объектам инфраструктуры линейного протяжения.*** Для цели анализа использовались материалы социальной исследовательской сети ResearchGate (<https://www.researchgate.net/>), а также собственные разработки датасетов лесотранспортной инфраструктуры.
- Мировой опыт представлен анализом датасетов открытого доступа по дорожной сети на территорию ряда стран (Великобритания, Япония, Индия, Чехия, Норвегия, США, Китай), опубликованным на разных платформах в 2021-2025 гг. и доступным по лицензии CC BY 4.0 (Подольская, 2025).

## How to Do Machine Learning with Small Data? – A Review from an Industrial Perspective

Ivan Kraljevski, Yong Chul Ju, Dmitrij Ivanov, Constanze Tschöpe, and Matthias Wolff

**Abstract**—Artificial intelligence experienced a technological breakthrough in science, industry, and everyday life in the recent few decades. The advancements can be credited to the ever-increasing availability and miniaturization of computational resources that resulted in exponential data growth. However, because of the insufficient amount of data in some cases, employing machine learning in solving complex tasks is not straightforward or even possible. As a result, machine learning with small data experiences rising importance in data science and application in several fields. The authors focus on interpreting the general term of “small data” and their engineering and industrial application role. They give a brief overview of the most important industrial applications of machine learning and small data. Small data is defined in terms of various characteristics compared to big data, and a machine learning formalism was introduced. Five critical challenges of machine learning with small data in industrial applications are presented: unlabeled data, imbalanced data, missing data, insufficient data, and rare events. Based on those definitions, an overview of the considerations in domain representation and data acquisition is given along with a taxonomy of machine learning approaches in the context of small data.

**Index Terms**—machine learning, small data, industrial applications, engineering applications

domain-specific data samples, and the  $P$  is a measure of success achieved in a particular task  $T$ . A less formal definition is given in [7], machine learning (ML) “is a sub-field of computer science, but is often also referred to as predictive analytics, or predictive modeling”. Its goal and usage is to build new and/or leverage existing *algorithms* to *learn* from data, in order to *build generalizable models* that give *accurate predictions*, or to find *patterns*, particularly with *new and unseen similar data*”.

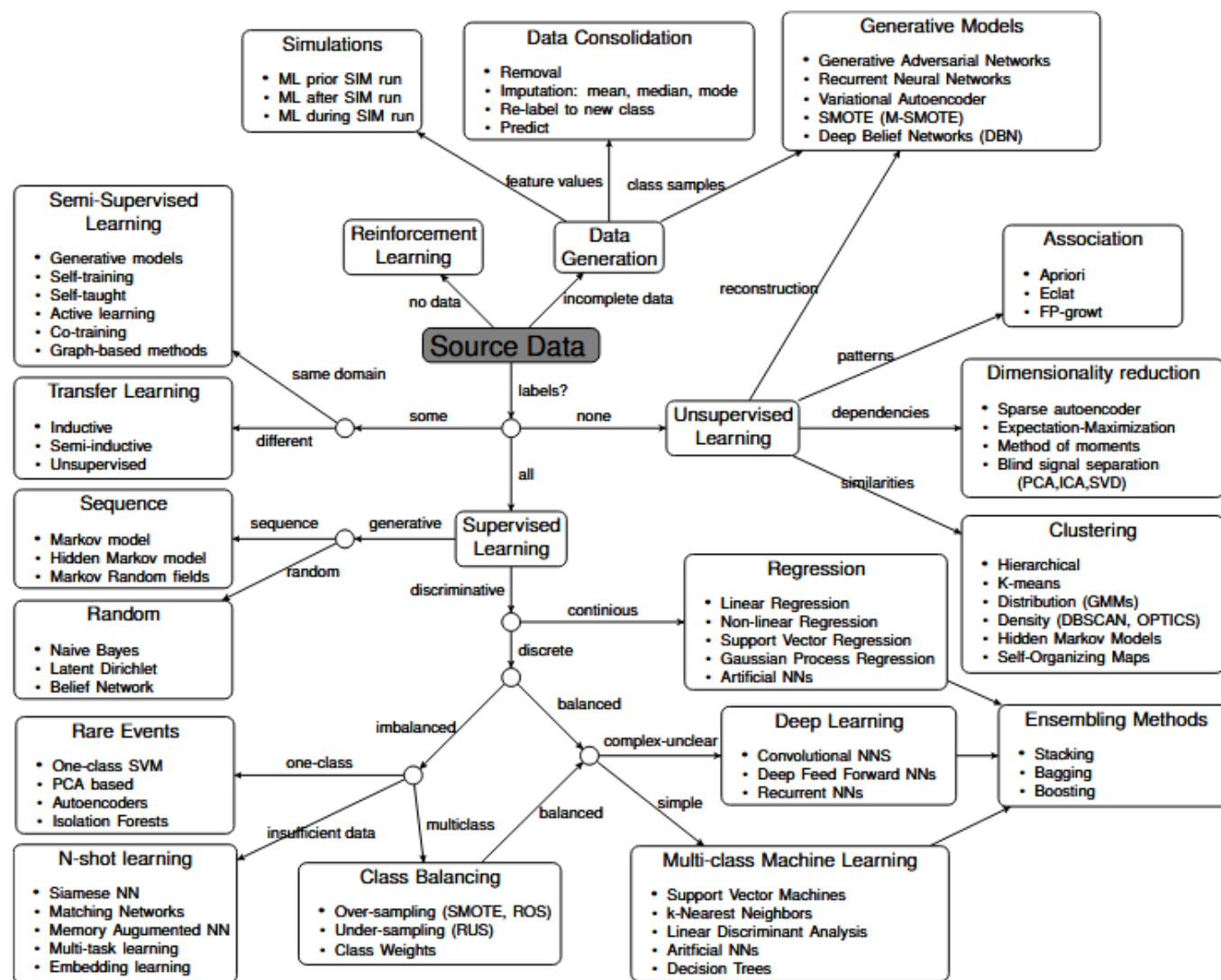
In the last two decades, considerable progress was made in ML-science leveraging the growing computational power [4], [8], [9] and the vast amounts of data increasingly being available (big data), see [10]. Consequently, data-driven ML methods are applied to social, economic, industrial, environmental, and even political tasks [11]. A large variety of comprehensive ML publications appeared, e.g., [1], [4], [5] with references therein related to one such as search engines, machine translation, online retailers, social networking services, financial modeling, marketing, education, and policy-making. Further detailed understanding of the underlying mechanisms of ML methods was gained as in statistical learning theory

[arXiv:2311.07126v1 \[cs.LG\] 13 Nov 2023](https://arxiv.org/abs/2311.07126v1)

- Small data is defined in terms of various characteristics compared to big data, and a machine learning formalism was introduced.

- Five critical challenges of machine learning with small data in industrial applications are presented: unlabeled data, imbalanced data, missing data, insufficient data, and rare events.

## Machine learning algorithms in context of small data





## Making the most of small Software Engineering datasets with modern machine learning

Julian Aron Prenner, Romain Robbes

**Abstract**—This paper provides a starting point for Software Engineering (SE) researchers and practitioners faced with the problem of training machine learning models on small datasets. Due to the high costs associated with labeling data, in Software Engineering, there exist many small (< 1 000 samples) and medium-sized (< 100 000 samples) datasets. While deep learning has set the state of the art in many machine learning tasks, it is only recently that it has proven effective on small-sized datasets, primarily thanks to pre-training, a semi-supervised learning technique that leverages abundant unlabelled data alongside scarce labelled data. In this work, we evaluate pre-trained Transformer models on a selection of 13 smaller datasets from the SE literature, covering both, source code and natural language. Our results suggest that pre-trained Transformers are competitive and in some cases superior to previous models, especially for tasks involving natural language; whereas for source code tasks, in particular for very small datasets, traditional machine learning methods often has the edge.

In addition, we experiment with several techniques that ought to aid training on small datasets, including active learning, data augmentation, soft labels, self-training and intermediate-task fine-tuning, and issue recommendations on when they are effective. We also release all the data, scripts, and most importantly pre-trained models for the community to reuse on their own datasets.

**Index Terms**—Small Datasets, Transformer, BERT, RoBERTA, Pre-training, Fine-Tuning, Data Augmentation, Back Translation, Soft Labels, Active Learning.

*arXiv:2106.15209v1 [cs.SE] 29 Jun 2021*



Editorial

## Machine Learning Methods with Noisy, Incomplete or Small Datasets

Cesar F. Caiafa <sup>1,\*</sup>, Zhe Sun <sup>2</sup>, Toshihisa Tanaka <sup>3</sup>, Pere Marti-Puig <sup>4</sup> and Jordi Solé-Casals <sup>4,\*</sup>

<sup>1</sup> Instituto Argentino de Radioastronomía—CCT La Plata, CONICET/CIC-PBA/UNLP, 1894 V. Elisa, Argentina

<sup>2</sup> Computational Engineering Applications Unit, Head Office for Information Systems and Cybersecurity, RIKEN, Wako-Shi 351-0198, Japan; zhe.sun.vk@riken.jp

<sup>3</sup> Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan; tanakt@cc.tuat.ac.jp

<sup>4</sup> Data and Signal Processing Research Group, University of Vic-Central University of Catalonia, 08500 Barcelona, Spain; pere.marti@uvic.cat

\* Correspondence: ccaiafa@gmail.com (C.F.C.); jordi.sole@uvic.cat (J.S.-C.)

**Abstract:** In this article, we present a collection of fifteen novel contributions on machine learning methods with low-quality or imperfect datasets, which were accepted for publication in the special issue “Machine Learning Methods with Noisy, Incomplete or Small Datasets”, Applied Sciences (ISSN 2076-3417). These papers provide a variety of novel approaches to real-world machine learning problems where available datasets suffer from imperfections such as missing values, noise or artefacts. Contributions in applied sciences include medical applications, epidemic management tools, methodological work, and industrial applications, among others. We believe that this special issue will bring new ideas for solving this challenging problem, and will provide clear examples of application in real-world scenarios.

**Keywords:** artificial intelligence; imperfect dataset; imperfect dataset; machine learning

*Caiafa, C.F.; Sun, Z.; Tanaka, T.; Marti-Puig, P.; Solé-Casals, J. Machine Learning Methods with Noisy, Incomplete or Small Datasets. Appl. Sci. 2021, 11, 4132. <https://doi.org/10.3390/app11094132>*

Tutorial

WSDM '20, February 3–7, 2020, Houston, TX, USA

## Learning with Small Data

Zhenhui Li  
Pennsylvania State University  
State College, Pennsylvania  
JessieLi@psu.edu

Huaxiu Yao  
Pennsylvania State University  
State College, Pennsylvania  
huaxiuyao@psu.edu

Fenglong Ma  
Pennsylvania State University  
State College, Pennsylvania  
fenglong@psu.edu

**ABSTRACT**

In the era of big data, it is easy for us collect a huge number of image and text data. However, we frequently face the real-world problems with only small (labeled) data in some domains, such as healthcare and urban computing. The challenge is how to make machine learn algorithms still work well with small data? To solve this challenge, in this tutorial, we will cover the state-of-the-art machine learning techniques to handle small data issue. In particular, we focus on the following three aspects: (1) Providing a comprehensive review of recent advances in exploring the power of knowledge transfer, especially focusing on meta-learning; (2) introducing the cutting-edge techniques of incorporating human/expert knowledge into machine learning models; and (3) identifying the open challenges to data augmentation techniques, such as generative adversarial networks. We believe this is an emerging and potentially high-impact topic in computational data science, which will attract both researchers and practitioners from academia and industry.

advanced material (50% for beginners and 50% for intermediate and advanced).

**2 PREVIOUS OFFERING OF THE TUTORIAL**

The topic of this tutorial is relatively new to machine learning and data mining area, and this is the first tutorial to systematically focus on small data. We have offered this tutorial internally at Penn State (<https://sites.psu.edu/daml/>) and have drawn a full class of students and have received very positive feedback.

**3 PRESENTERS**

**3.1 Contacts of the Presenters**

**Zhenhui (Jessie) Li** (main contact), Email: [JessieLi@psu.edu](mailto:JessieLi@psu.edu), Tel: 814-863-6317, Website: <https://faculty.ist.psu.edu/jessieli/>

**Huaxiu Yao**, Email: [huaxiuyao@psu.edu](mailto:huaxiuyao@psu.edu), Tel: 814-862-8330, Website: <http://huaxiuyao.mystrikingly.com/>

<https://dl.acm.org/doi/pdf/10.1145/3336191.3371874>

Practical machine learning is learning on small samples

Marina Sapir

**Abstract**

Based on limited observations, machine learning discerns a dependence which is expected to hold in the future. What makes it possible? Statistical learning theory imagines indefinitely increasing training sample to justify its approach. In reality, there is no infinite time or even infinite general population for learning.


Here I argue that practical machine learning is based on an implicit assumption that underlying dependence is relatively “smooth” : likely, there are no abrupt differences in feedback between cases with close data points.

From this point of view learning shall involve selection of the hypothesis “smoothly” approximating the training set. I formalize this as Practical learning paradigm. The paradigm includes terminology and rules for description of learners. Popular learners (local smoothing, k-NN, decision trees, Naive Bayes, SVM for classification and for regression) are shown here to be implementations of this paradigm.


Review

SCIENCE PROGRESS

Science Progress  
2022, Vol. 105(1) 1–16  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/00368504211029777  
journals.sagepub.com/home/sci



## Machine learning on small size samples: A synthetic knowledge synthesis

Peter Kokol<sup>1</sup> , Marko Kokol<sup>2</sup> and Sašo Zagoranski<sup>2</sup>

<sup>1</sup> Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia  
<sup>2</sup> Semantika, Maribor, Slovenia

### Abstract

Machine Learning is an increasingly important technology dealing with the growing complexity of the digitalised world. Despite the fact, that we live in a ‘Big data’ world where, almost ‘everything’ is digitally stored, there are many real-world situations, where researchers are still faced with small data samples. The present bibliometric knowledge synthesis study aims to answer the research question ‘What is the small data problem in machine learning and how it is solved?’ The analysis a positive trend in the number of research publications and substantial growth of the research community, indicating that the research field is reaching maturity. Most productive countries are China, United States and United Kingdom. Despite notable international cooperation, the regional concentration of research literature production in economically more developed countries was observed. Thematic analysis identified four research themes. The themes are concerned with to dimension reduction in complex big data analysis, data augmentation techniques in deep learning, data mining and statistical learning on small datasets.

### Keywords

Machine learning, small data sets, knowledge synthesis, bibliometrics

*В чем состоит проблема малых данных в машинном обучении и как она может быть решена?*

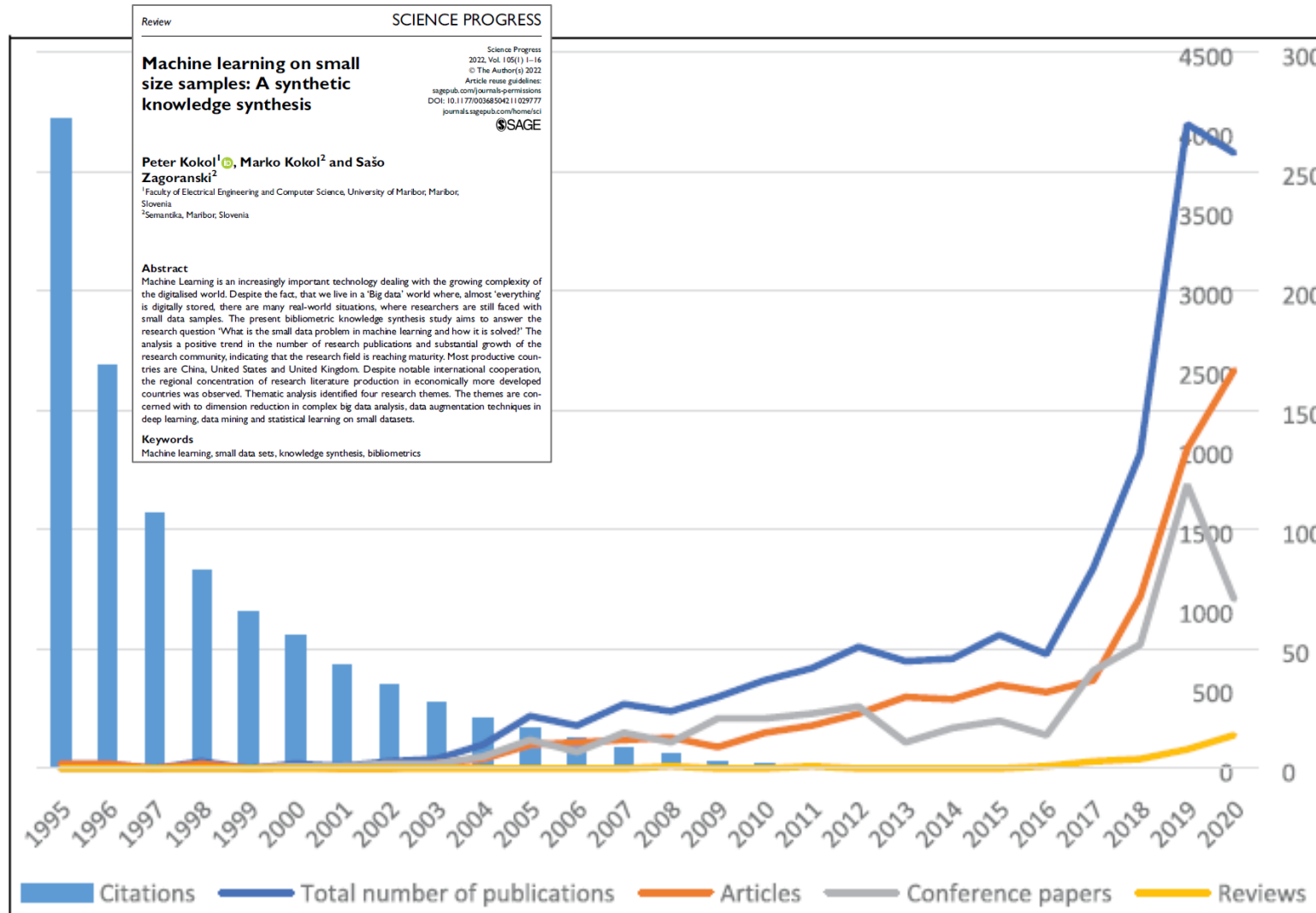
Preprint · January 2025  
DOI: 10.48550/arXiv.2501.01836

<https://doi.org/10.1177/00368504211029777>

Большие и малые геодатасеты машинного обучения открытого доступа для линейных объектов инфраструктуры

6

## Research literature production dynamics



## Будущее Больших данных – это малые датасеты

- The most frequently reported difficulties causing the small data problem are the small size of the dataset, high/low dimensionality of datasets and unbalanced data.
- Our analysis also revealed that small datasets may result from hardware requirements to cope with limited processing power or small storage size of devices like for example Raspberry PI.
- Some research also shows that high quality small sample can be better than a low quality large sample in the case of statistical machine learning

<https://doi.org/10.1177/00368504211029777>



## Less (Data) Is More: Why Small Data Holds the Key to the Future of Artificial Intelligence

Ciro Greco\*, Andrea Polonioli\* and Jacopo Tagliabue\*

*Tooso Labs,  
San Francisco, CA, U.S.A.*

**Keywords:** Artificial Intelligence, Big Data, Probabilistic Programming, Concept Learning, Machine Learning.

**Abstract:** The claims that big data holds the key to enterprise successes and that Artificial Intelligence (AI) is going to replace humanity have become increasingly more popular over the past few years, both in academia and in the industry. However, while these claims may indeed capture some truth, they have also been massively oversold, or so we contend here. The goal of this paper is two-fold. First, we provide a qualified defence of the value of less data within the context of AI. This is done by carefully reviewing two distinct problems for big data driven AI, namely a) the limited track record of Deep Learning (DL) in key areas such as Natural Language Processing (NLP), b) the regulatory and business significance of being able to learn from few data points. Second, we briefly sketch what we refer to as a case of “A.I. with humans and for humans”, namely an AI paradigm whereby the systems we build are privacy-oriented and focused on human-machine collaboration, not competition. Combining our claims above, we conclude that when seen through the lens of cognitively inspired A.I., the bright future of the discipline is about less data, not more, and more humans, not fewer.

<https://doi.org/10.5220/0007956203400347>

Greco, C., Polonioli, A. and Tagliabue, J. (2019). Less (Data) Is More: Why Small Data Holds the Key to the Future of Artificial Intelligence. In *Proceedings of the 8th International Conference on Data Science, Technology and Applications - DATA*; ISBN 978-989-758-377-3; ISSN 2184-285X, SciTePress, pages 340-347. DOI: 10.5220/0007956203400347



## Каковы основные проблемы малых датасетов в публикациях?

### Темы исследований малых датасетов


Theme	Colour	More frequent codes	Prevailing sub-categories
Dimension reduction in complex big data analysis	Red	Machine learning (339), Classification (51), Feature selection (50), Artificial neural network (28), Neural network (19), Natural language processing (16), Ensemble learning (14), Bioinformatics (13), Decision tree (12), Breast cancer (11), Machine learning algorithms (10), Computer vision (10), Microarray (9), Small datasets (9), Text classification (9), Big data (9), Dimensionality reduction (9), Social media (7),	Machine learning algorithms in bioinformatics, Classification on small samples with feature selection, Ensemble learning for solving under-sampling in personalised medicine, Solving missing data in breast cancer using deep neural networks, Dimensionality reduction, with feature extraction in cancer classification, Natural language processing in social media and text analysis
Data augmentation techniques in deep learning for pattern recognition and classification	Blue and yellow	Deep learning (89), Convolutional neural networks (45), Transfer learning (38), Pattern recognition (14), Image classification (14), data augmentation (12), Generative adversarial networks (8), LSTM (6), Resnet – Residual neural network (n=4)	Deep and transfer learning using in image classification using data augmentation and synthetic data, complex neural networks, Pattern recognition in mobile sensing using Principal component analysis and LSTM, Resnet and Desnet based deep learning using auto encoder on case of glaucoma,
Support vector machines, Random forest and Genetic algorithms in statistical learning	Green	Support vector machines (109), Random forest (24), Prediction (18), Fault diagnosing (18), Forecast (10), Genetic algorithms (8), Regression (8), Gaussian process (7), Optimisation (7), Time series (6), Statistical learning theory (6)	Support vector machines and random forests in prediction, forecasting and fault diagnosis, Genetic algorithms in optimisation and time series analysis
Data mining on small datasets with support vectors regression	Violet	Small datasets (2), Data mining (15), Support vector regression (11), Virtual sample (7)	Small datasets augmentation with virtual samples, Data mining with support vector regression
Numbers in parenthesis present the number of papers in which an author keyword occurred			

Theme	Small data size problem	Machine learning algorithms	Small – data pre-processing technique	Application area
Dimension reduction in complex big data analysis	Missing data (3), Unbalanced data (8), Under sampling (3)	Bagging (9), Bayesian networks (7), Boosting (6), Decision trees and forests (43), deep neural network (12), Ensemble learning (21), manifold learning (5), neural network (30), naive Bayes (2), semi supervised learning (17), supervised learning (19), unsupervised learning (4), FMRI (7)	Dimensionality reduction (9), Feature extraction (12), engineering (4), fusion (2), Lasso (6), Linear and multiple discriminant analysis (13), PCA (16), autoencoder (10)	Image analysis (18), genetics and bioinformatics (27), personalised and precision medicine (6) neuroimaging (7), mental health (25)
Data augmentation for deep learning in pattern recognition and classification		Convolutional neural network (42), Dense CNN (3), Generative adversarial networks (8), LSTM (6), Resnet (4), transfer learning (40), semi-supervised learning (17)	Data augmentation (12), Domain adaption (5), data fusion (3), synthetic data (3), autoencoder (10)	Image classification (22), genetics (3), mobile sensing (10), predictive modelling (6), semantic segmentation (3)
Statistical based machine learning	Small sample (98)	Support vector machines (89), Least square SVM (10), genetic algorithms, random forest (24), Particle swarm (27), support vector regression (11)	Feature engineering (3), meta-learning (5), Monte Carlo (3), Small sample learning (5), virtual sample (11)	Forecasting (13), Fault diagnosing (18), regression (19), prediction (26), pattern recognition (17), rock mechanics (5), security (7), computer vision (10), data mining (15)
Data mining on small datasets	Big data (9)			Text and social media analysis (21), Precision medicine (3), Natural language processing (16), Sentiment analysis (9), clustering (8)
Numbers in parenthesis present the number of publications.				



Greco, C., Polonioli, A. and Tagliabue, J. (2019). Less (Data) Is More: Why Small Data Holds the Key to the Future of Artificial Intelligence. In *Proceedings of the 8th International Conference on Data Science, Technology and Applications - DATA*; ISBN 978-989-758-377-3; ISSN 2184-285X, SciTePress, pages 340-347. DOI: 10.5220/0007956203400347

# Датасеты машинного обучения для распознавания дорог

- Сбор датасетов по распознаванию дорог был проведен по материалам социальной исследовательской сети ResearchGate (<https://www.researchgate.net/>): сделаны запросы на русском («датасеты машинного обучения для распознавания дорог») и английском (“machine learning datasets road recognition”) языках.
- Согласно современным представлениям о размере датасета, собранные примеры можно отнести к большим.



XII Международная научная конференция  
 «Региональные проблемы дистанционного зондирования Земли»  
 9 сентября - 12 сентября 2025 г., Красноярск, Россия

## ДАТАСЕТЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАСПОЗНАВАНИЯ ДОРОГ

Секция «Технологии искусственного интеллекта и Big Data»  
 Аудитория УЛК-112. Председатель Гулютин Н.Н.

Подольская Екатерина Сергеевна, к. т. н.,  
 ведущий научный сотрудник Лаборатории мониторинга лесных экосистем ЦЭПЛ РАН,  
 доцент Факультета географии и геоинформационных технологий НИУ ВШЭ  
[ekaterina.podolskaia@gmail.com](mailto:ekaterina.podolskaia@gmail.com)  
[epodolskaya@hse.ru](mailto:epodolskaya@hse.ru)

[https://rprs.sfu-kras.ru/sites/default/files/nauchnaya\\_programma\\_koferencii\\_r\\_p\\_dzz\\_2025\\_3.pdf](https://rprs.sfu-kras.ru/sites/default/files/nauchnaya_programma_koferencii_r_p_dzz_2025_3.pdf)

Оригинальное название и год публикации	Количество изображений	География данных	Ссылка
Oxford Road Boundaries (2021)	62605	Великобритания (Oxford)	Датасет: <a href="https://oxford-robotics-institute.github.io/road-boundaries-dataset/">https://oxford-robotics-institute.github.io/road-boundaries-dataset/</a> ; статья: [11]
RDD2022 - The multi-national Road Damage Dataset (2022)	47420	Япония, Индия, Чехия, Норвегия, США, Китай	Датасет: <a href="https://figshare.com/articles/dataset/RDD2022_-_The_multi-national_Road_Damage_Dataset_released_through_CRDDC_2022/21431547/1">https://figshare.com/articles/dataset/RDD2022_-_The_multi-national_Road_Damage_Dataset_released_through_CRDDC_2022/21431547/1</a> ; статья: [12]
Dataset of Unpaved and Paved Road Surface with Seasons (2022)	8484	Индия	Датасет: <a href="https://data.mendeley.com/datasets/tj2m7zz4rg/2">https://data.mendeley.com/datasets/tj2m7zz4rg/2</a> ; статья: [13]
Road Surface Reconstruction Dataset (2024)	16000	Китай (Beijing, Qingdao)	Датасет: <a href="https://github.com/ztsrxh/RSRD_dev_toolkit">https://github.com/ztsrxh/RSRD_dev_toolkit</a> ; статья: [14]
HighRPD: A high-altitude drone dataset of road pavement distress (2025)	11696	Китай (Shanxi)	Датасет: <a href="https://data.mendeley.com/datasets/sywswj7dji/1">https://data.mendeley.com/datasets/sywswj7dji/1</a> ; статья: [15]

Изученные датасеты доступны по лицензии CC BY 4.0 (<http://creativecommons.org/licenses/by-nc-sa/4.0>) с обязательным указанием разработчика



# Датасеты инфраструктуры линейных объектов (1)



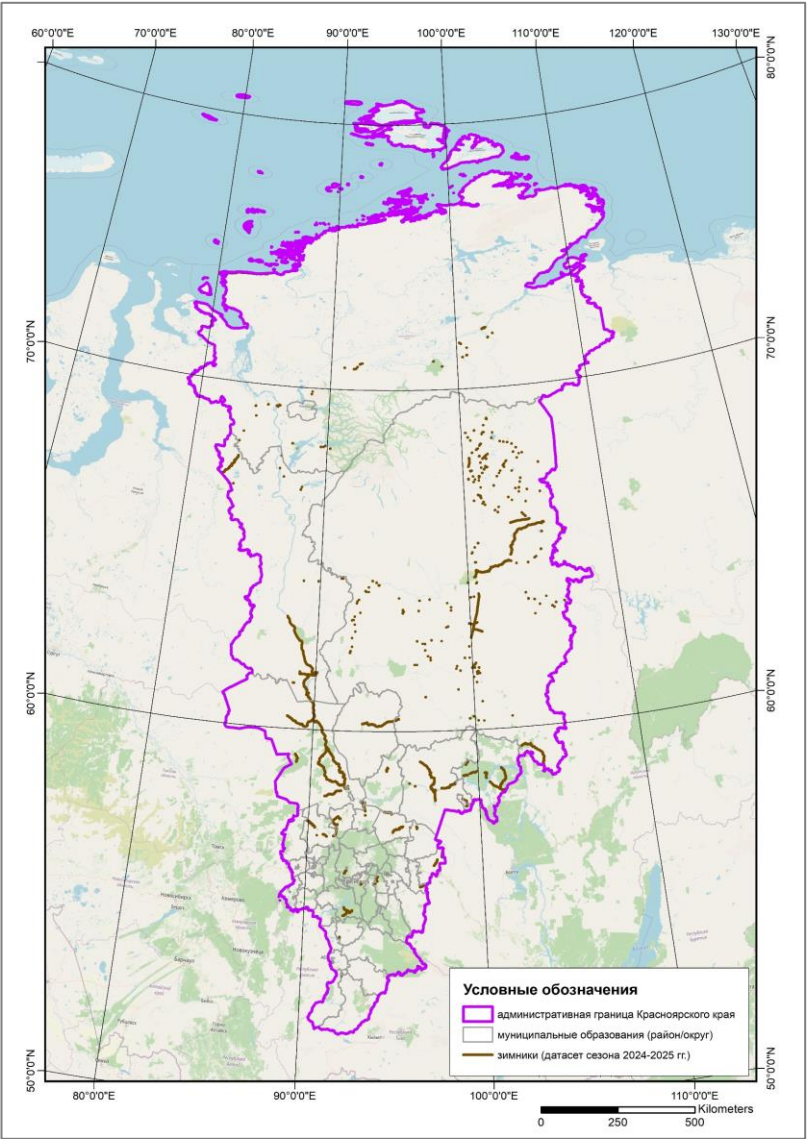
## Датасет зимников Красноярского края: сбор данных из разных источников

По состоянию на: 29.04.2025

Интервал функционирования зимников в сезоне 2024-2025: ноябрь- апрель  
490 записей

№	Муниципальное образование район	Краевые и региональные автодороги общего пользования(км)	Иные автодороги(км)			Район	Название Дорога	Дата открытия	Протяженность (км)	Тоннаж	Дата закрытия
			Дороги с грунтовыми покрытием	Неклассифицированные дороги	Зимники						
1	Абанский муниципальный район	148,891	0	0	8,655	1	Абанский	Дорога Подъезд к п. Тулень (зимник)	-	8,7	05 февраля 2025
2	Ачинский муниципальный район	33,018	23,919	23,919	0	2	Абанский	Дорога Покатеево - Хиндичет (зимник)	-	0,29	28 марта 2017
3	Балахтинский муниципальный район	68,088	1,039	1,039	94,266	3	Балахтинский	Дорога "Черемушки - Тюлюпта" - Березовая	18 декабря 2024	27,1	-
4	Березовский муниципальный район	62,445	30,432	30,432	0	4	Балахтинский	Дорога Черемушки - Тюлюпта (зимник)	18 декабря 2024	46,59	-
5	Бирюсский муниципальный район	15,1	3,09	3,09	73,158	5	Бирюсский	Дорога Маталассы - Нивкифорвка (зимник)	-	-	-
6	Боготольский муниципальный район	32,889	14,51	14,51	0	6	Бирюсский	Дорога Мендельский - Малая Кеть (зимник)	-	-	-
7	Богучанский муниципальный район	438,409	67,515	67,515	320,324	7	Бирюсский	Дорога Промбор - Проточное (зимник)	-	-	-
8	Большемуртинский муниципальный район	70,131	8,378	8,378	0	8	Богучанский	Дорога Ангарский-Беляки (зимник)	09 декабря 2020	-	24 октября 2024
9	Большелудинский муниципальный район	67,415	0	0	0	9	Богучанский	Дорога Ангарский-Беляки (зимник)	09 декабря 2020	-	-
10	город Ачинск	41,804	0,521	0,521	0	10	Богучанский	Дорога Обход Богучан	11 декабря 2017	-	-
11	город Боготол	9,895	0,227	0,227	0						
12	город Бородино	8,604	0,535	0,535	0						
13	город Дивногорск	7,829	9,222	9,222	0						
14	город Енисейск	14,092	0,996	0,996	7,511						
15	город Канск	20,038	0,399	0,399	0						
16	город Красноярск	287,68	140,761	140,761	0						
17	город Лесосибирск	36	7,365	7,365	3,016						
18	город Минусинск	35,864	2,623	2,623	0						
19	город Назарово	13,643	3,016	3,016	0						
20	город Норильск	103,077	19,732	19,732	0						
21	город Шарыпово	13,582	2,43	2,43	0						
22	Дзержинский муниципальный район	58,771	0	0	0						
23	Емельяновский муниципальный район	95,086	37,009	37,009	0						
24	Енисейский муниципальный район	238,209	277,688	600,245	1192,235						
25	Ермаковский муниципальный район	9,085	11,119	11,119	0						
26	ЗАТО город Железногорск	43,78	6,892	6,892	0						
27	ЗАТО город Зеленогорск	21,943	46,18	46,18	0						

Район	Название Дорога	Тип переправы	Дата открытия	Грузоподъемность (тонн)	Дата закрытия
Бирюсский	Дорога Бирюссы-Биктимиров юсский ка (1+451)	Ледовая	24 февраля 2025	20	-
Бирюсский	Дорога Бирюссы-Биктимиров юсский ка (1+451)	Понтонная	-	-	24 октября 2024
Бирюсский	Дорога Бирюссы-Биктимиров юсский ка (1+451)	Паромная	-	-	18 июля 2024
Бирюсский	Дорога Шпагино 2-Подкаменка юсский (28+400)	Ледовая	10 января 2025	10	-



<https://miigaik.ru/scidata/section2/#table-section>

# Датасеты инфраструктуры линейных объектов (2)

## Датасет просек под линиями электропередач (ЛЭП):

## распознавание по снимкам Sentinel при помощи нейросети YOLO

## 2446 снимков с учетом аугментации

### Карта района исследования и расположения тестовых участков

Современные проблемы дистанционного зондирования Земли из космоса. 2025. Т. 22. № 4. С. 11–26

#### Автоматическое детектирование просек под линиями электропередач на снимках Sentinel-2 с помощью методов машинного обучения и компьютерного зрения

Я. О. Бахрамхан<sup>1</sup>, Д. М. Ермаков<sup>2,3</sup>, Е. С. Подольская<sup>1,4</sup>

<sup>1</sup> Национальный исследовательский университет «Высшая школа экономики»  
Москва, 109028, Россия

E-mails: ybahramhan@gmail.com, podols\_kate@mail.ru

<sup>2</sup> Институт космических исследований РАН, Москва, 117997, Россия  
E-mail: d.m.ermakov@cosmos.ru

<sup>3</sup> Фрязинский филиал Института радиотехники и электроники  
им. В. А. Котельникова РАН, Фрязино, Московская обл., 141190, Россия

<sup>4</sup> Центр по проблемам экологии и продуктивности лесов им. А. С. Исаева РАН  
Москва, 117997, Россия

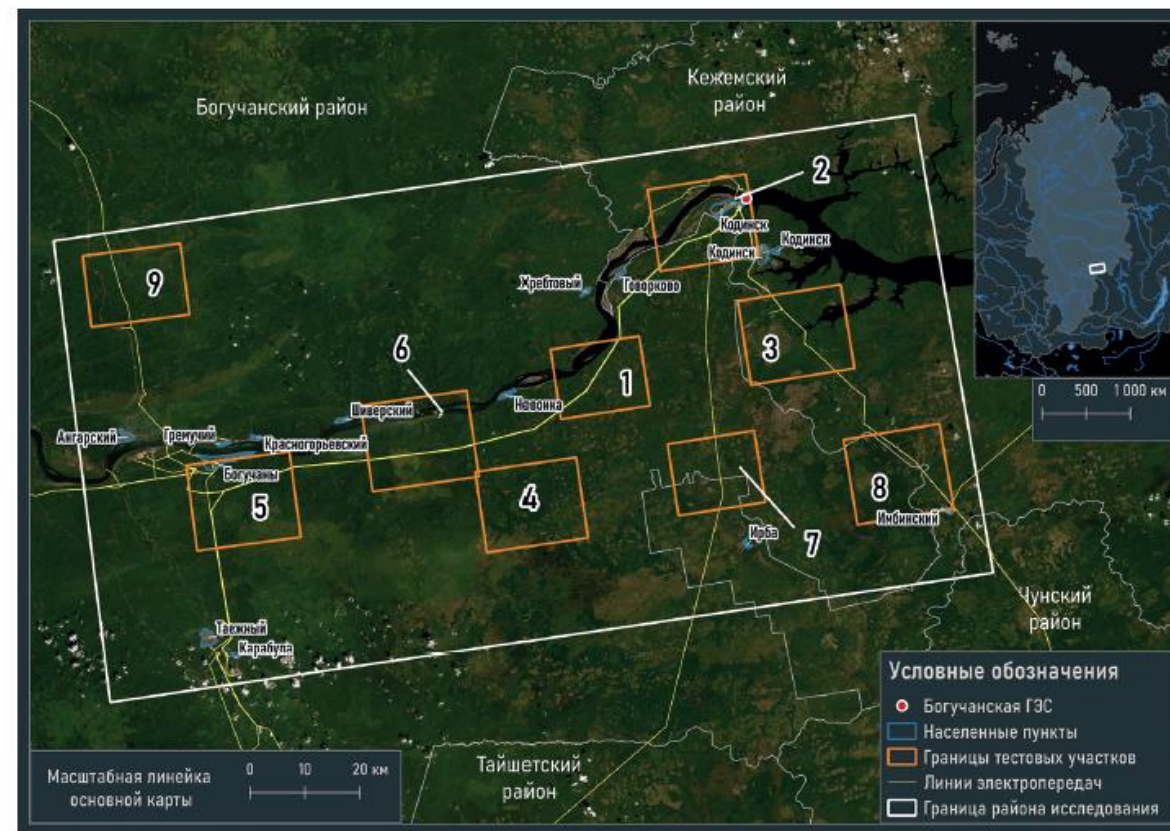
Регулярный мониторинг состояния линий электропередач (ЛЭП) необходим для бесперебойного электроснабжения населённых пунктов и объектов инфраструктуры. Однако на лесных территориях с низкой плотностью населения и дорог задача постоянного мониторинга усложняется. Для дистанционного мониторинга просек под ЛЭП в труднодоступных лесных территориях настоящим исследованием предлагается космических снимках Sentinel-2. Такой алгоритм автоматизации дистанционного мониторинга, детектирования изменений (прокладки, з. Логика работы алгоритма имеет наглядную инт. интереса в пространстве дешифрованных п. стик модель логистической регрессии выделяя секам, затем вероятностное преобразование X. маске этих участков. Разработанный алгоритм не требует индивидуальной настройки параметров созданы метрики качества выделения пр. алгоритм корректно определяет области распо. ных спутниковых изображениях в пределах рег.

**Ключевые слова:** просеки, ЛЭП, дистанционный интерпретируемость алгоритма, преобразование

Конференция молодых ученых «Фундаментальные и прикладные космические исследования»

#### Сегментация лесных просек под ЛЭП на снимках Sentinel-2 с помощью нейронной сети YOLO

Бахрамхан Я.О., НИУ ВШЭ  
Ермаков Д.М., ИКИ РАН, ФИРЭ им. В.А. Котельникова РАН  
Подольская Е.С., ЦЭПЛ РАН, НИУ ВШЭ



[https://kmu.cosmos.ru/sites/default/files/2025/presentations/Bakhramkhan\\_DZZ.pdf](https://kmu.cosmos.ru/sites/default/files/2025/presentations/Bakhramkhan_DZZ.pdf)

<http://jr.rse.cosmos.ru/article.aspx?id=3045>



# Выводы исследования

- Согласно современным научным работам, *размер малого датасета может варьироваться и определяться порядком в 100-1000* записей/объектов, нет строгого разделения на большой/малый.
- *Методическими ограничениями малого датасета является набор следующих признаков:* неаннотированные данные, отсутствие баланса в данных, пропущенные данные, редкие и недостаточные данные для определенного явления.
- Вместе с работами по необходимости создания больших датасетов, создаваемых и поддерживаемых ввиду трудоемкости и стоимости преимущественно большим компаниям, есть перспективы малых, но качественных датасетов (международный опыт в статьях 2019-2020-2025 гг.).
- Для территории России в открытом доступе есть недостаток данных по объектам инфраструктуры линейного протяжения (просеки, дороги); дорожной сети общего и специального (например, лесные), а также сезонного использования (зимники и летники), что может быть решено путем сбора, обработки и обновления разнородных данных исследовательскими/учебными группами в рамках региональных проектов.
- Полученные результаты для зимников и просек (Красноярский край) демонстрируют практическую значимость и могут быть использованы для поддержки принятия решений в ряде отраслей хозяйства региона и страны при условии поддержки их актуальности.

*Работа выполнена в рамках государственного задания ЦЭПЛ РАН по теме  
«Биоразнообразие и экосистемные функции лесов» (Регистрационный номер НИОКТР 124013000750-1)*



- Бахрамхан Я. О., Ермаков Д. М., Подольская Е. С. Автоматическое детектирование просек под линиями электропередач на снимках Sentinel-2 с помощью методов машинного обучения и компьютерного зрения // Современные проблемы дистанционного зондирования Земли из космоса. 2025. Т. 22. № 4. С. 11–26.  
<http://jr.rse.cosmos.ru/article.aspx?id=3045>
- Подольская Е. С. Датасеты машинного обучения для распознавания дорог // Региональные проблемы дистанционного зондирования Земли: материалы XII Международной научной конференции. Красноярск, 09–12 сентября 2025 г. / Сиб. федер. ун-т, Ин-т космич. и информ. технологий; Красноярск: СФУ, 2025. [https://rprs.sfu-kras.ru/sites/default/files/sbornik\\_2025.pdf](https://rprs.sfu-kras.ru/sites/default/files/sbornik_2025.pdf)
- Синицина А. Н., Подольская Е. С. Опыт создания датасета зимников для лесного хозяйства и инфраструктурных проектов Красноярского края // Пространственные данные (28-29 мая 2025 г.), Москва, Россия. М.: МИИГАиК, 2025. <https://miigaik.ru/scidata/section2/>