Вестник Череповецкого государственного университета, 2026, № 5 (134), с. 7–17.
Cherepovets State University Bulletin, 2026, no. 5 (134), pp. 7–17.

Научная статья
УДК 004.89; 004.032.26
https://doi.org/10.23859/1994-0637-2026-5-134-1
https://elibrary.ru/vwufau

Разработка методики формирования наборов данных для обучения нейросетевых систем автоматического анализа радиосигналов космических аппаратов

Алина Владимировна Кобенко[1]✉, Роман Александрович Алешко[2]
[1,2]Северный (Арктический) федеральный университет им. М. В. Ломоносова, Архангельск, Россия
[1]✉a.kobenko@narfu.ru, https://orcid.org/0000-0001-6857-1799
[2]r.aleshko@narfu.ru, https://orcid.org/0000-0002-1235-8937

Аннотация. В работе предложена методика синтеза верифицируемых наборов данных для обучения нейросетей задачам радиомониторинга. Особенностью подхода является использование изображений дистанционного зондирования Земли в качестве источника полезного сигнала, что сохраняет естественную энтропию радиообмена. Авторами реализован программный комплекс, моделирующий канал с аддитивным шумом и эффектом Доплера, а также экспериментально подтверждена корректность физической модели: для аналоговых сигналов достигнута высокая корреляция между исходным и восстановленным изображениями, для цифровых – совпадение вероятности ошибок с теорией.

Ключевые слова: радиосигналы, синтез данных, машинное обучение, нейронные сети, дистанционное зондирование Земли, модуляция, радиомониторинг

Благодарность. Исследование проводилось под руководством кандидата технических наук, доцента Р. А. Алешко.

Для цитирования: Кобенко А. В., Алешко Р. А. Разработка методики формирования наборов данных для обучения нейросетевых систем автоматического анализа радиосигналов космических аппаратов**.** *Вестник Череповецкого государственного университета*, 2026, № 5 (134), с. 7–17. https://doi.org/10.23859/1994-0637-2026-5-134-1; EDN: VWUFAU

© Кобенко А. В., Алешко Р. А., 2026

Development of a methodology for generating datasets for training neural network systems for automatic analysis of spacecraft radio signals

Alina V. Kobenko[1]✉, Roman A. Aleshko[2]

[1,2]Northern (Arctic) Federal University named after M. V. Lomonosov, Arkhangelsk, Russia

[1]✉a.kobenko@narfu.ru, https://orcid.org/0000-0001-6857-1799

[2]r.aleshko@narfu.ru, https://orcid.org/0000-0002-1235-8937

Abstract. The paper proposes a methodology for synthesizing verifiable datasets for training neural networks in radio monitoring tasks. A specific feature of the approach is the use of Earth remote sensing images as a source of the useful signal, which preserves the natural entropy of the radio exchange. The authors have implemented a software package that simulates a channel with additive noise and Doppler effect. The correctness of the physical model was experimentally confirmed: a high correlation between the initial and restored images was achieved for analog signals, and the error probability matched the theory for digital signals.

Keywords: radio signals, data synthesis, machine learning, neural networks, Earth remote sensing, modulation, radio monitoring

Acknowledgments. The research was conducted under the supervision of Candidate of Technical Sciences, Associate Professor R. A. Aleshko.

For citation: Kobenko A. V., Aleshko R. A. Development of a methodology for generating datasets for training neural network systems for automatic analysis of spacecraft radio signals. *Cherepovets State University Bulletin*, 2026, no. 5 (134), pp. 7–17. (In Russ.) https://doi.org/10.23859/1994-0637-2026-5-134-1; EDN: VWUFAU

Введение

Современные системы радиомониторинга все активнее используют методы машинного обучения[1] для автоматического анализа сигналов, эффективность которых напрямую зависит от объема и качества обучающих данных. Однако формирование больших, достоверно размеченных наборов данных сопряжено с дороговизной сбора реальных сигналов и ошибками ручной разметки[2].

[1] Abd-Elaziz O., Abdalla M., Elsayed R. Deep learning-based automatic modulation classification using robust CNN architecture for cognitive radio networks. *Sensors*, 2023, vol. 23, no. 23, pp. 9467. https://doi.org/10.3390/s23239467; Wang D., Lin M., Xiaoxu Z., et al. Automatic modulation classification based on CNN-transformer graph neural network. *Sensors*, 2023, vol. 23, no. 16, pp. 7281. https://doi.org/10.3390/s23167281; Snoap J. A., Popescu D. C., Latshaw J. A., Spooner C. M. Deep-learning-based classification of digitally modulated signals using capsule networks and cyclic cumulants. *Sensors*, 2023, vol. 23, iss. 12, pp. 5735. https://doi.org/10.3390/s23125735; Snoap J. A., Popescu D. C., Spooner C. M. Deep-learning-based classifier with custom feature-extraction layers for digitally modulated signals. *IEEE Transactions on Broadcasting*, 2024, vol. 70, no. 3, pp. 763–773. https://doi.org/10.1109/TBC.2024.3391056

[2] Scholl S. RF signal classification with synthetic training data and its real-world performance. *arXiv*, 2022. https://doi.org/10.48550/arXiv.2206.12967

Альтернативой является использование синтетических данных. Существующие подходы, такие как наборы данных серии *RadioML* (*RML2016*, *RML2018*)[1], имеют ряд ограничений: сигналы в них часто представляют собой шум или используют нереалистичные последовательности. Кроме того, обучение нейросетей на синтетических данных без тщательного моделирования физических искажений канала связи ведет к значительному падению точности при работе с реальным эфиром.

Целью настоящей работы является разработка и верификация методики синтеза радиосигналов на основе растровых изображений дистанционного зондирования Земли (далее – ДЗЗ), пригодных для обучения нейросетевых систем.

Для достижения цели решаются задачи разработки гибридной математической модели, программной реализации комплекса генерации и экспериментальной верификации качества получаемых данных.

Основная часть

В основу разработанной методики положен принцип гибридного моделирования, при котором полезный сигнал формируется не на основе генераторов псевдослучайных чисел (PRNG), а из реальных структурированных объектов – растровых изображений ДЗЗ. Такой подход обеспечивает сохранение естественной информационной энтропии и внутриструктурных корреляционных связей (закономерностей между соседними отсчетами данных, унаследованных от исходного изображения) в формируемом радиосигнале, что является критически важным для обучения нейросетей задачам семантической сегментации и восстановления контента.

Программная реализация комплекса выполнена на языке высокого уровня Python 3, выбор которого обусловлен наличием развитой экосистемы библиотек для цифровой обработки сигналов (DSP). Вычислительное ядро построено на базе библиотеки NumPy, обеспечивающей векторные операции с многомерными массивами типа complex64, что является стандартом представления сигналов в базовой полосе частот (baseband signals). Для реализации алгоритмов цифровой обработки (преобразование Гильберта для однополосной модуляции, оконное преобразование Фурье (STFT), фильтрация) использован пакет SciPy (модуль signal). Обработка исходных изображений, включая конвертацию в оттенки серого и изменение разрядности, осуществляется средствами библиотеки Pillow (PIL), а визуализация частотно-временных спектрограмм с помощью Matplotlib.

Архитектура системы генерации построена по конвейерному принципу и включает четыре функциональных блока: блок предварительной обработки данных, блок модуляции, модель канала распространения и блок регистрации результатов (см. рис. 1).

[1] Spooner C. *DeepSig's 2018 Dataset: 2018.01.OSC.0001_1024x2M.h5.tar.gz*, 2020. Available at: https://cyclostationary.blog/2020/09/24/deepsigs-2018-data-set-2018-01-osc-0001_1024x2m-h5-tar-gz/ (accessed: 16.01.2026); Spooner C. *More on DeepSig's RML Datasets*, 2020. Available at: https://cyclostationary.blog/2020/08/17/more-on-deepsigs-rml-data-sets/ (accessed: 16.01.2026).

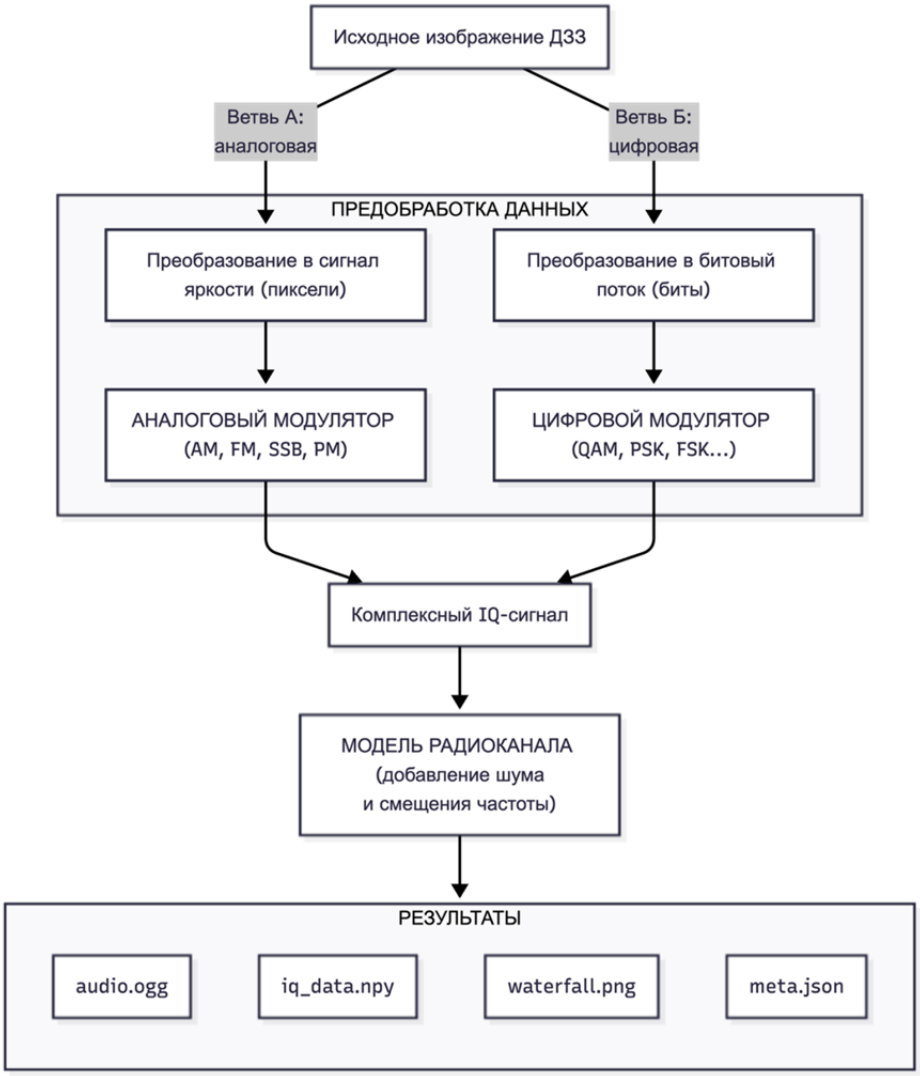


Рис. 1. Архитектура системы генерации синтетических радиосигналов

В зависимости от типа формируемого сигнала поток обработки разделяется на две независимые ветви.

В аналоговой ветви (AM, FM, SSB, PM) исходное изображение рассматривается как двумерный сигнал, подвергающийся строчному сканированию. Яркость пикселей $I(x, y)$ линейно отображается в мгновенные параметры несущего колебания. Предварительная обработка включает передискретизацию (изменение пространственного разрешения) изображения для согласования количества пикселей с часто-

той дискретизации F_S и длительностью сигнала T, а также нормировку динамического диапазона.

В цифровой ветви (PSK, QAM, FSK) реализован алгоритм адаптивного согласования информационного объема источника с пропускной способностью канала. Емкость канала C (в битах) рассчитывается исходя из заданной длительности записи T (4,0 с.), символьной скорости R_S (1200 Бод) и позиционности модуляции M по формуле: $C = \lfloor TR_2 \log_2 M \rfloor$.

На основе вычисленной емкости C производится адаптивная передискретизация исходного изображения с использованием фильтра Ланцоша для сохранения детализации, после чего изображение преобразуется в битовый поток и подвергается маппингу на точки сигнального созвездия.

Для моделирования условий распространения сигнала применяется параметрическая стохастическая модель канала. Она включает внесение частотного смещения Δf, имитирующего эффект Доплера и нестабильность гетеродинов, а также добавление аддитивного белого гауссовского шума (AWGN). Уровень шума определяется целевым отношением сигнал / шум (SNR), задаваемым в диапазоне от 0 до 25 дБ. Генерируемые выборки сохраняются в формате массивов NumPy (.npy) для непосредственной подачи на вход нейронных сетей, а также сопровождаются файлом метаданных (.json), содержащим эталонную разметку параметров генерации.

Для формализованной постановки задачи синтеза и верификации данных использованы математические модели, описывающие преобразование значений яркости пикселей исходного изображения в параметры (амплитуду, частоту или фазу) формируемого радиосигнала. Модель разделяется на две ветви в соответствии с классом модуляции.

В аналоговой ветви (AM, SSB, FM, PM) формируется непрерывный модулирующий сигнал $m[n]$ непосредственно из значений яркости пикселей исходного изображения I. Процесс включает операции изменения размера, векторизации и центрирования относительно нуля: $m[n] = a\left(\mathrm{Re}\, size\left(I, N_{samples}\right)\right)[n] - 0,5\Big)$, где a – коэффициент масштабирования (в реализации $a = 0,8$), $N_{samples} = F_s T$ – общее число отсчетов.

Для амплитудной модуляции (AM) комплексная огибающая $S_{AM}[n]$ формируется как: $S_{AM}[n] = 1 + m[n]$.

Для частотной модуляции (FM) сигнал определяется путем интегрирования модулирующего сообщения в показателе экспоненты:

$$S_{FM}[n] = \exp\left(j \cdot 2\pi \cdot k_f \cdot \sum_{i=0}^{n} m[i] \cdot \frac{1}{F_s} \right),$$

где k_f – коэффициент девиации частоты.

В цифровой ветви (PSK, QAM, FSK) изображение преобразуется в битовый поток *b*, который разбивается на упорядоченные кортежи длиной $k = \log_2 M$.

При квадратурной амплитудной модуляции (например, 16-QAM, $M = 16$) кортежи из четырех бит $(b_{4i},\ldots,b_{4i+3})$ отображаются на символы квадратного созвездия. Разделение битов производится для независимого кодирования синфазной (*I*) и квадратурной (*Q*) компонент с использованием симметричного алфавита уровней амплитуды $A = \{-3,-1,1,3\}$: $d_i = A\left[Map\left((b_{4i},b_{4i+1})\right)\right] + j \cdot A\left[Map\left((b_{4i+2},b_{4i+3})\right)\right]$.

При частотной манипуляции (M-FSK) кортеж бит определяет индекс частоты f_m, а сигнал для символьного интервала формируется как: $s_i(t) = \exp\left(j \cdot 2\pi \cdot f_m \cdot t\right),\ t \in [0, T_s]$.

Сформированный комплексный сигнал $s[n]$ поступает на вход модели канала распространения. В рамках разработанной методики используется аддитивная стохастическая модель, имитирующая искажения в канале связи «Космический аппарат – Земля». Математически принятый сигнал $r[n]$ описывается выражением: $r[n] = s[n] \cdot e^{j \cdot 2\pi \cdot \Delta f \frac{\pi}{F_s}} + w[n]$, где Δf – частотное смещение, $w[n]$ – отсчеты аддитивного белого гауссовского шума (AWGN).

Величина Δf выбирается случайным образом для каждого образца из равномерного распределения $\Delta f \sim U(-2500, 2500)$ Гц, что позволяет моделировать совокупное влияние нестабильности гетеродинов и эффект Доплера для низкоорбитальных спутников[1].

Уровень шума определяется целевым отношением сигнал / шум (SNR). Средняя мощность полезного сигнала P_s и требуемая мощность шума P_n рассчитываются следующим образом: $P_s = \frac{1}{N}\sum_{n=0}^{N-1}\left|s[n]\right|^2$, $P_n = \frac{P_s}{10^{SNR_{dB}/10}}$.

Комплексные отсчеты шума $w[n]$ генерируются так, чтобы энергия делилась поровну между действительной и мнимой составляющими: $\mathrm{Re}\{w\}, \mathrm{Im}\{w\} \sim \mathcal{N}\left(0, P_n/2\right)$. В разработанном датасете величина SNR варьируется в диапазоне от 0 до 25 дБ, охватывая ситуации от порогового приема до уверенной связи.

Для экспериментальной валидации методики был сформирован тестовый набор данных. В качестве источника полезного сигнала использованы 100 высокодетальных снимков облачного покрова Земли, полученных с геостационарного гидроме-

[1] Sharma E., Deo R., Davey C., Carter B. Artificial intelligence-empowered Doppler weather profile for low-earth-orbit satellites. *Sensors*, 2024, vol. 24, no. 16, pp. 5271. https://doi.org/10.3390/s24165271

теорологического спутника «Электро-Л» № 3 (GOMS-4). Выбор данного источника обусловлен сложной спектральной и текстурной структурой изображений, что является репрезентативным примером реального трафика систем ДЗЗ.

Общий объем экспериментального набора данных составил 2000 образцов радиосигналов (20 реализаций различных видов модуляции на каждое изображение). Параметры генерации включали частоту дискретизации 48 кГц, вариацию SNR от 0 до 25 дБ и случайное частотное смещение в диапазоне ±2500 Гц.

Оценка качества передачи визуальной информации проводилась на примере амплитудной модуляции (АМ). После прохождения сигнала через модель канала с шумом (AWGN) выполнялась программная демодуляция и восстановление изображения.

В качестве метрик использованы коэффициент корреляции Пирсона (R), оценивающий степень линейной зависимости и структурного подобия, и среднеквадратичная ошибка (MSE).

Для тестового образца при уровне шума, соответствующем SNR = 12,5 дБ, получены следующие показатели:

– коэффициент корреляции между исходным и декодированным изображением $R \approx 0{,}8091$;

– среднеквадратичная ошибка MSE ≈ 0,8018.

Визуальный анализ результатов (см. рис. 2) показывает, что несмотря на наличие зернистости, вызванной шумом канала, основные структурные элементы изображения (контуры облачности, границы терминатора) сохраняются. Значение корреляции $R > 0{,}8$ подтверждает, что в аналоговой ветви генератора сигнал физически переносит информацию о структуре изображения, а модель канала вносит адекватные стохастические искажения, не разрушающие полезный сигнал полностью.

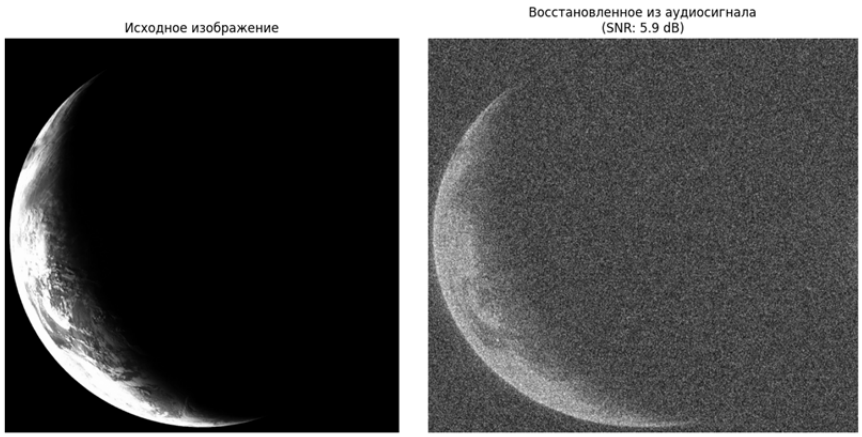


Рис. 2. Сравнительный анализ исходного изображения ДЗЗ
и результата демодуляции АМ-сигнала (SNR = 12,5 дБ)

Для цифровой ветви (PSK, QAM, FSK) критически важным критерием является корректность математической модели канала распространения. Верификация проводилась путем оценки помехоустойчивости модели на примере двоичной фазовой манипуляции (BPSK).

Была рассчитана зависимость вероятности битовой ошибки (Bit Error Rate, BER) от отношения сигнал / шум. Результаты эксперимента представлены в таблице.

Таблица

Зависимость вероятности битовой ошибки от SNR

Отношение сигнал / шум (SNR), дБ	Количество ошибок (на 100000 бит)	Вероятность ошибки (BER)
0	7782	$7{,}78 \cdot 10^{-2}$ $(7{,}8\,\%)$
3	2284	$2{,}28 \cdot 10^{-2}$ $(2{,}3\,\%)$
6	226	$2{,}26 \cdot 10^{-3}$ $(0{,}2\,\%)$
9	2	$2{,}00 \cdot 10^{-5}$
12	0	$< 10^{-5}$

Анализ результатов показывает строгое соответствие теоретическим ожиданиям для канала с аддитивным белым гауссовским шумом. При низком SNR (0 дБ) вероятность ошибки составляет порядка 7,8 %, что соответствует условиям порогового приема. При увеличении SNR до 6 дБ наблюдается снижение BER более чем на порядок (до 0,2 %), а при 12 дБ достигается безошибочный прием в рамках объема тестовой выборки.

Графическая зависимость (см. рис. 3) демонстрирует совпадение экспериментальной кривой с теоретической кривой помехоустойчивости BPSK, что подтверждает корректность реализации генератора шума и модулятора.

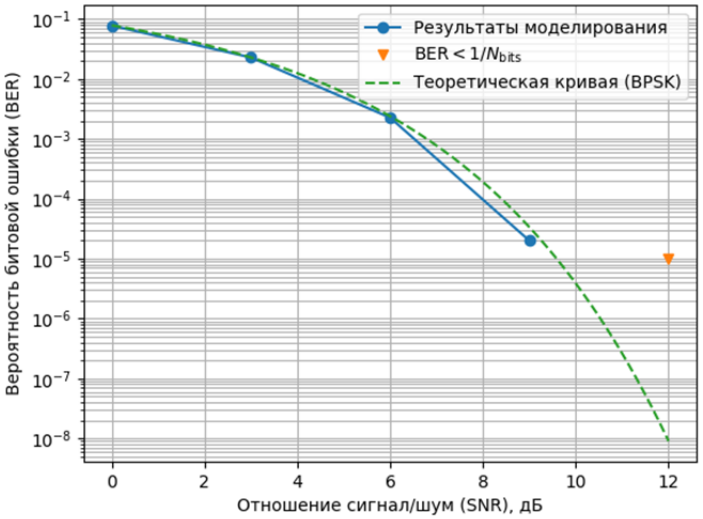


Рис. 3. Экспериментальная и теоретическая зависимость BER от SNR

По результатам работы комплекса сформирован верифицированный набор данных. Структура каждого образца унифицирована и содержит четыре компонента:

1) файл iq_data.npy – массив комплексных чисел (complex64), являющийся основным объектом для обучения нейросетей (сохраняет фазовую информацию);

2) файл meta.json – эталонная разметка (Ground Truth) с параметрами модуляции и SNR;

3) файл waterfall.png – спектрограмма (STFT) для визуального контроля и задач компьютерного зрения;

4) файл audio.ogg – аудиозапись демодулированного сигнала.

Использование изображений ДЗЗ обеспечило высокую энтропию цифровых сигналов и наличие корреляционных связей в аналоговых сигналах, что выгодно отличает полученный датасет от синтетических наборов, построенных на генераторах случайных чисел.

Выводы

На основании результатов теоретического анализа, программной реализации и экспериментальной верификации разработанной методики синтеза радиосигналов можно сделать следующие выводы.

1. Разработана гибридная методика синтеза радиосигналов, ключевой особенностью которой является отказ от использования псевдослучайных битовых последовательностей в пользу изображений дистанционного зондирования Земли, обладающих выраженной двумерной пространственной структурой (в отличие от равномерного белого шума, генерируемого математическими алгоритмами). Это позволило

сформировать сигналы с реалистичной информационной энтропией и сохранить корреляционные связи, необходимые для задач семантического анализа.

2. Экспериментально подтверждена корректность формирования аналоговых сигналов. При отношении сигнал / шум 12,5 дБ коэффициент корреляции между исходным и восстановленным изображениями составил $R > 0,8$, что свидетельствует о высокой степени сохранения структурного подобия полезного сигнала после прохождения канала.

3. Верифицирована математическая модель канала распространения для цифровых видов модуляции. Полученная экспериментальная зависимость вероятности битовой ошибки (BER) от отношения сигнал / шум полностью совпадает с теоретической кривой для канала с аддитивным белым гауссовским шумом (AWGN), что подтверждает физическую адекватность вносимых искажений.

4. Сформирован сбалансированный и размеченный набор данных объемом 2000 образцов, включающий комплексные квадратурные выборки, спектрограммы и файлы метаданных. Унифицированная структура датасета делает его готовым инструментом для обучения и тестирования нейросетевых моделей автоматической классификации радиосигналов.

Список литературы / References

Abd-Elaziz O., Abdalla M., Elsayed R. Deep learning-based automatic modulation classification using robust CNN architecture for cognitive radio networks. *Sensors*, 2023, vol. 23, no. 23, pp. 9467. https://doi.org/10.3390/s23239467

Scholl S. RF signal classification with synthetic training data and its real-world performance. *arXiv*, 2022. https://doi.org/10.48550/arXiv.2206.12967

Sharma E., Deo R., Davey C., Carter B. Artificial intelligence-empowered Doppler weather profile for low-earth-orbit satellites. *Sensors*, 2024, vol. 24, no. 16, pp. 5271. https://doi.org/10.3390/s24165271

Snoap J. A., Popescu D. C., Latshaw J. A., Spooner C. M. Deep-learning-based classification of digitally modulated signals using capsule networks and cyclic cumulants. *Sensors*, 2023, vol. 23, iss. 12, pp. 5735. https://doi.org/10.3390/s23125735

Snoap J. A., Popescu D. C., Spooner C. M. Deep-learning-based classifier with custom feature-extraction layers for digitally modulated signals. *IEEE Transactions on Broadcasting*, 2024, vol. 70, no. 3, pp. 763–773. https://doi.org/10.1109/TBC.2024.3391056

Spooner C. *DeepSig's 2018 Dataset: 2018.01.OSC.0001_1024x2M.h5.tar.gz*, 2020. Available at: https://cyclostationary.blog/2020/09/24/deepsigs-2018-data-set-2018-01-osc-0001_1024x2m-h5-tar-gz/ (accessed: 16.01.2026).

Spooner C. *More on DeepSig's RML Datasets*, 2020. Available at: https://cyclostationary.blog/2020/08/17/more-on-deepsigs-rml-data-sets/ (accessed: 16.01.2026).

Wang D., Lin M., Xiaoxu Z., et al. Automatic modulation classification based on CNN-transformer graph neural network. *Sensors*, 2023, vol. 23, no. 16, pp. 7281. https://doi.org/10.3390/s23167281

Сведения об авторах

Алина Владимировна Кобенко – аспирант, https://orcid.org/0000-0001-6857-1799, a.kobenko@narfu.ru, Северный (Арктический) федеральный университет им. М. В. Ломоносова (д. 17, наб. Северной Двины, 163002 Архангельск, Россия); **Alina V. Kobenko** – Postgraduate Student, https://orcid.org/0000-0001-6857-1799, a.kobenko@narfu.ru, Northern (Arctic) Federal University named after M. V. Lomonosov (17, nab. Severnoi Dviny, 163002 Arkhangelsk, Russia).

Роман Александрович Алешко – кандидат технических наук, доцент, https://orcid.org/0000-0002-1235-8937, r.aleshko@narfu.ru, Северный (Арктический) федеральный университет им. М. В. Ломоносова (д. 17, набережная Северной Двины, 163002 Архангельск, Россия); **Roman A. Aleshko** – Candidate of Technical Sciences, Associate Professor, https://orcid.org/0000-0002-1235-8937, r.alesko@narfu.ru, Northern (Arctic) Federal University named after M. V. Lomonosov (17, nab. Severnoi Dviny, 163002 Arkhangelsk, Russia).

Заявленный вклад авторов: авторы сделали разный вклад в подготовку публикации, что отражено в последовательности персоналий авторского коллектива. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors made different contributions to the preparation of the publication, which is reflected in the sequence of personalities of the authors' team. The authors declare no conflicts of interests.

Статья поступила в редакцию 06.04.2026; одобрена после рецензирования 25.05.2026; принята к публикации 10.06.2026.

The article was submitted 06.04.2026; Approved after reviewing 25.05.2026; Accepted for publication 10.06.2026.