16. *Huntington D.E., Lyrintzis C.S.* Improvements to and limitations of Latin hypercube sampling, *Probabilistic engineering mechanics*, 1998, Vol. 13, No. 4, pp. 245-253.
17. *Faure H., Lemieux C.* Generalized Halton sequences in 2008: A comparative study, *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 2009, Vol. 19, No. 4, pp. 1-31.
18. *Wang X., Hickernell F.J.* Randomized halton sequences, *Mathematical and Computer Modelling*, 2000, Vol. 32, No. 7-8, pp. 887-899.
19. *Faure H., Lemieux C.* Generalized Halton sequences in 2008: A comparative study, *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 2009, Vol. 19, No. 4, pp. 1-31.
20. *Pitzer E., Affenzeller M.* A comprehensive survey on fitness landscape analysis, *Recent advances in intelligent engineering systems*, 2012, pp. 161-191.

Пикалов Максим Вадимович – Национальный исследовательский университет ИТМО; e-mail: pikmaksim@gmail.com; г. Санкт-Петербург, Россия; тел.: 88126070283; аспирант.

Pikalov Maxim Vadimovich– ITMO University; e-mail: pikmaksim@gmail.com; St. Petersburg, Russia; phone: +78126070283; graduate student.

УДК 004.032.26 DOI 10.18522/2311-3103-2026-1-52-64

М.О. Доброхвалов, А.Ю. Филатов, Е.А. Чегодаева

НОВАЯ МЕТРИКА ВОСПРОИЗВОДИМОСТИ ДЛЯ СРАВНЕНИЯ КЛАССИФИКАТОРОВ ВРЕМЕННЫХ РЯДОВ

Воспроизводимость результатов экспериментов является критическим аспектом современного машинного обучения, однако выбор случайного инициализирующего сида существенно влияет на итоговое качество моделей, что создает проблему корректного сравнения различных архитектур и методов. Цель исследования заключалась в оценке влияния выбора случайного сида на результаты классификации временных рядов сверточными нейросетями и в разработке корректного способа сравнения моделей. Задачи включали измерение разброса метрик при множественных повторных запусках в рамках варьирующихся инициализаций, проверку нормальности распределений, введение метаметрики воспроизводимости RM и подбор ее параметра λ, а также проверку переносимости подхода на альтернативных архитектурах. Проведены эксперименты с двумя одномерными архитектурами (FCN, ResNet) сверточных нейронных сетей на семи открытых наборах данных временных рядов разной природы. Для каждой пары модель–датасет выполнено по 55 независимых запусков с фиксацией источников случайности и идентичными настройками обучения в PyTorch. Статистический анализ включал критерии Шапиро Уилка и Андерсона Дарлинга. Показано, что распределения аккуратности чаще всего не соответствуют нормальному закону, поэтому интервальные оценки, основанные на нормальности, некорректны. Варьирование сида приводит к различиям аккуратности до 12 процентных пунктов, причем величина разброса зависит от датасета и архитектуры. Предложенная метрика воспроизводимости (RM), штрафующая за дисперсию, при малом числе запусков на различных инициализирующих значениях приближает нижнюю наблюдаемую границу, при большом числе стремится к среднему. Предложенная RM позволяет сравнивать модели с учетом случайных "удачных" и "неудачных" инициализаций и ранжировать модели по устойчивости, задает стандарт отчетности, повышающий воспроизводимость экспериментов и надежность выводов. Эмпирическая проверка на архитектуре DenseNet подтвердила, что RM адекватно реагирует как на стабильные, так и на нестабильные наборы. Методика легко переносится на новые датасеты и архитектуры. Предложенная метрика может быть использована для стандартизации отчётности и повышения воспроизводимости исследований.

Классификация временных рядов; сверточные нейронные сети; случайный сид; чувствительность к инициализации; воспроизводимость экспериментов.

M.O. Dobrokhvalov, A.Yu. Filatov, E.A. Chegodaeva

A NEW REPRODUCIBILITY METRIC FOR COMPARING TIME SERIES CLASSIFIERS

Experimental reproducibility constitutes a critical cornerstone of modern machine learning research, yet random initialization seed selection substantially influences final model performance, creating challenges for principled comparison of different architectures and methods. Random seed effects on convolutional time series classifiers were quantified, and a principled comparison criterion was established. Two 1D architectures, FCN and ResNet, were trained on seven public datasets containing different data. 55 independent runs for each combination of model and dataset were performed nder controlled pseudorandomness in Python, NumPy, and PyTorch. Deterministic backends were enabled, and identical hyperparameters were used across runs. Normality of seed-wise accuracy distributions was assessed with the Shapiro–Wilk and Anderson–Darling tests. Accuracy variability attributable to seed choice reached up to 12 percentage points in some settings, with magnitude dependent on dataset and architecture. The distributions were found to be non-normal in most cases, indicating that confidence intervals predicated on normality are unreliable. To enable fair comparison across runs, a reproducibility meta-metric, RM, was introduced that subtracts a dispersion penalty from the mean and depends on the number of runs and a tunable coefficient λ. RM was shown to lie between the empirical minimum and the mean, to approach the lower bound for small sample sizes, and to converge toward the mean as the number of runs increases. Portability of the approach was examined on an additional architecture, DenseNet, confirming expected behavior. Practical value is provided by RM metric rankings reflect both performance and stability. In this way, reproducibility and the credibility of empirical conclusions are strengthened.

Time series classification; convolutional neural networks; random seed; initialization sensitivity; experimental reproducibility.

Введение. Машинное обучение продолжает активно развиваться, в частности, нейронные сети. До начала процесса обучения веса сети инициализируются определенным образом. Инициализация весов сверточной сети влияет на конечное состояния, к которому сходится модель. Существуют различные методы инициализации весов [1].

Однако, даже в рамках одного метода инициализации веса могут иметь различные значения, которые зависят от состояния генератора случайных чисел. Это может вносить различия в результаты обучения. В работе [2] рассматривается влияние инициализирующего сида на результаты классификации с использованием разновидностей модели VGG на наборе данных CIFAR-10 [3]. Авторами рассматривалось три способа выбора значения для сида: (i) последовательные значения [0, 19], (ii) получение псевдослучайного значения с помощью библиотеки numpy[1] с инициализацией состояния с помощью последовательного выбора значений [0, 19], (iii) использование стандартной инициализации без фиксирования состояния. Авторы демонстрируют, что разность между минимальным и максимальным значением аккуратности (accuracy) может составлять более 3, при среднем значении 90,584 и стандартном отклонении 0,216. В [4] рассматривают вероятностную модель скрытых состояний для описания динамики обучения нейросетей с использованием 40 сидов.

В работе [5] также рассматривается результат влияния инициализирующего сида на результаты предсказания. Автор рассматривает 3 различных архитектуры/инициализации на 2 датасетах: CIFAR-10 [6] и ImageNet [7]. Автор демонстрирует, что в рамках датасета CIFAR-10 разность между минимальным и максимальным значениями метрики аккуратности может достигать 1,82 при среднем значении 90,02 и стандартном отклонении 0,23. При использовании датасета ImageNet стандартное отклонение составляет около 0,1, а разность между минимальным и максимальным значениями составляет около 0,5 при среднем значении около 76.

В работе [8] предлагается механизм агрегации с вниманием для повышения точности сверточных сетей, при этом все эксперименты проводились с фиксированным сидом 0. В статье [9] для всех экспериментов в области генеративного искусственного интеллекта фиксировался сид 42.

[1] https://numpy.org/doc/stable/.

На основании [10] в ряде работ применяли множество сидов для количественной оценки устойчивости и статистической надёжности предлагаемых методов. В [11] использовали 11 сидов при решении задачи одновременного определения типов клеток и сегментации тканей, в [12] провели эксперименты с 5 сидами в задаче детектирования объектов, в [13] рассмотрели 100 сидов при оптимизации процедуры обучения ResNet для решения задачи классификации, в [14] оценили влияние 50 сидов на точность медицинской сегментации, в [15] применили 10 сидов для отбора наиболее качественных результатов преобразования текста в изображение. Также в работах [16] авторы указывают, что фиксируют сид для исключения случайного выбора “удачного сида”.

При сравнении исследований обычно используются результаты, указанные другими авторами. При этом повторить эксперимент часто не представляется возможным из-за недостаточного описания постановки.

В данной работе рассмотрено влияние инциализирующего сида на результаты классификации двух сверточных сетей, обрабатывающих временные ряды, взятые из 7 открытых наборов данных.

Формализация задачи. Пусть задан набор данных $D=\{(x_i, y_i)\}_{i=1}^{N}$,, где $x_i \in R^T$ – временной ряд длины T, а $y_i \in \{1,2,\dots,K\}$ – метка класса. Требуется обучить классификатор $f_\theta: R^T \to \{1,2,\dots,K\}$ с параметрами θ^0, минимизирующий функцию потерь. Итоговые параметры θ^* зависят от случайного сида $s \in N$, определяющего начальную инициализацию весов $\theta^0(s)$. Для фиксированного датасета и архитектуры получаем множество значений метрики качества $\{m(s_1), m(s_2), \dots, m(s_n)\}$, где $m(s_j)$ – значение метрики качества модели, обученной после инициализации весов s_j. Требуется оценить статистические значения набора метрик качества.

Модели и данные. В дальнейших исследованиях использовалось архитектуры сверточных нейронных сетей: Fully Convolutional Network (FCN) [17], ResNet [18].

Кодировщик архитектуры FCN, основан на более ранних сверточных нейронных сетях для классификации изображений, таких как VGG-16. Ключевой модификацией является замена полностью связных слоёв на сверточные слои, что позволяет сети работать с изображениями произвольного размера и сохранять пространственные характеристики данных. Энкодер, использующийся в данной работе как сеть для классификации, представляет собой серию сверточных и слоёв пулинга, которые постепенно уменьшают пространственное разрешение, извлекая высокоуровневые признаки из входного изображения. На выходе кодировщика формируются семантически богатые, но низкоразмерные карты признаков, которые затем используются для последующего восстановления пиксельных классификаций в декодере. Также в оригинальной работе рассматривается сеть для обработки изображений. В данной работе она была адаптирована для обработки одномерных сигналов, путём замены двумерных слоев на одномерные.

Также рассматривается адаптированная для решения задачи классификации одномерных временных рядов архитектура сети ResNet. Представленная модификация архитектуры сохраняет ключевые элементы исходной сети [19], при этом учитывает специфику временных данных. Кодировщик сети состоит из серии последовательно соединённых остаточных (residual) блоков, каждый из которых включает три свёрточных слоя с фильтрами размеров 1×1, 3×3 и снова 1×1. Особое внимание уделяется механизму остаточных соединений (skip-connections), которые обеспечивают прямую передачу информации между входом и выходом блока, который позволяет уменьшить эффект исчезновения градиентов и упрощает процесс обучения более глубоких сетей. После последовательности резидуальных блоков используется слой глобального среднего объединения (global average pooling), который агрегирует информацию вдоль временного измерения. Архитектура завершается полностью связанным выходным слоем. Данная архитектура также дополнительно адаптирована в рамках данной работы для работы с одномерными данными.

В рамках данного исследования было использовано 7 наборов данных[2].

Датасет Adiac состоит из временных рядов длиной 176 отсчётов, представляющих контуры изображений объектов, сгруппированных в 37 классов. Он содержит 390 обучающих и 391 тестовый экземпляр. Задача — классификация форм по контурам. ECG5000 основан на 20-часовой ЭКГ пациента с сердечной недостаточностью. Из записи были выделены и интерполированы сердечные циклы, после чего случайным образом отобрано 5000 циклов, разбитых на 5 классов. Датасет Fish включает ряды длиной 463 отсчёта, характеризующие 7 видов рыб (по 175 экземпляров в обучающем и тестовом наборах). Lightning2 содержит 2 класса спектрограмм молний длиной 637 отсчётов, всего 60 обучающих и 61 тестовых экземпляров. Lightning7 усложнён за счёт 7 классов и более коротких рядов (319 отсчётов), содержит 70 обучающих и 73 тестовых экземпляра. SonyAIBORobotSurface1 и SonyAIBORobotSurface2 содержат сигналы акселерометров робота Sony AIBO, длиной 70 и 65 отсчётов соответственно, с двумя классами. Тренировочных экземпляров 20 и 27, тестовых 601 и 953 соответственно.

Эксперимент

Постановка.

Фреймворк PyTorch[3], активно применяемый в современных исследованиях, предоставляет инструменты для улучшения воспроизводимости результатов. Основным методом, позволяющим уменьшить неопределённость результатов экспериментов, является фиксация случайного сида (инициализирующего состояния генератора псевдослучайных чисел). Однако, помимо фиксации сида в самом PyTorch, необходимо также фиксировать сиды стандартной библиотеки Python и библиотеки работы с массивами NumPy, поскольку эти библиотеки также генерируют случайные числа, используемые в вычислениях. Кроме того, важным шагом для достижения стабильной воспроизводимости является активация режима детерминированных операций PyTorch и отключение бенчмаркинга, который в противном случае может автоматически выбирать наиболее быстрые алгоритмы для операций, вводя дополнительную неопределённость (листинг 1).

Листинг 1

Фиксирование сида

```
random.seed(seed)
np.random.seed(seed)
torch.manual_seed(seed)
torch.backends.cudnn.deterministic = True
torch.backends.cudnn.benchmark = False
```

Также для подачи данных в едином порядке, сид генератора загрузчика данных дополнительно фиксировался значением 2025.

В качестве функции потерь использовалась перекрестная энтропия (CrossEntropyLoss), оптимизатор – Adam (шаг – 0.001). Для обучения использовалось 500 эпох. Все эксперименты проводились с использованием графического процессора NVIDIA AD106M (GeForce RTX 4070 Max-Q, 8188 MiB).

Результаты эксперимента.

В табл. 1 представлены статистические характеристики аккуратности моделей, решающих задачу классификации, обученных на датасетах с временными рядами. Для исследования взяты простые и хорошо изученные архитектуры. Рассмотрено 55 сидов [1971, 2025].

[2] https://www.timeseriesclassification.com/dataset.php.
[3] https://docs.pytorch.org/docs/stable/index.html.

Для сети FCN наиболее устойчивые результаты наблюдаются на Fish и SonyAIBORobotSurface1. В первом случае стандартное отклонение составило 0,005 при размахе 0,023, во втором 0,007 при размахе 0,028. Близкий уровень стабильности показан на ECG5000, стандартное отклонение 0,008 при размахе 0,036. Наиболее нестабильное поведение FCN демонстрирует на Lightning7, где стандартное отклонение достигает 0,024, а размах 0,110. Повышенную вариативность видно и на Lightning2, стандартное отклонение 0,015 при размахе 0,082.

Для сети ResNet отклонение и размах также зависят от датасета. Наименьшая вариативность наблюдается на Fish и SonyAIBORobotSurface2, где стандартные отклонения составили 0,005 и 0,005 при размахе 0,017 и 0,023. Сходный уровень стабильности получен на SonyAIBORobotSurface1, стандартное отклонение 0,008 при размахе 0,037. Наибольшая нестабильность зафиксирована на Lightning2 и ECG5000, где стандартное отклонение достигло 0,032 и 0,026 соответственно, а размах 0,115 и 0,123.

Таким образом, для обеих архитектур наибольшую устойчивость обеспечивают Fish, SonyAIBORobotSurface1 и SonyAIBORobotSurface2, тогда как Lightning2 и ECG5000 характеризуются высокой вариативностью результатов. Для FCN к этой группе нестабильных наборов добавляется Lightning7. Стандартное отклонение и размах значений точности остаются ключевыми параметрами при оценке надежности моделей на различных задачах.

Таблица 1

Характеристики аккуратности моделей, обученных на разных сидах

Сеть	Датасет	Средняя	Стандартное отклонение	Минимальная	Максимальная	Размах
FCN	Adiac	0,808	0,015	0,772	0,847	0,074
	ECG5000	0,924	0,008	0,899	0,935	0,036
	Fish	0,969	0,005	0,954	0,977	0,023
	Lightning2	0,696	0,015	0,656	0,738	0,082
	Lightning7	0,772	0,024	0,712	0,822	0,110
	SonyAIBORobotSurface1	0,957	0,007	0,943	0,972	0,028
	SonyAIBORobotSurface2	0,958	0,009	0,933	0,979	0,046
ResNet	Adiac	0,783	0,013	0,757	0,806	0,049
	ECG5000	0,884	0,026	0,798	0,922	0,123
	Fish	0,979	0,005	0,971	0,989	0,017
	Lightning2	0,712	0,032	0,656	0,771	0,115
	Lightning7	0,788	0,024	0,740	0,822	0,082
	SonyAIBORobotSurface1	0,977	0,008	0,957	0,993	0,037
	SonyAIBORobotSurface2	0,970	0,005	0,957	0,980	0,023

Таким образом, исследования, утверждающие, что значение метрики превышает аналоги на десятые или сотые доли аккуратности, в некоторых случаях могут являться примером удачной инициализации весов сети.

Нормальность распределения метрик.

На рис. 1 представлено распределение аккуратности классификации моделей, указанных выше. Из графиков видно, что чаще всего плотность распределения имеет один пик, вокруг которого сосредоточено основное количество результатов.

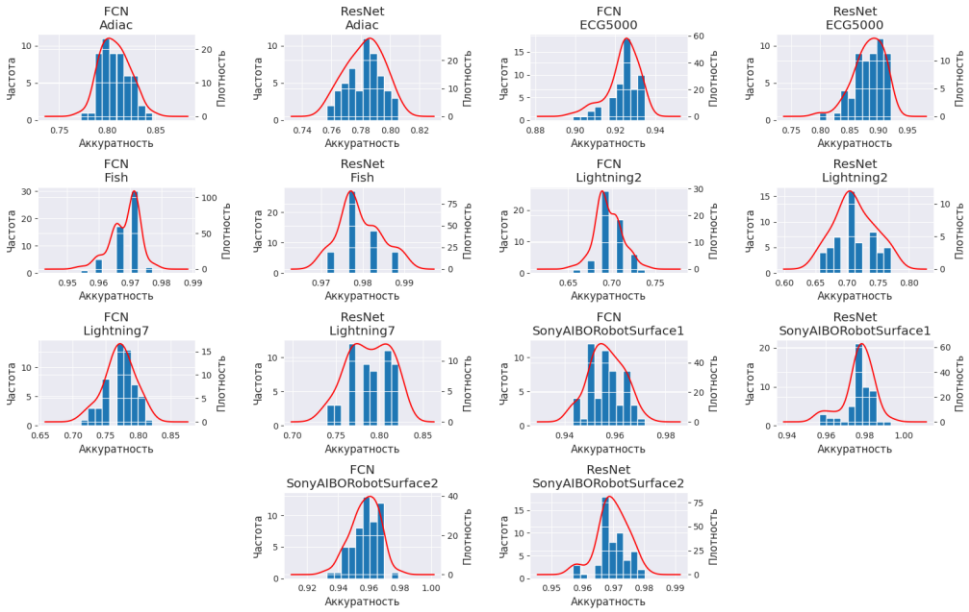

Рис. 1. Плотность распределения аккуратности классификации на различных датасетах

Проверка распределения аккуратности моделей на соответствие нормальному проводилась с целью выяснить, допустимо ли использование квантилей нормального закона для построения доверительных интервалов. Для оценки соответствия распределений были применены критерии Шапиро–Уилка (p-значение больше 0,05) и Андерсона–Дарлинга (уровень значимости 0,05). Результаты представлены в табл. 2. Модель FCN демонстрирует нормальность распределения на датасетах Adiac, SonyAIBORobotSurface1 и SonyAIBORobotSurface2 по обоим критериям. Для Lightning7 критерий Шапиро–Уилка дал положительный результат $p = 0{,}0813$, однако по критерию Андерсона–Дарлинга значение статистики 0,9266 превысило критический уровень 0,739, что указывает на отсутствие нормальности. На остальных наборах данных (ECG5000, Fish, Lightning2) оба критерия опровергли гипотезу нормальности.

Результаты модели ResNet показали меньшее соответствие нормальному распределению по критериям Шапиро–Уилка и Андерсона–Дарлинга, подтвердив его только на Adiac и SonyAIBORobotSurface2. По критерию Андерсона–Дарлинга распределение соответствовало нормальному для ECG5000, несмотря на низкое p-значение 0,0172 по Шапиро–Уилку. На остальных датасетах (Fish, Lightning2, Lightning7, SonyAIBORobotSurface1) оба критерия показали отклонение от нормального распределения.

Таким образом, распределения метрик моделей FCN и ResNet в большинстве случаев не являются нормальными. Обе модели на обоих критерия отклонили гипотезу нормальности на датасетах Fish и Lightning2. На датасатетах Adiac и SonyAIBORobotSurface2 обе модели удовлетворяют условиям нормальности обоих тестов. В остальных комбинациях датасетов и моделей наблюдается расхождение.

Таблица 2

Нормальность распределения значений аккуратности, обученных на разных сидах

Сеть	Датасет	Р-значение критерия Шапиро–Уилка	Критерий Андерсона–Дарлинга
FCN	Adiac	0,6573	0,3593
	ECG5000	0,0002	1,7574
	Fish	0,0000	5,1043
	Lightning2	0,0001	3,2711
	Lightning7	0,0813	0,9266
	SonyAIBORobotSurface1	0,3206	0,4098
	SonyAIBORobotSurface2	0,4734	0,3895
ResNet	Adiac	0,2832	0,4118
	ECG5000	0,0172	0,6345
	Fish	0,0000	3,6117
	Lightning2	0,0187	0,9820
	Lightning7	0,0025	1,3357
	SonyAIBORobotSurface1	0,0002	2,1110
	SonyAIBORobotSurface2	0,0913	0,5973

Новая метрика воспроизводимости. Наиболее часто используемые метрики для оценки качества моделей: аккуратность (accuracy), точность (precision), полнота (recall), F1-мера. В связи с наблюдаемым варьированием результатов при повторных запусках с использованием различных состояний генератора случайных чисел предлагается ввести некоторый способ учета при сравнении результатов.

Можно использовать среднее значение и стандартное отклонение, однако в данном случае отсутствует универсальность при варьирующемся количестве запусков на разных сидах. Поэтому имеет смысл ввести мета-метрику воспроизводимости:

$$RM(\lambda, n) = \mu_m - \lambda \frac{\sigma_m}{\sqrt{n}},$$

где $\lambda \geq 0$ – коэффициент штрафа за нестабильность, который нужно выбирать эмпирически, n – количество повторных запусков на различных инициализирующих сидах. Чем выше значение λ, тем сильнее учитывается влияние разброса результатов между запусками.

Предлагаемая формула структурно аналогична нижней границе доверительного интервала для среднего значения. Однако оригинальную формулу доверительного интервала нельзя использовать из-за того, что распределение значений метрик при обучении на различных инициализирующих сидах не является нормальным.

На основе предложенной мета-метрики RM и наиболее часто используемых в задаче классификации метик можно рассмотреть 4 производных метрики: (i) $RM_{acc}(\lambda, n)$ – метрика воспроизводимости аккуратности; (ii) $RM_{precision}(\lambda, n)$ – метрика воспроизводимо-

сти точности; (iii) $RM_{recall}(\lambda, n)$ – метрика воспроизводимости полноты; (iv) $RM_{F1}(\lambda, n)$ – метрика воспроизводимости F1-меры. Аналогичным способом можно ввести метрику воспроизводимости для других метрик, представленных числовыми значениями.

Формализация выбора коэффициента

Для подбора коэффициента λ, определяющего величину штрафа в метрике воспроизводимости, рассматривается множество значений метрики качества, полученное из нескольких независимых запусков модели: $M = \{m_1, \dots, m_N\}$. Формируется набор X подмножеств x фиксированного размера n, для которого вычисляется значение $RM_x(\lambda, n)$. После вычисляется относительная ошибка

$$\delta_x = \frac{|RM(\lambda, n) - min\ (x)|}{min\ (x)}.$$

Коэффициент λ для фиксированного размера n определяется минимизацией средней ошибки по множеству подвыборок X:

$$\bar{\Delta_n}(\lambda) = \frac{1}{|X|}\sum_{x \in X} \delta_x(\lambda)$$

$$\lambda_n^\star = argmin_{\lambda \in \Lambda} \bar{\Delta_n}(\lambda),$$

где Λ – заданный диапазон допустимых значений λ, определенных в следующем разделе, а x – множество выбранных подвыборок размера n. Таким образом, $\lambda_n^\star$ – это такое значение коэффициента, при котором метрика RM в среднем наиболее близка к минимальному наблюдаемому значению метрики качества и, следовательно, наиболее строго и надёжно штрафует модели с высоким разбросом результатов.

Выбор параметров

Проведен эксперимент по выбору значения коэффициента λ. Для каждой пары «датасет–сеть» из первичной выборки, состоящей из 55 независимых запусков, сформировано множество значений воспроизводимости X. Далее для каждого фиксированного размера подвыборки $n \in \{5, 10, 15\}$ методом бутстрэппинга было сгенерировано по 1000 случайных подвыборок из X. Также были рассмотрены значения λ в диапазоне от 1 до 15 с шагом 0,1. Для каждой подвыборки и каждого λ вычислялась метрика воспроизводимости RM. Для оценки качества каждого λ были вычислены абсолютные значения относительной ошибки которая показывает, насколько близко оценка RM приближается к эмпирическому минимуму выборки X. Средняя величина Δ_n по 1000 значениям δ_n подвыборок и каждой пары «датасет–сеть» позволила определить λ, минимизирующее среднюю относительную ошибку при заданном n.

В табл. 3 представлены полученные значения λ_5, λ_{10}, λ_{15} с соответствующими средними ошибками Δ_5, Δ_{10}, Δ_{15}. Из таблицы виден рост значений и снижение значений при увеличении размера подвыборок. Также следует отметить, что для всех пар «датасет–сеть» соблюдается $\lambda_5 \leq \lambda_{10} \leq \lambda_{15}$, при этом $\Delta_5 \geq \Delta_{10} \geq \Delta_{15}$. При меньших n дисперсия оценок метрики возрастает, что соответствует условиям с наибольшей статистической неопределенностью. Это позволяет определить такое значение λ, при котором метрика воспроизводимости будет близка к эмпирическому минимуму даже при самом ограниченном числе запусков.

Таблица 3

Значение, λ при котором достигается минимальная разность с наименьшим значением

Датасет	Сеть	λ_5	Δ_5	λ_{10}	Δ_{10}	λ_{15}	Δ_{15}
Adiac	FCN	4,9	0,0125	7,4	0,0081	9,0	0,0060
Adiac	ResNet	4,1	0,0102	6,1	0,0069	7,6	0,0048
ECG5000	FCN	5,9	0,0118	9,1	0,0075	11,4	0,0060

Окончание табл. 3

Датасет	Сеть	λ_5	Δ_5	λ_{10}	Δ_{10}	λ_{15}	Δ_{15}
ECG5000	ResNet	7,0	0,0381	10,4	0,0268	12,3	0,0215
Fish	FCN	5,9	0,0056	8,8	0,0039	11,0	0,0029
Fish	ResNet	3,1	0,0023	4,8	0,0014	5,9	0,0011
Lightning2	FCN	6,0	0,0186	8,4	0,0125	9,9	0,0098
Lightning2	ResNet	3,6	0,0262	5,4	0,0167	6,7	0,0124
Lightning7	FCN	5,0	0,0281	7,8	0,0182	9,5	0,0135
Lightning7	ResNet	4,3	0,0199	6,1	0,0122	7,5	0,0094
SonyAIBORobotSurface1	FCN	3,8	0,0045	5,9	0,0027	7,3	0,0021
SonyAIBORobotSurface1	ResNet	4,5	0,0097	7,4	0,0059	9,5	0,0046
SonyAIBORobotSurface2	FCN	5,5	0,0083	8,1	0,0057	10,0	0,0043
SonyAIBORobotSurface2	ResNet	5,4	0,0051	7,3	0,0032	9,2	0,0025

Рекомендуемое значение λ вычисляется как средневзвешенное усредненных значений по размеру подвыборки для всех пар «датасет–сеть». Весовые коэффициенты $w_i = 1/\Delta$. Полученное значение составляет λ=4,51.

Результат

В табл. 4 представлены характеристики относительной ошибки, вычисленные для подвыборок размера 5 с использованием λ=4,51. Средние значения относительной ошибки находятся в диапазоне {-0.0071, 0.0460}, демонстрируя, что оценка RM в большинстве случаев близка к минимальному значению, полученному на полном множестве. Исключение составляет пара «ECG5000–ResNet», для которой зафиксировано максимальное среднее значение 0,046 и наибольшая дисперсия (0,0326), что указывает на нестабильность модели на данном датасете. На парах «Fish–ResNet» и «Lightning2–ResNet» средние ошибки являются отрицательными, что означает, что значение RM с предложенным λ в среднем превосходит эмпирический минимум на полном множестве. Медианные значения также остаются вблизи нуля, подтверждая отсутствие систематического смещения оценок в сторону переоценки или недооценки. Стандартное отклонение не превышает 0,0353. Таким образом, можно SonyAIBORobotSurface1заключить, что предложенное значение параметра сглаживания обеспечивает устойчивое и достаточно точное приближение к нижней границе метрики воспроизводимости.

Таблица 4

Характеристики относительной ошибки RM и минимального значения

Датасет	Сеть	mean	std	min	median	max
Adiac	FCN	0,0082	0,0138	-0,0382	0,0093	0,0506
Adiac	ResNet	0,0017	0,0128	-0,0363	0,0007	0,0426
ECG5000	FCN	0,0122	0,0099	-0,013	0,0125	0,0299
ECG5000	ResNet	0,046	0,0326	-0,0493	0,0512	0,1168

Окончание табл. 4

Датасет	Сеть	mean	std	min	median	max
Fish	FCN	0,0062	0,0051	-0,009	0,0066	0,0179
Fish	ResNet	-0,0018	0,0035	-0,0125	-0,0024	0,0089
Lightning2	FCN	0,0192	0,0179	-0,0443	0,0199	0,075
Lightning2	ResNet	-0,0071	0,0353	-0,1095	-0,0073	0,1129
Lightning7	FCN	0,0205	0,0286	-0,0602	0,023	0,0962
Lightning7	ResNet	0,0031	0,0247	-0,0582	0,0032	0,0926
SonyAIBORobotSurface1	FCN	0,0003	0,0053	-0,0151	0,0008	0,0144
SonyAIBORobotSurface1	ResNet	0,0068	0,0091	-0,0196	0,0083	0,0229
SonyAIBORobotSurface2	FCN	0,0082	0,0083	-0,0169	0,0076	0,0292
SonyAIBORobotSurface2	ResNet	0,0038	0,0048	-0,0103	0,005	0,0143

В табл. 5 представлены наименьшее и среднее значение метрик, а также RM(4.51, 55) на полном множестве для пар «датасет–сеть». Во всех парах значение RM, занимает промежуточное положение между минимальным и средним значением, что свидетельствует о сбалансированности оценки. Разность между RM и минимальным значением варьируется от 0,0049 (Fish–ResNet) до 0,0698 (ECG5000–ResNet). При этом разность между RM и среднем значением варьируется от 0,0027 (Fish–FCN) до 0,0195 (Lightning2–ResNet1D), что значительно меньше, чем до минимального. При этом средняя разность RM и минимального значения составляет 0.0252, а среднего значения и RM – 0.0084, что говорит о том, что RM значительно ближе к среднему значению, чем к минимальному. Таким образом, предложенное значение λ=4,51 позволяет при малых n проводить оценку нижней границы значения метрики. При этом, при увеличении количества повторений, позволяет сохранять значение метрики, близкое к среднему.

Таблица 5

Сравнение наименьшего значения выборки, метрики RM и среднего значения

Датасет	Сеть	min	RM	mean
Adiac	FCN	0,7724	0,7987	0,8078
Adiac	ResNet	0,757	0,775	0,7826
ECG5000	FCN	0,8989	0,9192	0,924
ECG5000	ResNet	0,7982	0,868	0,8838
Fish	FCN	0,9543	0,9658	0,9685
Fish	ResNet	0,9714	0,9763	0,9793
Lightning2	FCN	0,6557	0,6871	0,6963
Lightning2	ResNet	0,6557	0,6929	0,7124
Lightning7	FCN	0,7123	0,7576	0,7718

Окончание табл. 5

Датасет	Сеть	min	RM	mean
Lightning7	ResNet	0,7397	0,7736	0,7883
SonyAIBORobotSurface1	FCN	0,9434	0,9525	0,9566
SonyAIBORobotSurface1	ResNet	0,9567	0,9723	0,977
SonyAIBORobotSurface2	FCN	0,9328	0,952	0,9575
SonyAIBORobotSurface2	ResNet	0,957	0,9669	0,9698

Таким образом, результирующая новая предлагаемая мета-метрика воспроизводимости:

$$RM(\lambda) = \mu_m - 4{,}51\frac{\sigma_m}{\sqrt{n}}\ .$$

Валидация на DenseNet

Эксперименты по воспроизводимости дополнительно были проведены на тех же 7 датасетах с использованием сети DenseNet [20]. В таблице 6 представлены значения аккуратности и метрики RM для сети DenseNet с наибольшим и наименьшим стандартным отклонением аккуратности. Из таблицы видно, что при относительно небольшом стандартном отклонении 0,011 (датасет Fish) штраф также небольшой и среднее значение RM по 1000 подвыборок (размер 5) близко к среднему значению аккуратности. С другой стороны, на данных Lightning7 наблюдается большая нестабильность 0,033 и размер штрафа составил 0,064.

Таблица 6

Значения аккуратности и RM сети DenseNet

Датасет	Средняя аккуратность	Стандартное отклонение аккуратности	Размах	Среднее RM	Штраф
Fish	0,969	0,011	0,046	0,949	0,021
Lightning7	0,701	0,033	0,137	0,637	0,064

Заключение. Данное исследование подчеркивает значимость выбора случайного сида при обучении сверточных нейронных сетей, решающих задачи классификации временных рядов. Полученные результаты показывают, что варьирование состояния генератора случайных чисел может вызывать разницу в метрике качества – до 12% в отдельных случаях. Сравнение моделей, результаты которых получены на единственном запуске, могут являться некорректными. Демонстрируется необходимость систематической оценки устойчивости результатов.

В работе демонстрируется, что распределение значений аккуратности (accuracy), полученных при множественных запусках модели, в большинстве случаев не подчиняется нормальному закону. Это делает традиционные статистические методы оценки (такие как доверительные интервалы) ненадежными. В связи с этим предложена новая мета-метрика воспроизводимости (RM), сочетающая среднее значение метрики и штраф за дисперсию, задаваемый параметром λ. Подбор λ выполнен эмпирически на основе бутстрэппинга, что позволило обеспечить устойчивость оценки даже при малом числе запусков. Анализ показал, что RM находится между минимальным и средним значением метрики при большом количестве повторений и ближе к минимальному при малом количестве запусков.

Таким образом, работа не только дополнительно подтверждает существование проблемы нестабильности моделей при различной инициализации, но и предлагает практически применимый способ её учета. Предложенная методика RM может стать шагом к стандартизации оценки воспроизводимости в машинном обучении, особенно в контексте сравнений между новыми архитектурами и методами.

В дальнейшем перспективным представляется исследование поведения RM-метрики в задачах регрессии, генерации и сегментации, а также её адаптация к случаям мультиклассовой и многозадачной классификации. Ещё одним направлением может стать автоматическая настройка λ на основе свойств данных или архитектуры модели.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Glorot X., Bengio Y.* Understanding the difficulty of training deep feedforward neural networks, *Proceedings of the thirteenth international conference on artificial intelligence and statistics: JMLR Workshop and Conference Proceedings*, 2010, pp. 249-256.
2. *Fellicious C., Weissgerber T., Granitzer M.* Effects of random seeds on the accuracy of convolutional neural networks, *International Conference on Machine Learning, Optimization, and Data Science*. Cham: Springer International Publishing, 2020, pp. 93-102.
3. *Krizhevsky A. et al.* Learning multiple layers of features from tiny images, 2009.
4. Hu M. Y. et al. Latent state models of training dynamics, *arXiv preprint arXiv:2308.09543*, 2023.
5. *Picard D.* Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, *arXiv preprint arXiv:2109.08203*, 2021.
6. *Krizhevsky A. et al.* Learning multiple layers of features from tiny images, 2009.
7. *Deng J. et al.* Imagenet: A large-scale hierarchical image database, *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248-255.
8. Touvron H. et al. Augmenting convolutional networks with attention-based aggregation, *arXiv preprint arXiv:2112.13692*, 2021.
9. *Guo Z. et al.* A grey-box attack against latent diffusion model-based image editing by posterior collapse, *arXiv preprint arXiv:2408.10901*, 2024.
10. *Picard D.* Torch. manual_seed (3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision, *arXiv preprint arXiv:2109.08203*, 2021.
11. *Singhal V. et al.* BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis, *Nature genetics*, 2024, Vol. 56. No. 3, pp. 431-441.
12. *Chen S. et al.* Diffusiondet: Diffusion model for object detection, *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 19830-19843.
13. Wightman R., Touvron H., Jégou H. Resnet strikes back: An improved training procedure in timm //arXiv preprint arXiv:2110.00476. – 2021.
14. *Åkesson J., Töger J., Heiberg E.* Random effects during training: Implications for deep learning-based medical image segmentation, *Computers in Biology and Medicine*, 2024, Vol. 180, pp. 108944.
15. *Karthik S. et al.* If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection, *arXiv preprint arXiv:2305.13308*, 2023.
16. *Touvron H. et al.* Augmenting convolutional networks with attention-based aggregation, *arXiv preprint arXiv:2112.13692*, 2021.
17. *Long J., Shelhamer E., Darrell T.* Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.
18. *Wang Z., Yan W., Oates T.* Time series classification from scratch with deep neural networks: A strong baseline, *2017 International joint conference on neural networks (IJCNN*). IEEE, 2017, pp. 1578-1585.
19. *He K. et al.* Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
20. *Huang G. et al.* Densely connected convolutional networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

Доброхвалов Максим Олегович – Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина); e-mail: night1337bot@gmail.com; г. Санкт-Петербург, Россия; тел.: +78122342682; кафедра математического обеспечения и применения ЭВМ; ORCID: 0000-0002-0571-5836.

Филатов Антон Юрьевич – Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина); e-mail: aifilatov@etu.ru; г. Санкт-Петербург, Россия; тел.: +78122342682; к.т.н.; доцент кафедра математического обеспечения и применения ЭВМ; ORCID: 0000-0003-4298-8523.

Чегодаева Елизавета Александровна – Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина); e-mail: elizaveta.cheg@gmail.com; г. Санкт-Петербург, Россия; тел.: +78122342682; магистрант кафедры математического обеспечения и применения ЭВМ; ORCID: 0009-0005-9968-3108.

Dobrokhvalov Maksim Olegovich – Saint Petersburg Electrotechnical University "LETI"; e-mail: night1337bot@gmail.com; Saint Petersburg, Russia; phone: +78122342682; the Department of Software Engineering and Computer Applications; ORCID: 0000-0002-0571-5836.

Filatov Anton Yuryevich – Saint Petersburg Electrotechnical University "LETI"; e-mail: aifilatov@etu.ru; Saint Petersburg, Russia; phone: +78122342682; cand. of eng. sc.; associate professor Department of Software Engineering and Computer Applications; ORCID: 0000-0003-4298-8523.

Chegodaeva Elizaveta Aleksandrovna – Saint Petersburg Electrotechnical University "LETI"; e-mail: elizaveta.cheg@gmail.com; Saint Petersburg, Russia; phone: +78122342682; master's student of the Department of Software Engineering and Application of Computers; ORCID: 0009-0005-9968-3108.