Научная статья
УДК 004.732.056
https://doi.org/10.31854/1813-324X-2026-12-2-113-120
EDN:OETOUI

Синтез реалистичных синтетических данных с помощью диффузионной модели TabDDPM в задачах обнаружения сетевых атак

Олег Иванович Шелухин, sheluhin@mail.ru
Фёдор Андреевич Маторин✉, f.a.matorin@mtuci.ru

Московский технический университет связи и информатики,
Москва, 111024, Российская Федерация

Аннотация

Актуальность. *Для исследований и тестирования средств обнаружения атак требуются репрезентативные наборы сетевого трафика, однако их сбор и разметка трудозатратны, а распространение ограничено требованиями конфиденциальности и рисками утечек. Синтетические данные позволяют увеличить объем выборок и моделировать редкие сценарии и сценарии «нулевого дня» при сохранении статистических свойств сетевого трафика.*

Цель: *повышение качества и воспроизводимости формирования синтетических табличных признаков сетевого трафика на примере Android-приложений за счет применения диффузионной модели Tabular Denoising Diffusion (TabDDPM) и комплексной валидации результатов генерации на согласованном наборе метрик.*

Методы. *Использована диффузионная генеративная модель TabDDPM, применимая к произвольным табличным наборам данных. Эффективность генерации оценивается методами статистического анализа: сравнение распределений и зависимостей признаков, проверка полезности в прикладной задаче, а также оценка отличия синтетических данных от реальных.*

Результат. *Выполнен комплексный анализ качества TabDDPM при генерации табличных признаков сетевого трафика атак или нежелательных приложений. Показана возможность формирования контролируемых синтезированных наборов данных, сохраняющих характерные паттерны трафика и обеспечивающих масштабирование обучающих выборок без прямого копирования исходных записей.*

Новизна. *Предложен согласованный протокол постгенерационной валидации синтетического трафика, сочетающий метрики реалистичности, полезности и неотличимости, что снижает риск некорректных выводов при разрозненной оценке. Для количественной оценки качества генерации синтезируемых данных предложен интегральный показатель качества на основе партициальных метрик.*

Теоретическая значимость *состоит в развитии методического подхода к верификации табличных диффузионных моделей для задач кибербезопасности.*

Практическая значимость *заключается в возможности применения получаемых синтетических наборов данных для моделирования компьютерных атак, сценариев «нулевого дня», стресс-тестирования и обучения / проверки систем обнаружения вторжений.*

Ключевые слова: *машинное обучение, генеративно-состязательные сети, TabDDPM, диффузионные модели, синтезированный трафик, Android-приложения, табличные данные, метрики, обнаружение вторжений*

Ссылка для цитирования: Шелухин О.И., Маторин Ф.А. Синтез реалистичных синтетических данных с помощью диффузионной модели TabDDPM в задачах обнаружения сетевых атак // Труды учебных заведений связи. 2026. Т. 12. № 2. С. 113–120. DOI:10.31854/1813-324X-2026-12-2-113-120. EDN:OETOUI

Original research
https://doi.org/10.31854/1813-324X-2026-12-2-113-120
EDN:OETOUI

Realistic Synthetic Data Generation Using the TabDDPM Diffusion Model for Network Attack Detection

Oleg I. Sheluhin, sheluhin@mail.ru
Fedor A. Matorin✉, f.a.matorin@mtuci.ru

Moscow Technical University of Communications and Informatics,
Moscow, 111024, Russian Federation

Annotation

Relevance. *Representative network-traffic datasets are required for research and testing of attack detection tools; however, their collection and annotation are labor-intensive, and data sharing is constrained by confidentiality requirements and the risk of leakage. Synthetic data can increase sample sizes and enable modeling of rare and zero-day scenarios while preserving the statistical properties of network traffic.*
Objective. *To improve the quality and reproducibility of generating synthetic tabular features of network traffic, using Android applications as a case study, by applying the Tabular Denoising Diffusion model (TabDDPM) and performing comprehensive validation of the generated data using a consistent set of metrics.*
Methods. *We employ the TabDDPM diffusion-based generative model, which is applicable to arbitrary tabular datasets. Generation performance is assessed via statistical analysis methods, including comparisons of feature distributions and inter-feature dependencies, evaluation of utility in a downstream task, and estimation of the discrepancy between synthetic and real data.*
Results. *A comprehensive quality assessment of TabDDPM is conducted for generating tabular features of network traffic associated with attacks or unwanted applications. The results demonstrate the feasibility of producing controlled synthetic datasets that preserve characteristic traffic patterns and enable scaling of training samples without directly copying the original records.*
Novelty. *We propose a unified post-generation validation protocol for synthetic traffic that integrates realism, utility, and indistinguishability metrics, thereby reducing the risk of misleading conclusions arising from fragmented evaluation. In addition, an integral quality indicator is introduced to quantify generation performance by aggregating partial metrics.*
The theoretical significance *lies in advancing a methodological framework for verifying tabular diffusion models in cybersecurity applications.*
The practical significance *is the ability to use the resulting synthetic datasets to model cyberattacks and zero-day scenarios, perform stress testing, and train and / or evaluate intrusion detection systems.*

Keywords: *machine learning, generative adversarial network; TabDDPM, diffusion models, synthetic traffic, Android applications, tabular data, metrics, intrusion detection*

For citation: Sheluhin O.I., Matorin F.A. Realistic Synthetic Data Generation Using the TabDDPM Diffusion Model for Network Attack Detection. *Proceedings of Telecommunication Universities.* 2026;12(2):113–120. (in Russ.) DOI:10.31854/1813-324X-2026-12-2-113-120. EDN:OETOUI

Введение

В условиях, когда получение большого объема реального сетевого трафика и сетевых атак сопряжено с юридическими, техническими и этическими трудностями, особую актуальность приобретают методы, позволяющие искусственно расширить обучающее множество без ухудшения его качества и репрезентативности. Генерация синтезированных данных является мощным и гибким подходом и заключается в создании новых, ранее не существовавших примеров на основе генеративных моделей.

Наибольшую известность получили несколько методов создания синтетических данных. Под «синтетическими» понимаются искусственно созданные данные, имитирующие реальные наблюдения и используемые для подготовки моделей машинного обучения, в условиях, когда получение реальных данных невозможно из-за сложности или дороговизны. Так, для увеличения объема обучающих данных и повышения эффективности машинного обучения используется метод аугментации данных [1], направленный на создание синтезированных данных на основе существующих наборов данных [2]. При использовании этого метода увеличивается как количество, так и разнообразие данных, доступных для обучения и тестирования моделей, устраняя необходимость в сборе новых данных. Увеличение объема данных может быть достигнуто либо за счет их генерации «с нуля» путем обучения генератора (например, с помощью генеративно-состязательных нейронных сетей (GAN, *аббр. от англ*. Generative Adversarial Nets) [3]), либо путем применения к имеющимся выборкам заранее определенного набора преобразований [4]. Оба подхода повышают эффективность моделей глубокого обучения за счет предоставления более разнообразного и обширного набора данных.

В контексте аугментации табличных данных в работах [5] с помощью GAN [6] и вариационных автокодировщиков (VAE*, аббр. от англ.* Variational Autoencoder) [7] используются разные методологии. Так, GAN хорошо справляются с формированием сложных распределений, что делает их высокоэффективными для создания реалистичных табличных данных. Однако процесс их обучения часто нестабилен и требует тщательной состязательной настройки. Напротив, VAE обеспечивают стабильное обучение с помощью своей вероятностной структуры и полезны при изучении скрытых представлений, позволяя интерполировать и исследовать данные. Несмотря на эти преимущества, как правило VAE генерируют менее четкие или реалистичные данные по сравнению с GAN, а балансировка возникающих потерь при реконструкции с регуляризацией остается сложной задачей.

В последние годы разработаны диффузионные модели [8], которые стали мощным инструментом для создания синтетических данных в условиях их дефицита. Диффузионные модели в задачах обнаружения атак (в т. ч. кибератак) применяются как генеративные модели для выявления аномалий. Их ключевая идея – научиться «восстанавливать» нормальные данные и фиксировать значительные отклонения при попытке реконструкции аномальных паттернов. В отличие от GAN и VAE диффузионные модели, также известные как вероятностные диффузионные модели или генеративные модели, основаны на оценке. Они опираются на теоретическую базу для стабильного обучения и позволяют генерировать высококачественные данные, устраняя некоторые ограничения генеративно-состязательных и вариативно-генеративно-состязательных моделей. Однако это преимущество достигается за счет более высоких требований к вычислительным ресурсам.

Табличную диффузионную вероятностную модель шумоподавления (TabDDPM, *аббр. от англ.* Tabular Denoising Diffusion Probabilistic Model) [9] можно универсально применять к любому табличному набору данных, работающему со всеми типами признаков. Она использует полиномиальную диффузию для категориальных и бинарных, а также гауссову диффузию для числовых признаков. Модель TabDDPM эффективно работает со смешанными типами данных и стабильно генерирует высококачественные синтетические данные.

Целью работы является генерация реалистичного синтетического сетевого трафика Android-приложений, представленного в виде нормализованных числовых признаков, с использованием генеративной диффузионной модели TabDDPM, адаптированной для табличных данных и комплексный анализ качества синтезированного трафика с помощью согласованного набора метрик.

Для достижения цели необходимо решить следующие задачи:

– выбрать исходный набор данных сетевого трафика Android-приложений и выполнить его предварительную обработку для обучения модели;

– обучить генеративную диффузионную модель TabDDPM на подготовленных данных и подобрать оптимальные гиперпараметры для улучшения качества синтетических выборок;

– оценить степень сходства между распределениями сгенерированного и реального трафика.

Обеспечение корректности и контролируемости генерации

Одной из важнейших задач при разработке генеративных моделей, предназначенных для синтеза сетевого трафика, является обеспечение корректности и воспроизводимости результатов генерации. Корректность означает, что выходные данные соответствуют логике и структуре реального трафика, не содержат аномалий, противоречий или артефактов. Под контролируемостью понимается, что генерация управляется заранее заданными параметрами и может быть повторена в идентичных условиях с предсказуемыми результатами.

Перед генерацией модель обучается на очищенном и строго структурированном датасете, в котором каждый признак имеет четко определенный тип, диапазон значений и статистическую значи-

мость. На этапе вывода модель использует идентичную структуру признаков, что гарантирует соответствие синтезированных записей формату исходных данных. Взаимно зависимые признаки (например, протокол и порт, флаги и тип соединения) обрабатываются с учетом их совместного распределения вероятности, что позволяет избежать появления комбинаций, не соответствующих установленным в системе правилам обработки.

Последующая валидация результатов генерации – сгенерированный пример проверяется по нескольким критериям:

– тип данных каждого признака должен соответствовать ожидаемому (например, целочисленные порты, категориальные метки протокола, вещественные интервалы);

– далее контролируется попадание значений в допустимые диапазоны, зафиксированные при анализе обучающего датасета;

– признаки, обладающие логической связью (например, TCP-флаг SYN и поле *init_win_bytes_forward*), проверяются на предмет совместимости, что позволяет исключить аномальные комбинации;

– отсутствие пропущенных или поврежденных значений данных (NaN, *аббр. от англ.* Not a Number) гарантируется за счет применения маскирования и восстановления их после нормализации.

Каждая процедура генерации характеризуется набором параметров, которые могут быть зафиксированы заранее или переданы через конфигурационный файл. Ключевыми из них являются: размер латентного вектора; объем выходных данных; класс генерируемого трафика (например, «Normal», «DDoS», «Botnet»); версия модели (номер эпохи, весовой чекпоинт); фиксированное значение Random Seed, обеспечивающее детерминированность результата.

Это дает возможность не только гибко управлять характеристиками выходных данных, но и гарантирует воспроизводимость эксперимента – важную для научных исследований и тестирования систем обнаружения вторжений.

Структура TabDDPM

Табличные наборы данных часто ограничены в объеме из-за проблем с конфиденциальностью при их сборе. TabDDPM может создавать новые синтетические данные без ущерба для конфиденциальности. Общая структура модели представлена на рисунке 1. Числовые и категориальные данные обычно представляются двумя ветвями: квантильным преобразователем для числовых данных и однократным кодированием для категориальных данных. Эти новые представления данных затем передаются на вход диффузионной вероятностной модели шумоподавления, реализованной на основе многослойных перцептронов (MLP, *аббр. от англ.* Multilayer Perceptron) для минимизации двух типов потерь L_{num} и L_{cat} с помощью функции Softmax. Блок One Hot кодирования осуществляет преобразование категориальных данных в числовой формат, при котором каждая уникальная категория становится отдельным бинарным признаком, принимающим значение 1 или 0.

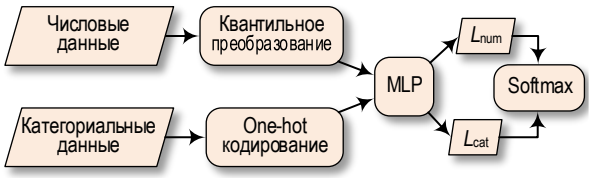

Рис. 1. Общая структура TabDDPM

Рис. 1. General Structure of TabDDPM

Диффузионная модель TabDDPM состоит из двух основных компонентов: прямого и обратного процесса [9]. Прямой процесс (Forward Process) заключается в постепенном добавлении шума удовлетворяющего распределению $q(x_{1:T}|x_0)$ к исходному образцу x_0 из распределения данных $q(x_0)$:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}).$$

Шум выбирается из заранее заданных распределений $q(x_t|x_{t-1})$ с дисперсиями $\{\beta_1, \dots, \beta_T\}$ до тех пор, пока итоговое распределение не станет неотличимым от случайного шума.

Обратный процесс (Backward Process), удовлетворяющий распределению $p(x_{0:T})$, постепенно устраняет шум в скрытой переменной $x_T \sim q(x_T)$ и позволяет порождать новые выборки данных из $q(x_0)$:

$$p(x_{0:T}) = \prod_{t=1}^{T} p(x_{t-1}|x_t).$$

Начиная с чистого шума, модель постепенно устраняет шумовую компоненту и тем самым восстанавливает по скрытым переменным $x_1, \dots, x_T$ исходные «чистые» $x_0 \sim q(x_0)$.

Совместное распределение $p_\theta(x_{0:T})$ называется обратным процессом и задается как цепь Маркова с обучаемыми гауссовыми переходами, начинающаяся с $p(x_T) = \mathcal{N}(x_T; 0, I)$:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t). \quad (1)$$

Отличие диффузионных моделей от других типов со скрытыми переменными состоит в том, что приближенное апостериорное распределение $q(x_{1:T}|x_0)$, называемое прямым процессом или диффузионным процессом, фиксируется как цепь Маркова, которая

постепенно добавляет к данным гауссов шум согласно расписанию дисперсий $\beta_1, \dots, \beta_T$:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}). \quad (2)$$

Гауссовские диффузионные модели работают в непрерывных пространствах $(x_t|\mathbb{R}^n)$, где прямой и обратный процессы задаются нормальными распределениями:

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \quad (3)$$

$$q(x_T) = N(x_T; 0, I), \quad (4)$$

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right). \quad (5)$$

Необходимо выбрать дисперсии β_t прямого процесса, а также архитектуру модели и параметризацию гауссовского распределения для обратного процесса. Дисперсии β_t могут быть получены либо с помощью повторной параметризации [10], либо зафиксированы как гиперпараметры. В [11] предлагают использовать диагональную матрицу $\Sigma_\theta(x_t, t)$ с постоянной β_t и вычислять среднее значение $\mu_\theta(x_t, t)$ как функцию от x_t и $\epsilon_\theta(x_t, t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right), \quad (6)$$

где $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i \le t}\alpha_i$, а $\epsilon_\theta(x_t, t)$ предсказывает «истинную» шумовую компоненту ϵ для зашумленного образца данных x_t.

На практике целевая функция может быть представлена в виде суммы среднеквадратичных ошибок между предсказанным шумом $\epsilon_\theta(x_t, t)$ и истинным шумом ϵ по всем временны́м шагам t:

$$L_t^{\text{simple}} = \mathbb{E}_{x_0, \epsilon, t} \parallel \epsilon - \epsilon_\theta(x_t, t) \parallel_2^2. \quad (7)$$

Исходный набор данных сетевого трафика Android-приложений

Генерация реалистичного сетевого трафика Android-приложений проводилась на примере экспериментальных данных мобильных приложений, приведенных в работах [12, 13]. Для формирования обучающей и тестовой выборок на мобильных устройствах под управлением ОС Android осуществлялся сбор необработанных данных сетевого трафика в виде IP-пакетов. Трафик различных типов мобильных приложений характеризовался набором из M = 23 атрибутов. Общее число экспериментально измеренных потоков каждого приложения составляло K = 5000.

В исследовании использовался подготовленный датасет сетевого трафика Android-приложений [12, 13], содержащий N = 90777 записей (сетевых сессий) и d = 21 числовых признаков. Каждый образец был помечен идентификатором приложения $y = app\_id$, сгенерировавшего данный трафик. Для оценки качества генерации синтетического трафика рассматривались три из 18 уникальных приложения из датасета: *Mail.ru* (*app_id* = 9), *SberMobile* (*app_id* = 12) и 4*PDA* (*app_id* = 2). Для каждого выбранного приложения формировалась модель TabDDPM с отдельной обучающей выборкой, аппроксимирующей эмпирическое распределение признаков данного класса.

На первом этапе к исходным данным применялась нормализация признаков, поскольку значения исходных числовых атрибутов имеют разную природу (размеры и количества пакетов, байтов, соотношения и т. д.) и характеризуются сильной вариативностью и асимметрией распределений.

На втором этапе производилось ограничение выбросов (*clip*). С этой целью для числовых признаков с наибольшим размахом вычисляются пороги $P_{0,99}(j)$ – 99-й перцентиль по обучающей выборке. Все значения, превышающие $P_{0,99}(j)$, заменялись пороговым значением: $x_j := \min(x_j, P_{0,99}(j))$.

На третьем этапе производилось масштабирование с помощью квантильного нормализатора $QT: \mathbb{R} \to [0,1]$, переводящего распределение каждого признака в равномерное на отрезке [0,1]. Формально для каждого признака j значение x_j заменялось на $u_j = F_j(x_j)$, где F_j – эмпирическая функция распределения данного признака на обучающей выборке.

Для построения и оценки модели датасет разбивался на обучающую, валидационную и тестовую выборки. Разделение производилось по классам для сохранения пропорций редких и частых классов. В проведенных экспериментах для обучения выделялось 80 % наблюдений, 10 % для подбора гиперпараметров и 10 % для финального тестирования генератора.

На основе особенностей используемых численных данных и метода их предобработки структурная схема модели TabDDPM может быть представлена в виде, представленном на рисунке 2.

Модель обучается путем минимизации суммы среднеквадратичной ошибки для гауссовой диффузионной части:

$$L_{\text{MSE}} = \frac{1}{M}\sum_{i=1}^{M} \parallel \hat{\epsilon}_\theta\left(x_{t_i}^{(i)}|t_i\right) - \epsilon^{(i)} \parallel^2. \quad (8)$$

На вход модели подается вектор $\{x_1, x_2, \dots, x_n\}$ числовых признаков сессии одного из приложений. Блок Quantile Transformer по каждому признаку выполняет масштабирование к диапазону [–1,1]. На выходе блока получаем нормализованный тензор x_{norm} размерности *batch×features*. После указанного преобразования данные попадают в скоринговую сеть диффузионной модели TabDDPM, в качестве которой используется MLP, формирующий оценку добавленного шума ε той же размерности.

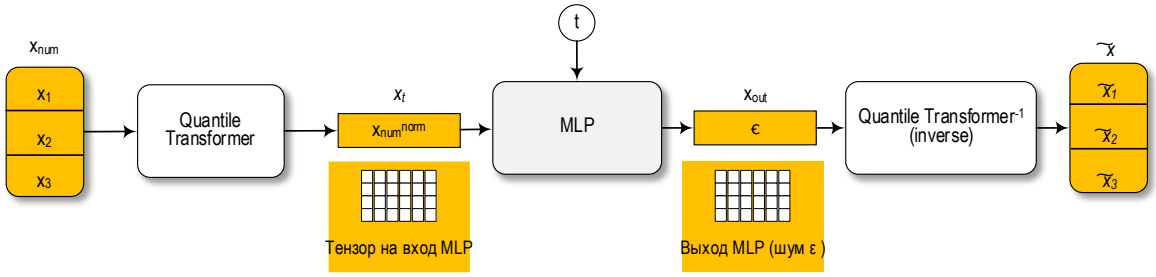

Рис. 2. Структурная схема модели TabDDPM

Рис. 2. Structural Diagram of the TabDDPM

На заключительном этапе выполняется обратное к исходному масштабу преобразование в блоке Quantile Transformer^{-1}. В результате на выходе модели формируется вектор синтезированных данных $\{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n\}$, сопоставимый по масштабу с реальными данными.

Оценка качества синтезированных данных

Для количественной оценки качества сходства синтезированных выборок, порожденных генеративной моделью TabDDPM и выбранного приложения реального трафика, использовался набор метрик, представленный в таблице 1.

ТАБЛИЦА 1. Метрики оценки сходства распределений

ТАБЛИЦА 1. Metrics for Assessing Distribution Similarity

№ п/п	Метрика	Формула	
1	Kolmogorov–Smirnov (KS)	$D_j = \sup_x \lvert F_j(x) - \hat{F}_j(x)\rvert$ $KS = \frac{1}{d}\sum D_j$	(9)
2	First-Order Wasserstein Distance (WD)	$WD = \frac{1}{d}\sum_{j=1}^{d}\int_0^1 \lvert F_j^{-1}(u) - \hat{F}_j^{-1}(u)\rvert\, du$	(10)
3	Reference Range Coverage (Coverage)	$\mathrm{cov}_j(W_k) = \frac{1}{\lvert W_k\rvert}\sum_{x\in W_k} 1\{F_j^{-1}(0{,}01) \le x^{(j)} \le F_j^{-1}(0{,}99)\}$	(11)
		$\mathrm{Coverage}(W_k) = \frac{1}{d}\sum_{j=1}^{d}\mathrm{cov}_j(W_k)$	(12)
4	Correlation Consistency (ΔCorr)	$\Delta\mathrm{Corr} = \frac{2}{d(d-1)}\sum_{1\le j<k\le d}\lvert \rho_{jk}^{(\mathrm{real})} - \rho_{jk}^{(\mathrm{synth})}\rvert$	(13)
5	Area Under the ROC Curve (AUC)	$AUC = \int_0^1 TPR(FPR^{-1}(t))dt$	(14)

В формулах, представленных в таблице 1, обозначениями $F_j(x)$ и $\hat{F}_j(x)$ представлены кумулятивные функции распределения реальных и синтезированных данных:

$F_j^{-1}(u)$ и $\hat{F}_j^{-1}(u)$ – функции распределения обратные к $F_j(x)$ и $\hat{F}_j(x)$;

$\rho_{jk}^{(\mathrm{real})}$ и $\rho_{jk}^{(\mathrm{synth})}$ – коэффициенты корреляции Пирсона между j-м и k-м признаками для реальных и синтетических данных;

В ходе экспериментов с генеративной моделью TabDDPM для всех метрик, представленных в таблице 1, и каждого рассматриваемого приложения были получены конкретные числовые значения, приведенные в таблице 2. Представленные численные значения метрик подтверждают, что генерация синтезированных с помощью модели TabDDPM данных воспроизводит как масштаб и форму распределений, так и существенную часть межпризнаковых зависимостей. Это позволяет применять синтез выбранных приложений для дальнейшей настройки и валидации детекторов, ориентированных на появление неизвестных классов.

ТАБЛИЦА 2. Сводная таблица метрик по приложениям

TABLE 2. Summary of Metrics by Application

app_id	AUC	KS	WD	ΔCorr	Coverage
2	0,6778	0,0520	0,0287	0,0485	1
9	0,6141	0,0501	0,0242	0,0433	1
12	0,5914	0,0658	0,0329	0,0679	0,9997

Каждая из представленных метрик покрывает «слепые зоны» других метрик. Так, метрики KS / WD служат для оценки сходства одномерных распреде-

лений и сдвиги массы, а ΔCorr оценивает многомерные связи. Метрика Coverage оценивает охват значений, а AUC – практическую различимость. Каждая из применяемых метрик AUC, Coverage, ΔCorr, KS и WD характеризует свой индивидуальный, принципиально отличный аспект качества генерации, синтезируемых данных с помощью модели TabDDPM и потому не является взаимозаменяемым. Анализ результатов моделирования показывает, что синтезированные данные почти полностью совпадают с реальными, поскольку центры и масштабы кластеров близки, а локальные окрестности идентичны. Незначительные отличия касаются главным образом глобального положения центров у приложений SberMobile и частично у 4PDA. При этом локальная структура соседств остается на уровне 95–98 %, что подтверждает корректную генерацию характерных паттернов трафика с помощью модели TabDDPM.

Учитывая разнонаправленный характер изменения величины указанных метрик, для характеристики индивидуальных особенностей перечисленных критериев предлагается ввести интегральный показатель качества синтезированных данных на основе модифицированных метрик, приведенных в таблице 3. Количественно эффективность моделирования может быть оценена показателем $\sum K_i$.

ТАБЛИЦА 3. Сводная таблица модифицированных метрик по приложениям

TABLE 3. Summary Table of Modified Metrics by Application

app_id	*K*1	*K*2	*K*3	*K*4	*K*5	$\sum K_i$
2	0,6444	0,948	0,971	0,951	1	4,192
9	0,7718	0,949	0,976	0,957	1	4,268
12	0,8172	0,934	0,967	0,922	0,999	4.232

Примечание: *K*2 = 1 – *KS*; *K*3 = 1 – *WD*; *K*4 = 1 – ΔCorr; *K*5 = Coverage

Для визуализации введенных метрик, характеризующих качественные показатели данных для различных типов анализируемых приложений и разработанного метода моделирования синтезированных данных, использовалась лепестковая диаграмма, представленной на рисунке 3. Для удобства метрика AUC дискриминатора, для которого целевым является значение 0,5, была преобразована в «индекс неотличимости», принимающий значение, равное 1 при AUC = 0,5, и убывающий к 0 по мере роста различимости: $K1 = 1 - 2 \mid AUC - 0{,}5 \mid$.

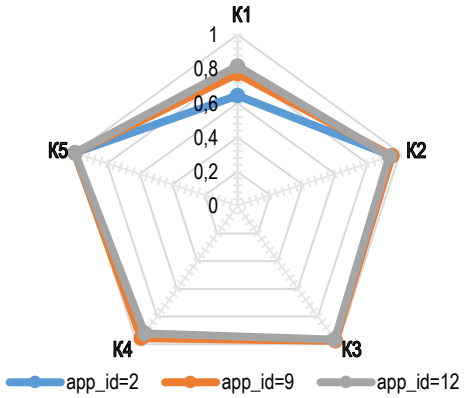

Рис. 3. Лепестковая диаграмма метрик для исследуемых приложений

Fig. 3. Radar Chart of Metrics for the Applications under Study

Заключение

Достоинством диффузионных моделей является возможность детализированной реконструкции данных и способности выявлять редкие аномалии. Представленные метрики подтверждают, что генерация синтезированных данных с помощью модели TabDDPM воспроизводит как масштаб и форму распределений, так и существенную часть межпризнаковых зависимостей. Это позволяет применять синтез выбранных приложений для дальнейшей настройки и валидации детекторов, ориентированных на появление неизвестных классов.

Каждая из представленных метрик покрывает «слепые зоны» других метрик. Так, метрики KS / WD служат для оценки сходства одномерных распределений и сдвиги массы, а ΔCorr оценивает многомерные связи. Метрика Coverage оценивает охват значений, а AUC – практическую различимость. Каждая из применяемых метрик AUC, Coverage, ΔCorr, KS и WD характеризует свой индивидуальный, принципиально отличный аспект качества генерации, синтезируемых данных с помощью модели TabDDPM и потому не являются взаимозаменяемыми. Для количественной оценки качества генерации синтезируемых данных предложено ввести в рассмотрение интегральной показатель качества на основе парциальных метрик.

Показано, что с помощью лепестковых диаграмм можно наглядно визуализировать интегральный показатель качества синтезируемых данных с помощью модели TabDDPM.

Список источников

1. Ding J., Li X., Kang X., Gudivada V.N. A Case Study of the Augmentation and Evaluation of Training Data for Deep Learning // Journal of Data and Information Quality. 2019. Vol. 11. Iss. 4. PP. 1–22. DOI:10.1145/3317573

2. Bansal A., Sharma R., Kathuria M. A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications // ACM Computing Surveys. 2022. Vol. 54. Iss. 10s. PP. 1–29. DOI:10.1145/3502287

3. Alqahtani H., Kavakli-Thorne M., Kumar G. Applications of Generative Adversarial Networks (GANs): An Updated Review // Archives of Computational Methods in Engineering. 2021. Vol. 28. Iss. 2. PP. 525–552. DOI:10.1007/s11831-019-09388-y. EDN:FTPNOL

4. Cubuk E.D., Zoph B., Mané D., Vasudevan V., Le Q.V. AutoAugment: Learning Augmentation Strategies From Data // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, Long Beach, USA, 2019). IEEE, 15–20 June 2019). PP. 113–123. DOI:10.1109/CVPR.2019.00020

5. Cui L., Li H., Chen K., Shou L., Chen G. Tabular data augmentation for machine learning: Progress and prospects of embracing generative AI // arXiv preprint arXiv:2407.21523. 2024. DOI:10.48550/arXiv.2407.21523
6. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., et al. Generative adversarial networks // Communications of the ACM. 2020. Vol. 63. Iss. 11. PP. 139–144. DOI:10.1145/3422622. EDN:SESCXD
7. Kingma D.P., Welling M. An Introduction to Variational Autoencoders // Foundations and Trends in Machine Learning. 2019. Vol. 12. Iss. 4. PP. 307–392. DOI:10.1561/2200000056
8. Yang L., Zhang Z., Song Y., Hong S., Xu R., Zhao Y., al. Diffusion Models: A Comprehensive Survey of Methods and Applications // ACM Computing Surveys. 2024. Vol. 56. Iss. 4. PP. 1–39. DOI:10.1145/3626235
9. Kotelnikov A., Baranchuk D., Rubachev I., Babenko A. TabDDPM: Modelling tabular data with diffusion models // arXiv preprint arXiv:2209.15421. 2022. DOI:10.48550/arXiv.2209.15421
10. Kingma D.P., Welling M. Auto-encoding variational Bayes // arXiv preprint arXiv:1312.6114. 2022. DOI:10.48550/arXiv.1312.6114
11. Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models // Advances in Neural Information Processing Systems 33 (NeurIPS). 2020.
12. Шелухин О.И., Маторин Ф.А. Снижение размерности массивов данных с помощью многослойных автокодировщиков в задаче классификации мобильных приложений // Труды учебных заведений связи. 2024. Т. 10. № 6. С. 111–120. DOI:10.31854/1813-324X-2024-10-6-111-120. EDN:TOPDUA
13. Шелухин О.И., Маторин Ф.А., Ванюшина А.В. Оценка свойств многослойных автокодировщиков в задаче обнаружения и классификации мобильных приложений // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16. № 6. С. 12–20. DOI:10.36724/2409-5419-2024-16-6-12-20. EDN:CRPLOR

References

1. Ding J., Li X., Kang X., Gudivada V.N. A Case Study of the Augmentation and Evaluation of Training Data for Deep Learning. *Journal of Data and Information Quality*. 2019;11(4):1–22. DOI:10.1145/3317573
2. Bansal A., Sharma R., Kathuria M. A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. *ACM Computing Surveys*. 2022;54(10s):1–29. DOI:10.1145/3502287
3. Alqahtani H., Kavakli-Thorne M., Kumar G. Applications of Generative Adversarial Networks (GANs): An Updated Review. *Archives of Computational Methods in Engineering*. 2021;28(2):525–552. DOI:10.1007/s11831-019-09388-y. EDN:FTPNOL
4. Cubuk E.D., Zoph B., Mané D., Vasudevan V., Le Q.V. AutoAugment: Learning Augmentation Strategies From Data. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 15–20 June 2019, Long Beach, USA. IEEE; 2019. p.113–123. DOI:10.1109/CVPR.2019.00020
5. Cui L., Li H., Chen K., Shou L., Chen G. Tabular data augmentation for machine learning: Progress and prospects of embracing generative AI. *arXiv preprint arXiv:2407.21523*. 2024. DOI:10.48550/arXiv.2407.21523
6. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., et al. Generative adversarial networks. *Communications of the ACM*. 2020;63(11):139–144. DOI:10.1145/3422622. EDN:SESCXD
7. Kingma D.P., Welling M. An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning*. 2019; 12(4):307–392. DOI:10.1561/2200000056
8. Yang L., Zhang Z., Song Y., Hong S., Xu R., Zhao Y., et al. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Computing Surveys*. 2024;56(4):1–39. DOI:10.1145/3626235
9. Kotelnikov A., Baranchuk D., Rubachev I., Babenko A. TabDDPM: Modelling tabular data with diffusion models. *arXiv preprint arXiv:2209.15421*. 2022. DOI:10.48550/arXiv.2209.15421
10. Kingma D.P., Welling M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*. 2022. DOI:10.48550/arXiv.1312.6114
11. Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems 33 (NeurIPS)*. 2020.
12. Sheluhin O.I., Matorin F.A. Reducing the Dimensionality of Data Arrays Using Multi-Layer Autoencoders in the Task of Classifying Mobile Applications. *Proceedings of Telecommunication Universities*. 2024;10(6):111–120. (in Russ.) DOI:10.31854/1813-324X-2024-10-6-111-120. EDN:TOPDUA
13. Sheluhin O.I., Matorin F.A., Vanyushina A.V. Evaluation of the properties of multilayer autoencoders in the task of detecting and classifying mobile applications. *H&ES Research*. 2024;16(6):12–20. (in Russ.) DOI:10.36724/2409-5419-2024-16-6-12-20. EDN:CRPLOR

Статья поступила в редакцию 30.01.2026; одобрена после рецензирования 19.02.2026; принята к публикации 24.02.2026
The article was submitted 30.01.2026; approved after reviewing 19.02.2026; accepted for publication 24.02.2026

Информация об авторах:

ШЕЛУХИН Олег Иванович — доктор технических наук, профессор, заведующий кафедрой «Информационная безопасность» Московского технического университета связи и информатики
https://orcid.org/0000-0001-7564-6744

МАТОРИН Фёдор Андреевич — аспирант кафедры «Информационная безопасность» Московского технического университета связи и информатики
https://orcid.org/0009-0002-4897-2338

Авторы сообщают об отсутствии конфликтов интересов.
The authors declare no conflicts of interests.