НОВЫЕ ВЫЗОВЫ
НОВЫЕ РЕШЕНИЯ

Риски мошеннических действий, использующих ИИ, и способы их минимизации

Килячков Анатолий Анатольевич, старший эксперт компании ООО «Б1 – консалт»

Килячков Александр Анатольевич, управляющий директор, ПАО «Сбербанк»

Декабрьские дебаты 2024,

18 декабря 2024 г.

Постановка задачи

- Согласно **Национальной стратегии** развития искусственного интеллекта (ИИ) к 2030 году:
  - **95% ведущих предприятий** должны внедрить технологии ИИ
  - **80% работников** должны иметь навыки работы с ИИ
- **Контрольные процедуры** являются важной частью технологических процессов предприятий и перспективным направлением для внедрения ИИ
- Однако, если ИИ обучен **на «инфицированных» данных**, то он выдаёт **неправильные результаты**, которые в отдельных случаях сложно выявить, т.е. **дипфейки**
- Специалистам по безопасности бизнеса следует **действовать на упреждение** и выявлять такие уязвимости в обучении и работе ИИ

Решаемая задача

- Решалась задача: «Каким образом можно **скрыть мошеннические транзакции** в контрольных процедурах, т.е. **создать дипфейк**, и , главное, понять, **как его можно выявить»**

- Для решения этой задачи **необходимо** было **понять**:
  - **как скрыть** от специалиста, готовящего базу данных, **«инфицированную»** обучающую информацию
  - **как «инфицировать»** обучающую информацию таким образом, чтобы **результаты** работы обученной нейронной сети **не вызывали подозрений** у специалистов контрольных подразделений
  - **как выявить** «инфицированные» обучающие данные

Содержание выступления

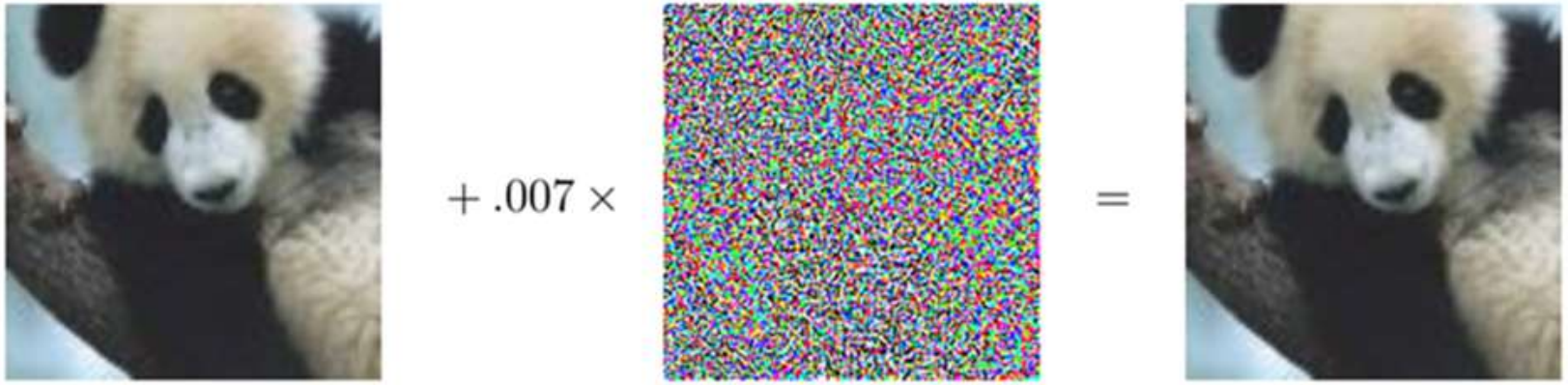
В сообщении рассмотрены:

1. **Примеры** чувствительности результатов работы нейронной сети (НС) к качеству обучающих данных
2. **Архитектура** нейронной сети, используемая для выявления мошеннических сделок
3. **Оценка чувствительности** предложенной НС к качеству обучающей информации
4. **Обнаруженная уязвимость** нейронной сети
5. **Способ выявления** попыток **«инфицирования»** обучающих данных, которые используют обнаруженную уязвимость

НОВЫЕ ВЫЗОВЫ
НОВЫЕ РЕШЕНИЯ

Чувствительность нейронной сети к качеству обучающей информации

Визуальные дипфейки



Панда,

достоверность

распознавания 57,7%

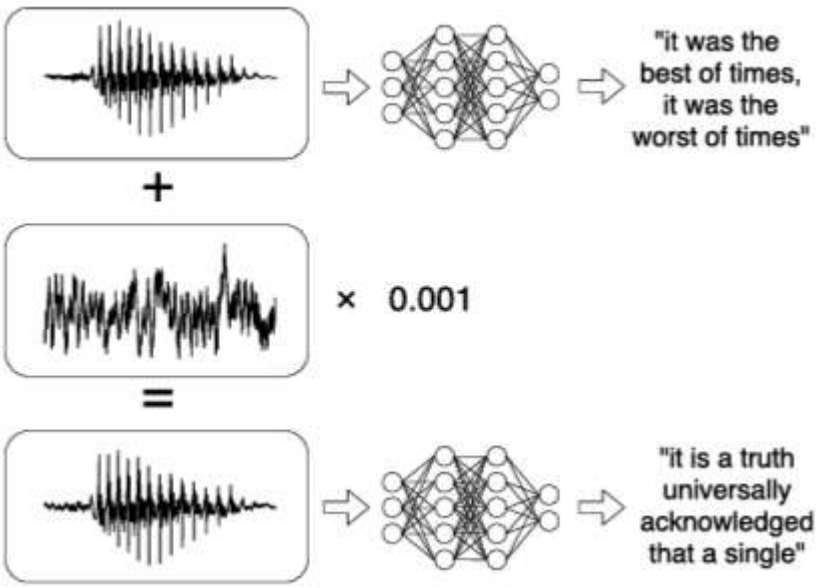
Гиббон,

достоверность

распознавания 99,3%

https://habr.com/ru/companies/vk/articles/348140/

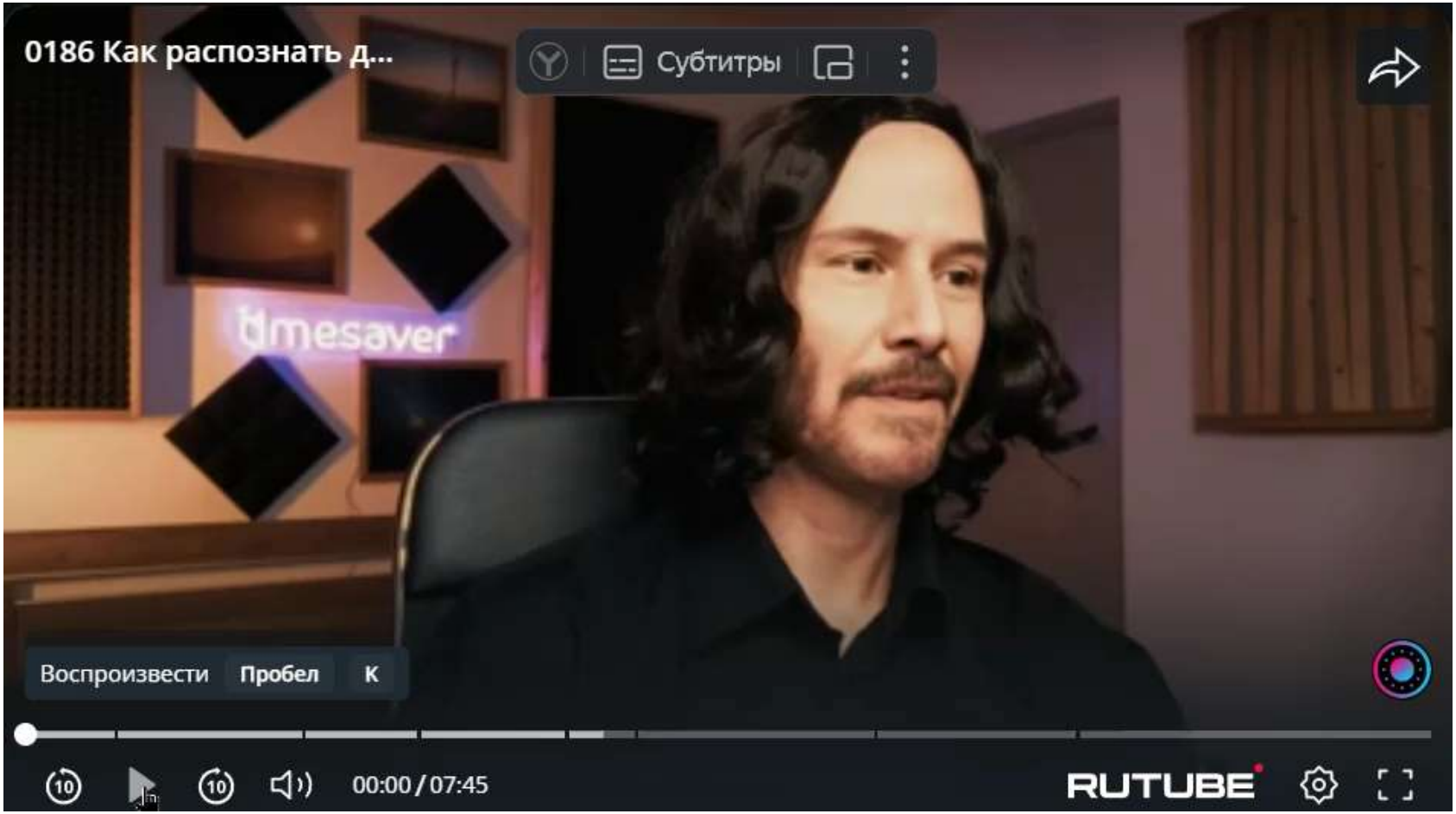
Аудио дипфейки



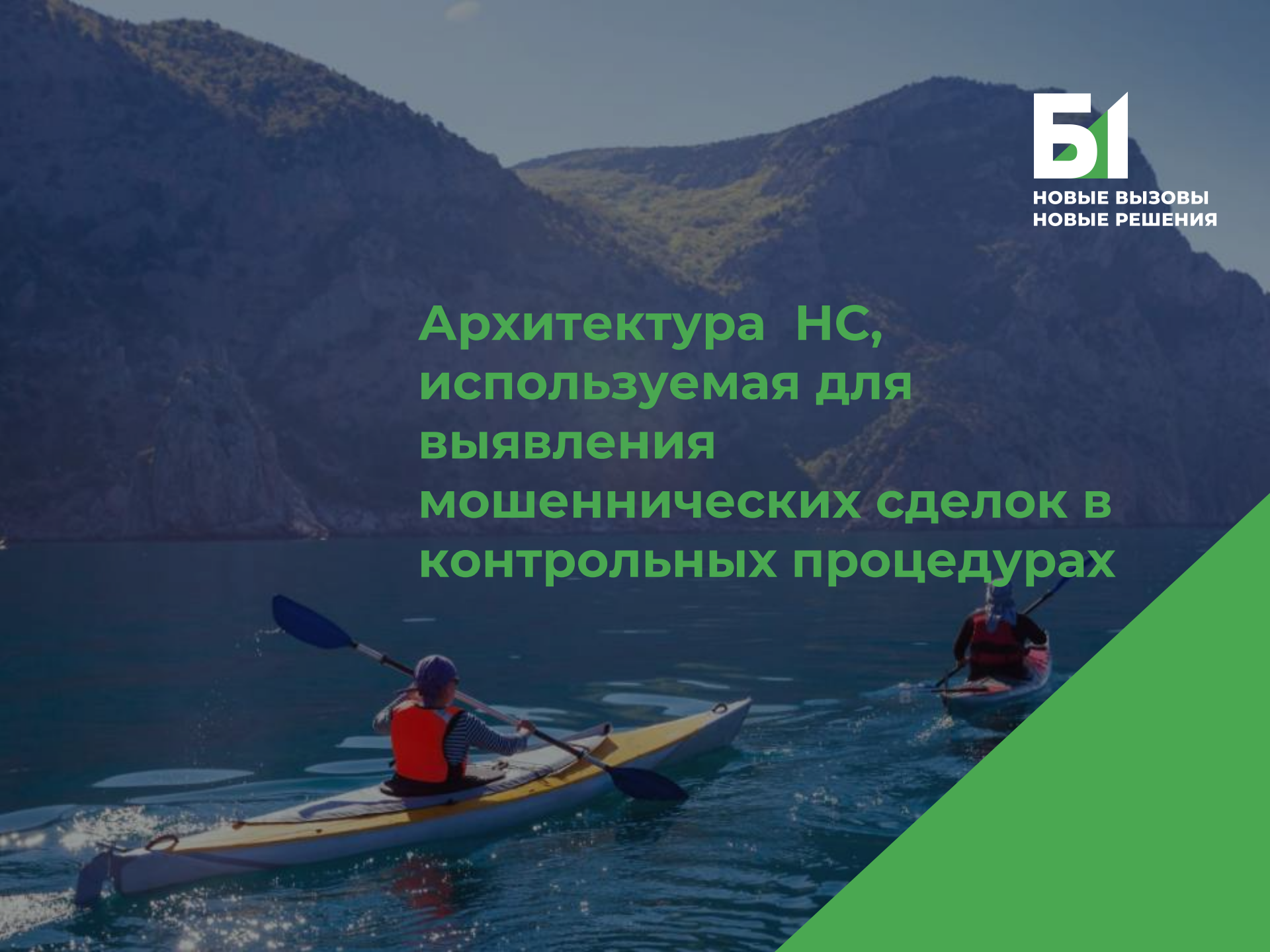
В 2018 г. исследователи из университета Беркли (Nicholas Carlini и David Wagner) показали, как можно взламывать голосовые помощники. Они отправляли голосовым помощникам набор звуков, находящихся за пределами человеческой слышимости, и тайно активировали системы ИИ на смартфонах

https://arxiv.org/pdf/1801.01944

Видео дипфейки



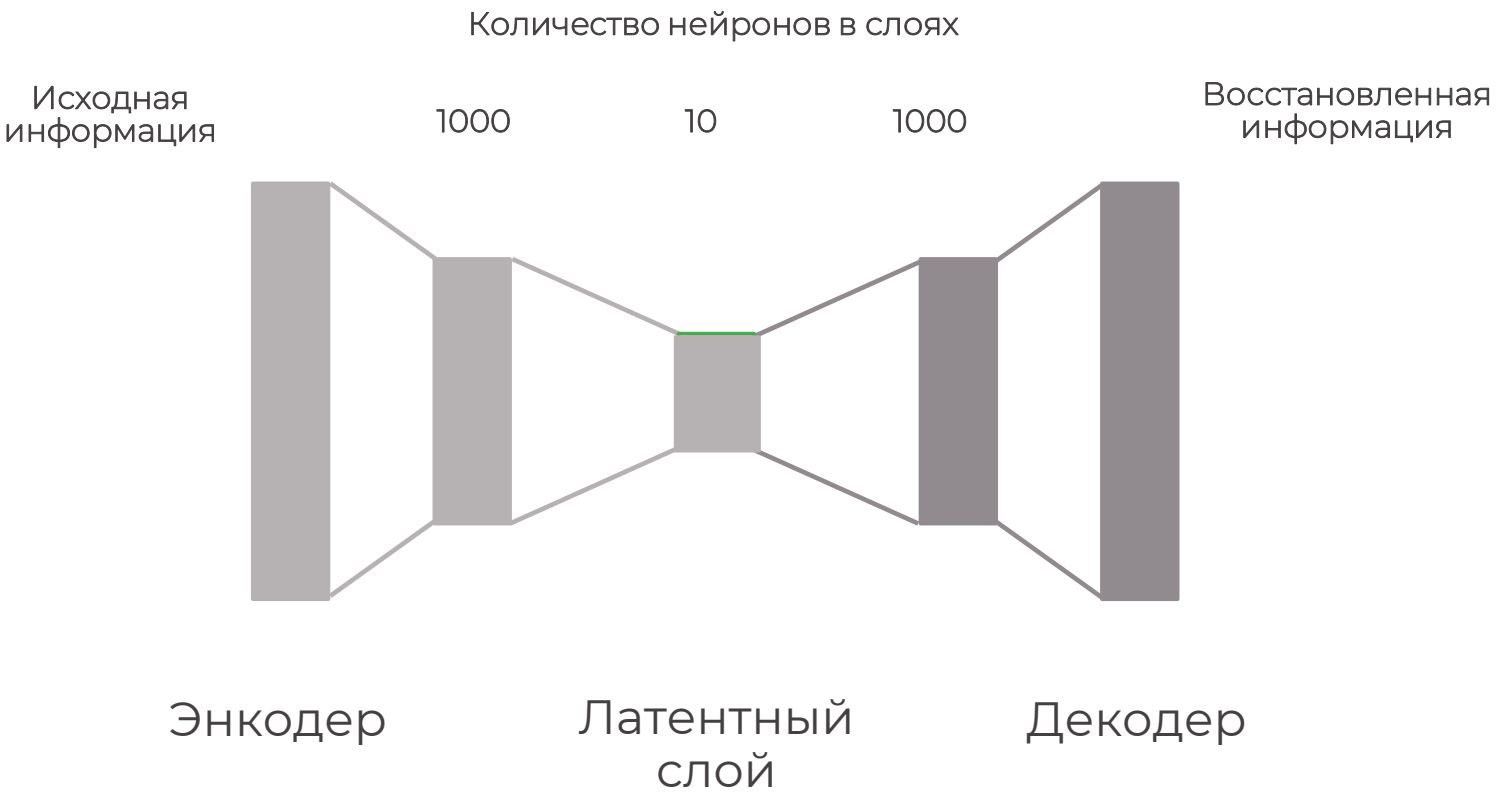
https://yandex.ru/video/preview/16523450493572788486



НОВЫЕ ВЫЗОВЫ
НОВЫЕ РЕШЕНИЯ

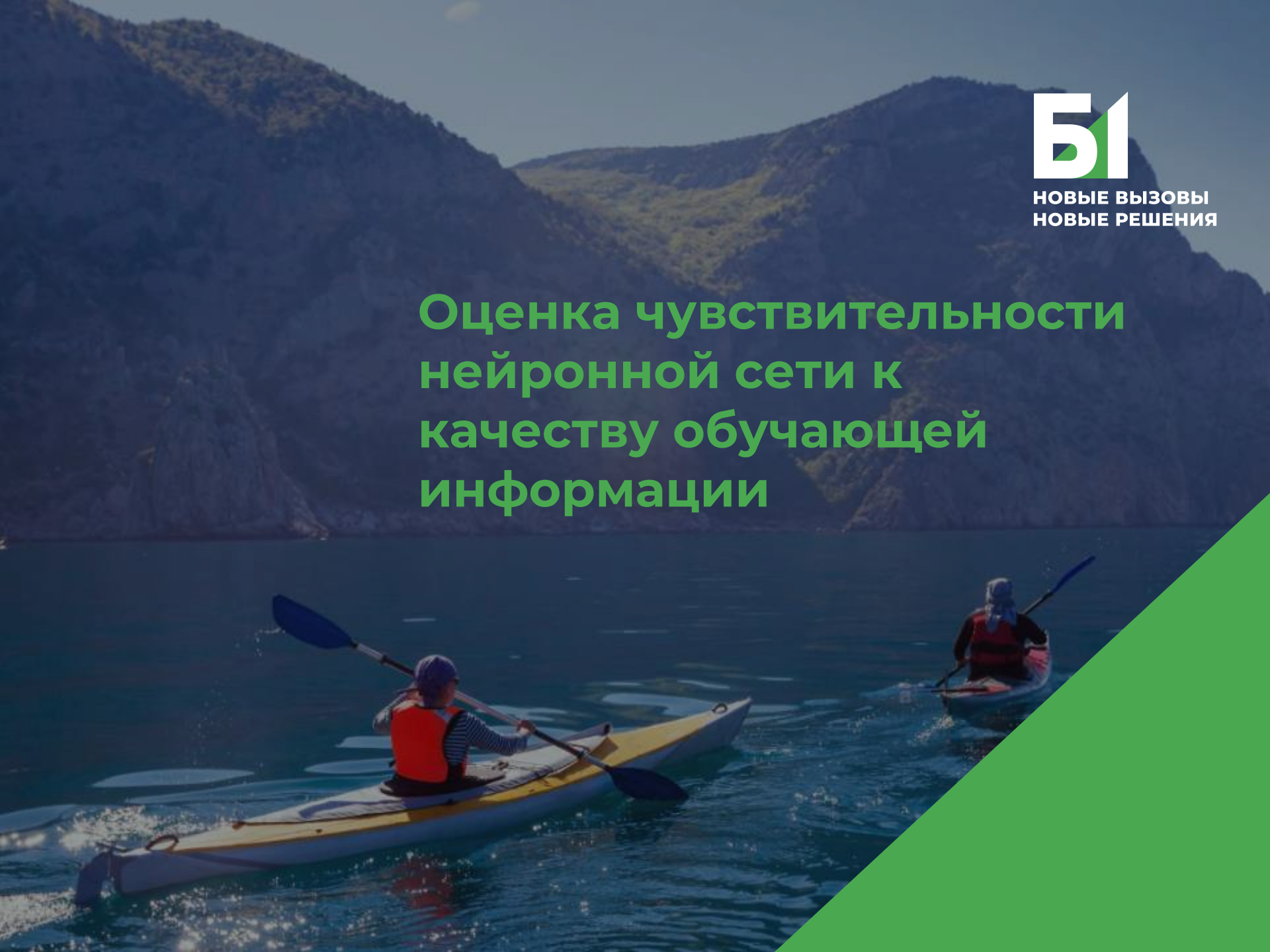
Архитектура НС, используемая для выявления мошеннических сделок в контрольных процедурах

Архитектура нейронной сети (автокодировщик)



Особенности обучения и работы нейронной сети

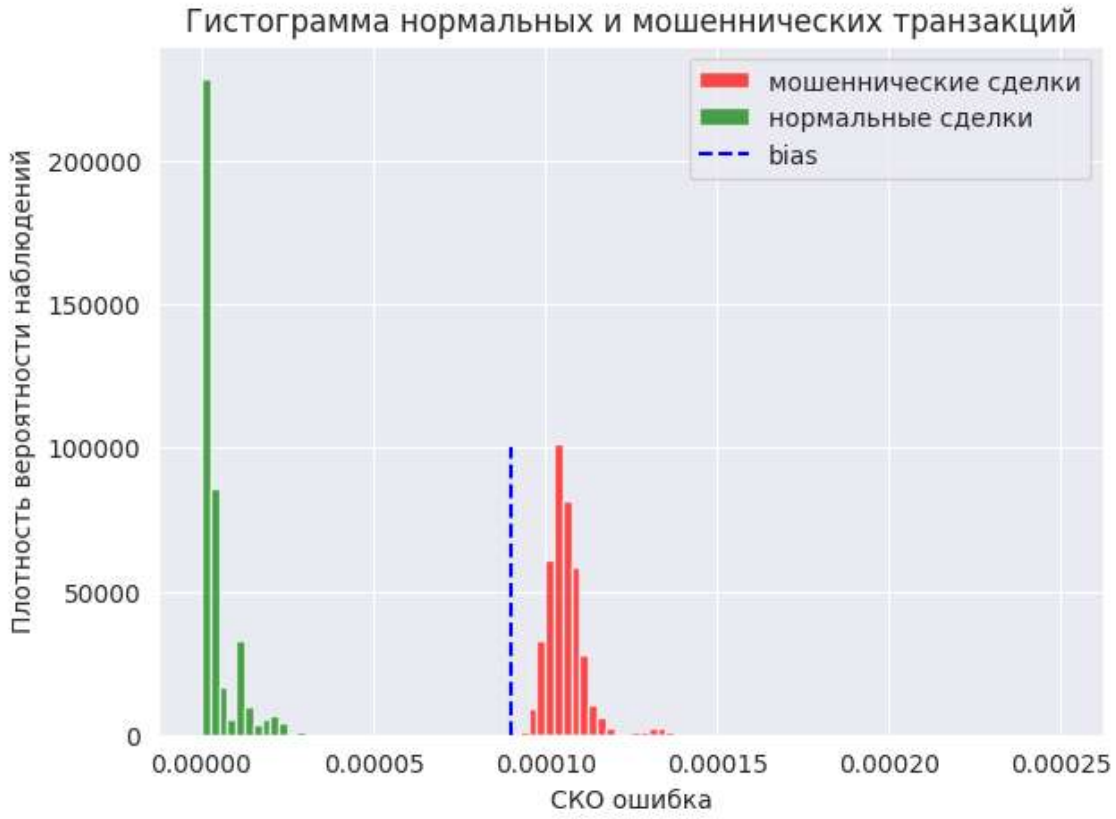
- **Для обучения** автокодировщика было использовано **3992** транзакций, из которых - 352 мошеннических
- В штатной ситуации **обучение** НС осуществлялось на информации, **не содержащей мошеннических сделок**
- При обучении нейронной сети:
  - исходная информация сжималась, что позволяло НС **выявить характерные черты** нормальных сделок
  - подбиралась такая **линия bias**, которая оптимальным образом **разделяла** исходные данные **на две группы**, на имеющие и не имеющие черты нормальных транзакций



НОВЫЕ ВЫЗОВЫ
НОВЫЕ РЕШЕНИЯ

Оценка чувствительности нейронной сети к качеству обучающей информации

Полученные результаты (штатная ситуация)



Средняя точность распознавания: 100%

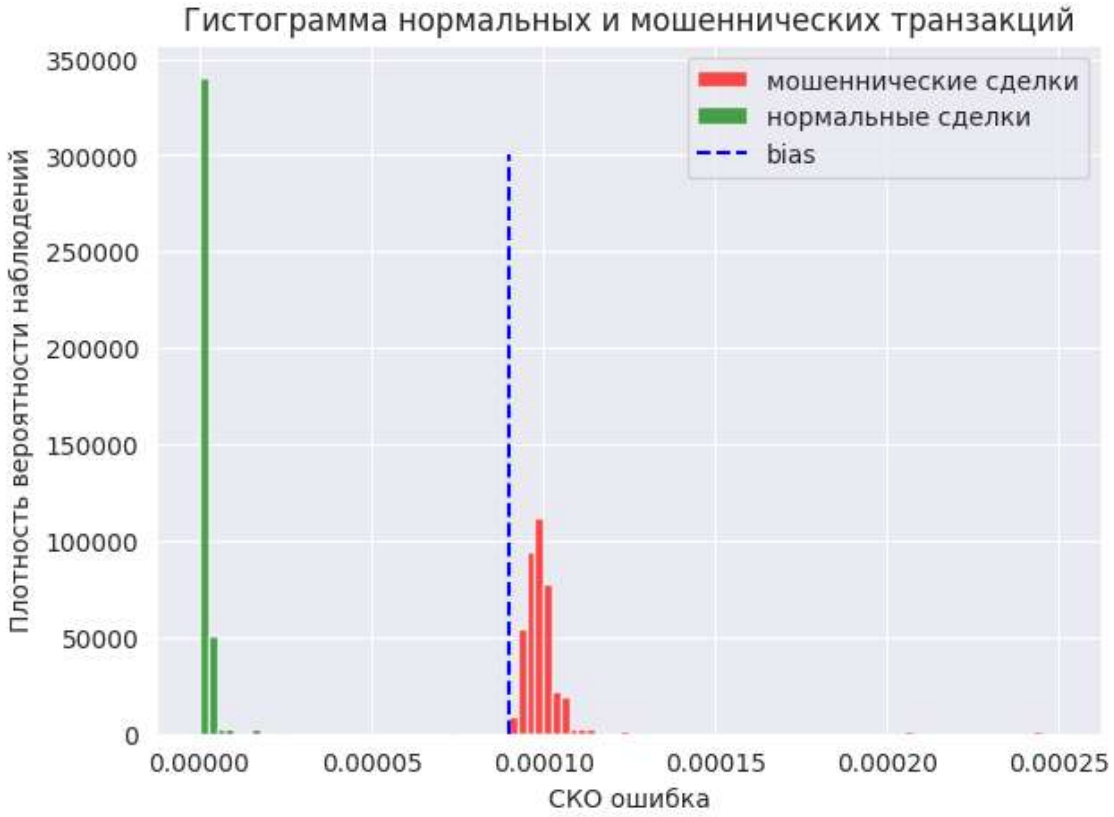
Для мошеннических транзакций: 100%

Для нормальных транзакций: 100%

Длительность обучения 100 эпох

bias = 0,00009

Результаты работы НС
(«подмешано» 1% мошеннических данных)



Средняя точность распознавания: 100 %

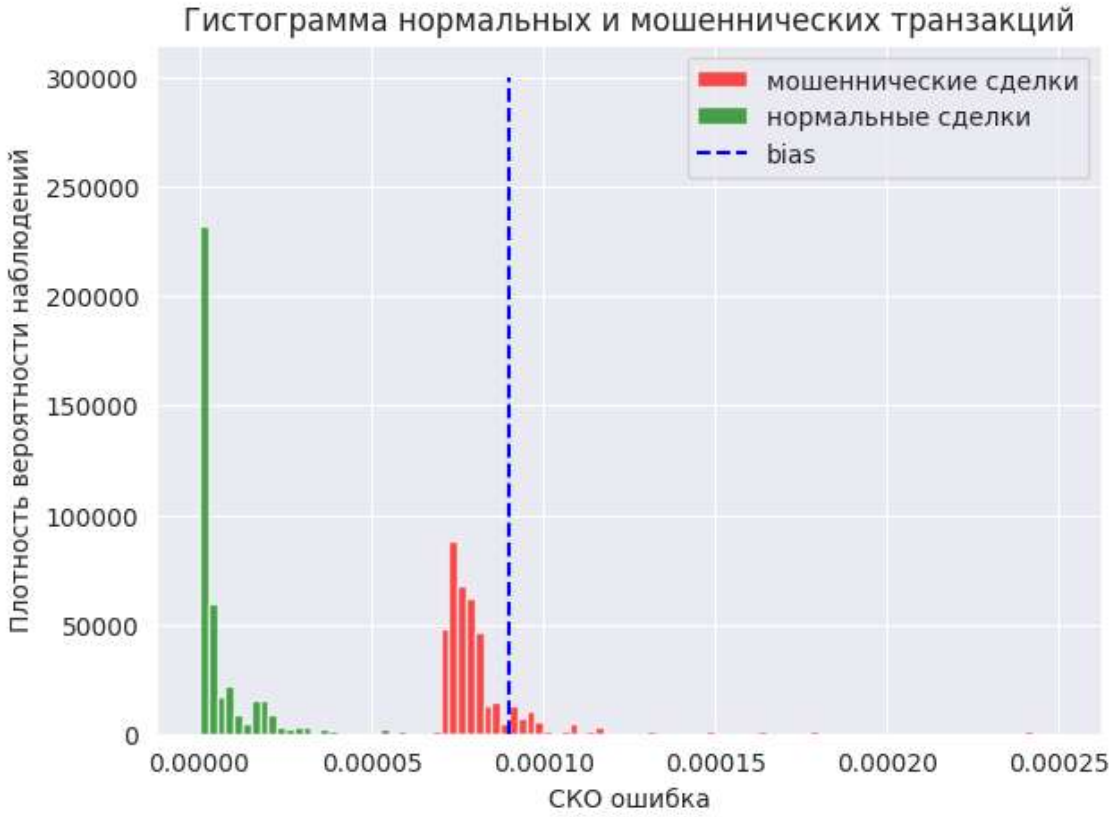
Для мошеннических транзакций: 100 %

Для нормальных транзакций: 100%

Длительность обучения 100 эпох

bias = 0,00009

Результаты работы НС
(«подмешано» 2% мошеннических данных)



Средняя точность распознавания: 58%

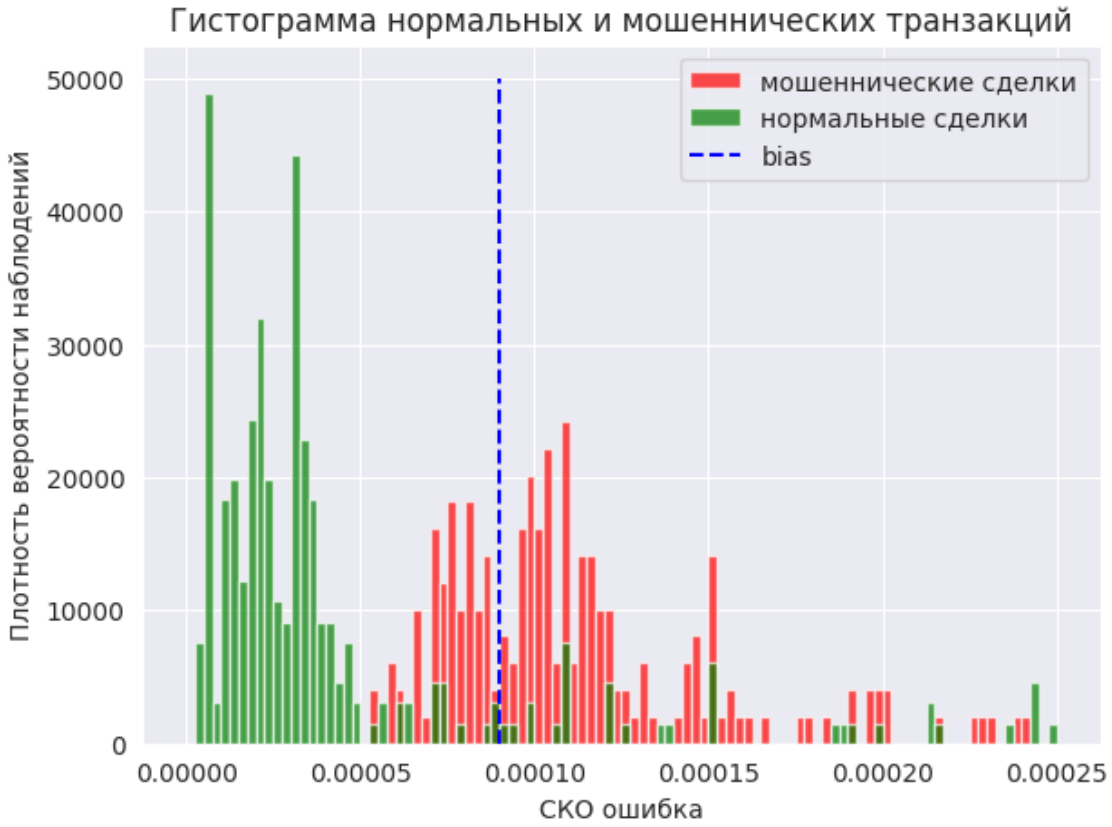
Для мошеннических транзакций: 15%

Для нормальных транзакций: 100%

Длительность обучения 100 эпох

bias = 0,00009

Результаты работы НС
(«подмешано» 3% мошеннических данных)



Средняя точность распознавания: 68%

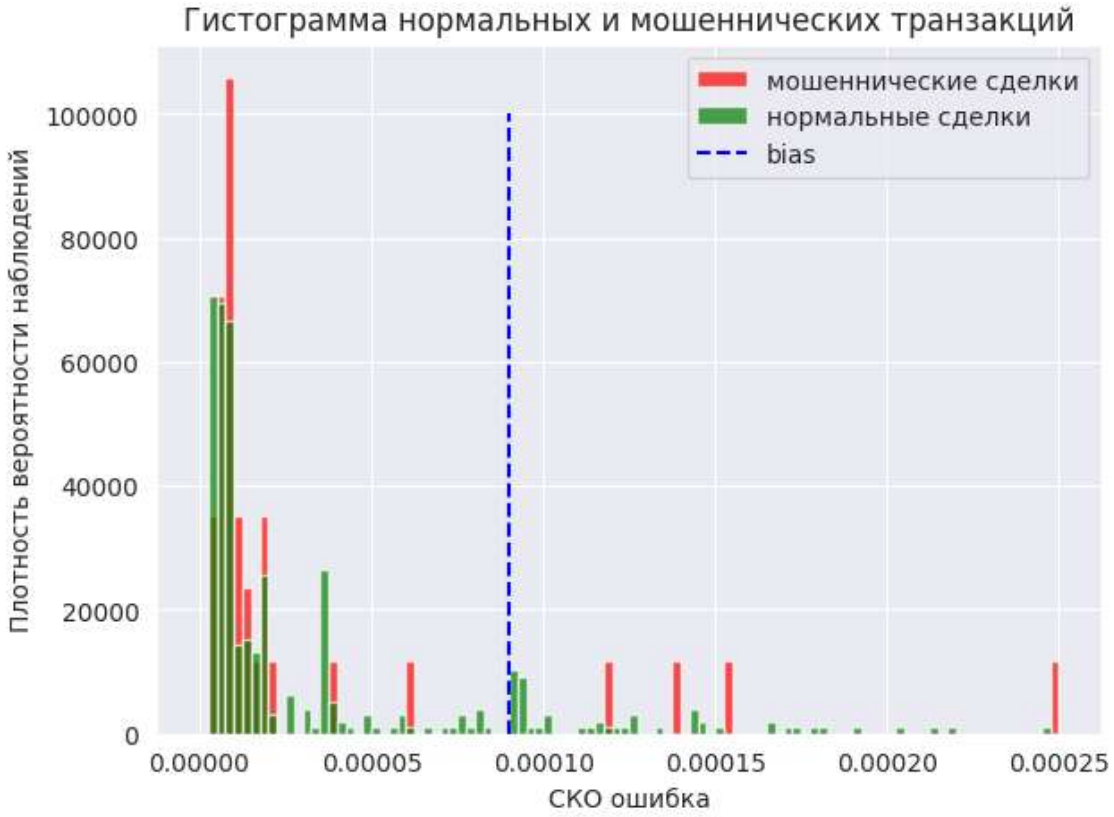
Для мошеннических транзакций: 74%

Для нормальных транзакций: 61%

Длительность обучения 100 эпох

bias = 0,00009

Результаты работы НС
(«подмешано» 9% мошеннических данных)



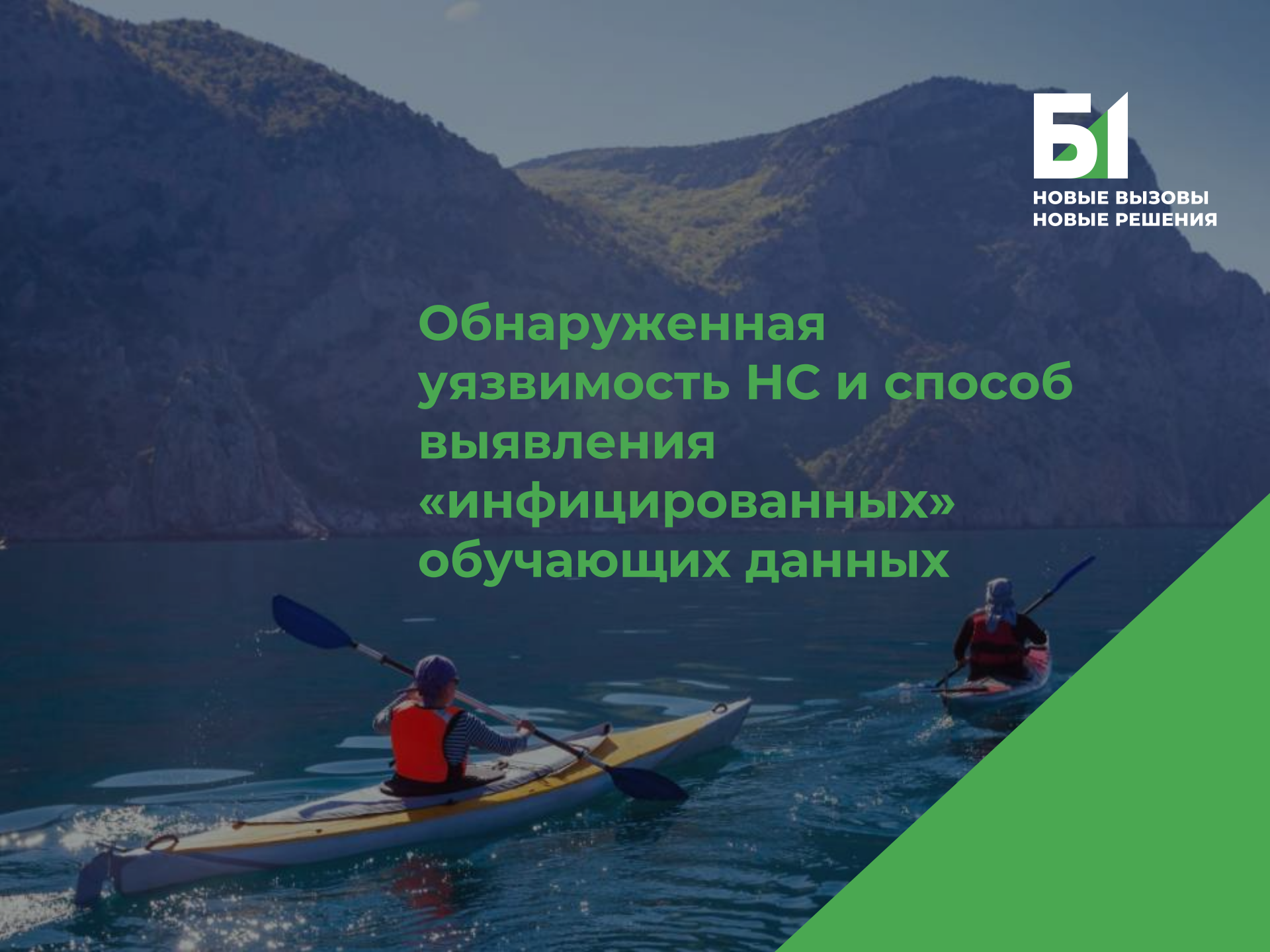
Средняя точность распознавания: 51%

Для мошеннических транзакций: 17%

Для нормальных транзакций: 85%

Длительность обучения 100 эпох

bias = 0,00009



НОВЫЕ ВЫЗОВЫ
НОВЫЕ РЕШЕНИЯ

Обнаруженная уязвимость НС и способ выявления «инфицированных» обучающих данных

Требования к дипфейку табличных данных

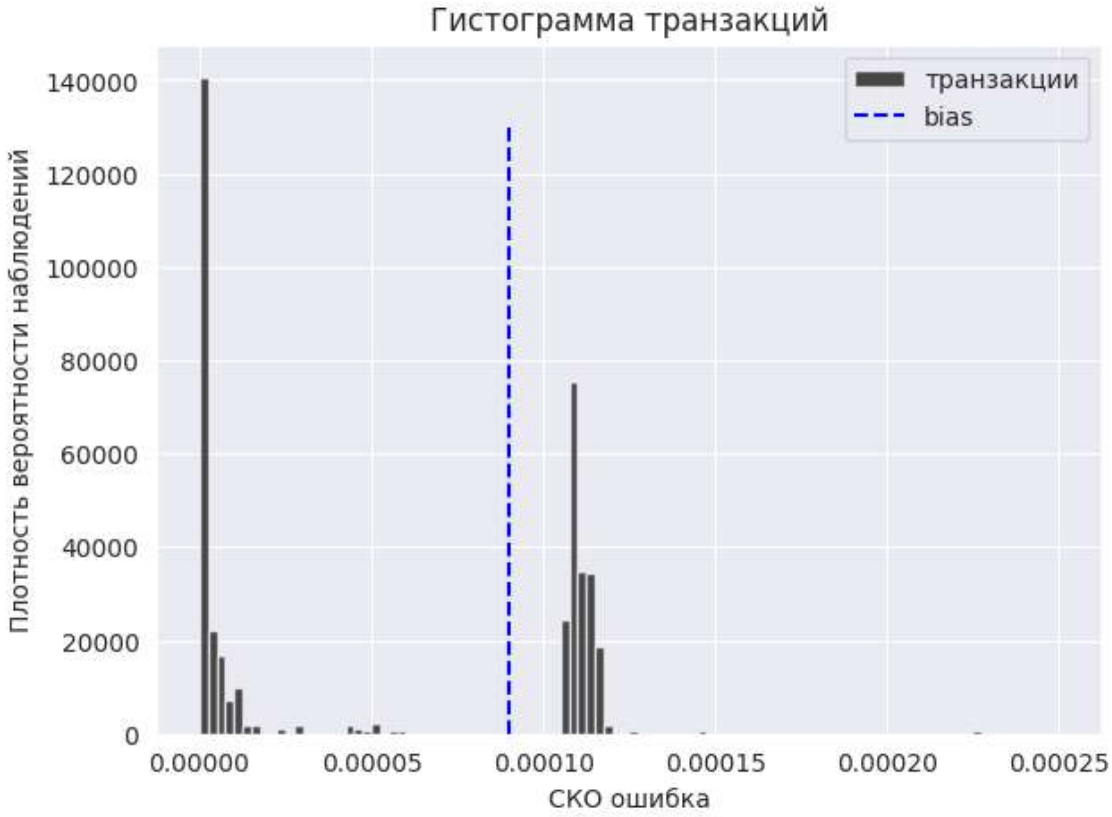
- Способ должен быть ЗАМЕТНЫМ для нейронной сети, чтобы она смогла обучиться на небольшом количестве (<1%) специальным образом подготовленных «инфицированных» данных, считая их нормальными, но при этом разделяла сделки на две компактные группы, содержащие остальные мошеннические сделки и нормальные транзакции
- Способ должен быть НЕЗАМЕТНЫМ для специалиста, анализирующего обучающую информацию и результаты работы нейронной сети, т.е. **быть дипфейком**
- Должна существовать УЯЗВИМОСТЬ нейронной сети, которая удовлетворяла бы этим двум требованиям

Уязвимость текстовых полей табличных данных

Существуют шрифты, **написание некоторых букв** в которых на латинской и русской раскладках (в кодировке UTF-8) **не отличаются друг от друга** (омоглифы):

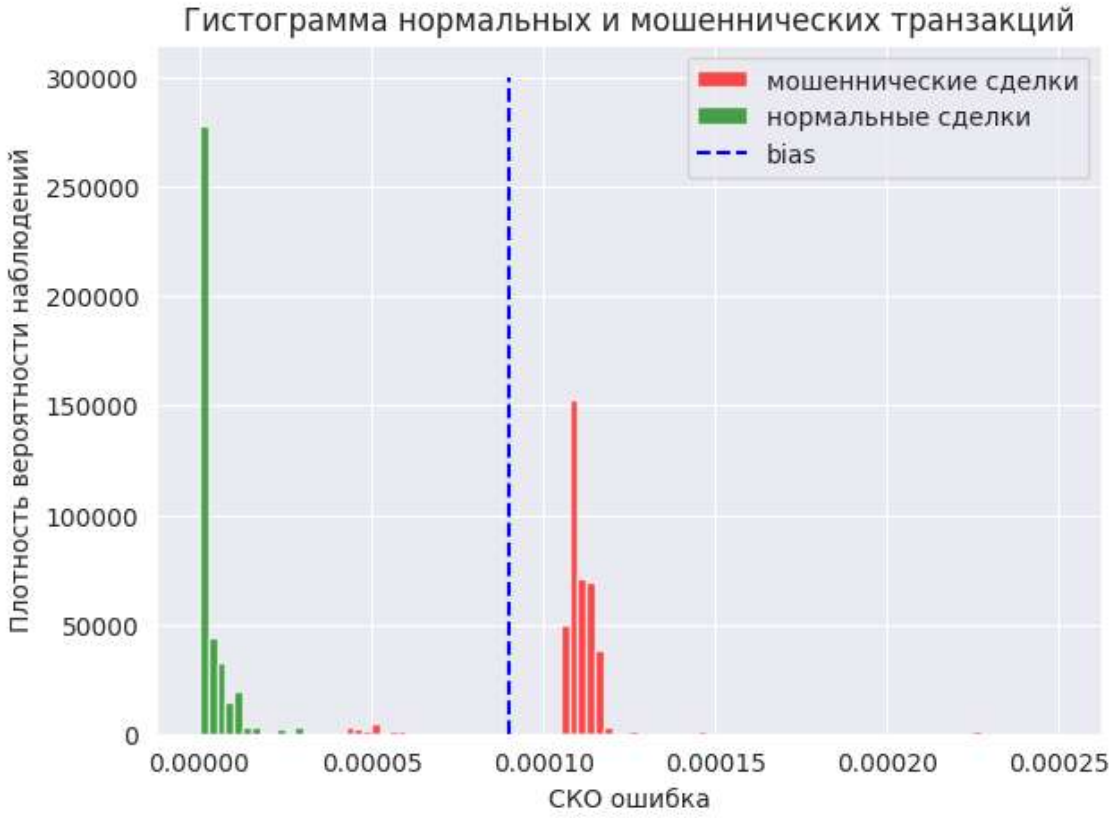
A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y	Calibri (Eng)
A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y	Calibri (Рус)
A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y	Arial (Eng)
A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y	Arial (Рус)
A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y	Times New Roman (Eng)
A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y	Times New Roman (Рус)
A,a,B,C,c,E,e,H,K,M,O,o,P,p,X,x,y	Courier (Eng)
A,a,B,C,c,E,e,H,K,M,O,o,P,p,X,x,y	Courier (Рус)

Результаты работы НС (испорчено 0,33% сделок)



Для обучения НС было использовано всего 12 записей-дипфейков

Результаты работы НС (испорчено 0,33% сделок)



Средняя точность распознавания: 98%

Для мошеннических транзакций: 97%

Для нормальных транзакций: 100%

Длительность обучения 100 эпох

bias = 0,00009

Для обучения НС было использовано всего 12 записей-дипфейков

Как обнаружить дипфейк

- Можно обнаружить подмену шрифта при обучении нейронной сети, если использовать шрифт, который не имеет русского аналога, например:

A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y - Bradley Hand ITC (Eng)
А, а, В, С, с, Е, е, Н, К, М, О, о, Р, р, Х, х, у - Bradley Hand ITC (Рус)

A, a, B, C, c, E, e, H, K, M, O, o, P, p, X, x, y - Chiller (Eng)
А, а, В, С, с, Е, е, Н, К, М, О, о, Р, р, Х, х, у - Chiller (Рус)

- Если нейронная сеть уже обучена, то выявить подобный дипфейк невозможно:
  - небольшое количество испорченной информации маскируется на фоне компактной группы других мошеннических сделок
  - анализ весов, используемых для связей нейронов в нейронной сети, практически невозможен из-за их огромного количества, т.к. в процессе обучения обучающая информация преобразуется в числовые значения нескольких миллионов весов НС

НОВЫЕ ВЫЗОВЫ
НОВЫЕ РЕШЕНИЯ

Выводы и рекомендации

Выводы

1. **Существует возможность создания дипфейков** для табличных данных в текстовом формате
2. Они характеризуются тем, что **их трудно выявить при подготовке** обучающей информации и **практически невозможно обнаружить после** того, как нейронная сеть уже обучена на испорченной информации
3. Подобные дипфейки могут быть **использованы в случае большого количества транзакций с относительно небольшими суммами** и с верификацией только с помощью искусственного интеллекта, без участия эксперта
4. При внедрении искусственного интеллекта в контрольные процедуры **возрастает риск недобросовестности специалистов комплаенс-подразделений и их руководителей**, которые могут быть заинтересованы во внедрении подобного дипфейка

Рекомендации

Специалистам по безопасности бизнеса необходимо:

1. **При подготовке к внедрению нейронных сетей** в практику работы комплаенс-подразделений **действовать на упреждение** и выявлять уязвимости в обучении и работе нейронных сетей, чтобы к моменту широкого внедрения искусственного интеллекта в контрольные процедуры у них был инструментарий для выявления подобных уязвимостей
2. **При использовании НС** в комплаенс-процедурах **выявлять недобросовестных сотрудников** и **осуществлять выборочную ручную и автоматическую проверки работы** подразделений с использованием нейронных сетей, заново обученных на специально подготовленной и тщательно проверенной информации

О ГРУППЕ КОМПАНИЙ Б1

Группа компаний Б1 (ранее компания EY в России и Беларуси) предлагает полный спектр профессиональных услуг, включая услуги в области аудита, налогообложения, права, стратегии, сделок и консалтинга.

За более чем 30-летний период работы в России и 20-летний период в Беларуси в компаниях группы создана сильнейшая команда специалистов, обладающих обширной экспертизой и опытом реализации сложнейших проектов, в 10 городах: Москве, Минске, Владивостоке, Екатеринбурге, Казани, Краснодаре, Новосибирске, Ростове-на-Дону, Самаре, Санкт-Петербурге и Тольятти.

Группа компаний Б1 помогает клиентам находить новые решения, расширять, трансформировать и успешно вести свою деятельность, а также повышать свою финансовую устойчивость и кадровый потенциал.

B1.RU | B1.BY

НОВЫЕ ВЫЗОВЫ
НОВЫЕ РЕШЕНИЯ