# Состязательные атаки на системы с искусственным интеллектом и методы борьбы с ними

Игорь Саенко[1] и **Владимир Садовников**[1]

[1] Лаборатория вычислительной техники Безопасность Проблемы, Санкт-Петербургский федеральный исследовательский центр Российской академии наук (НПЦ РАН)

XXIX научно-практическая конференция «КЗИ» 15 - 17 мая 2024 года г. Санкт-Петербург

# Содержание (1)

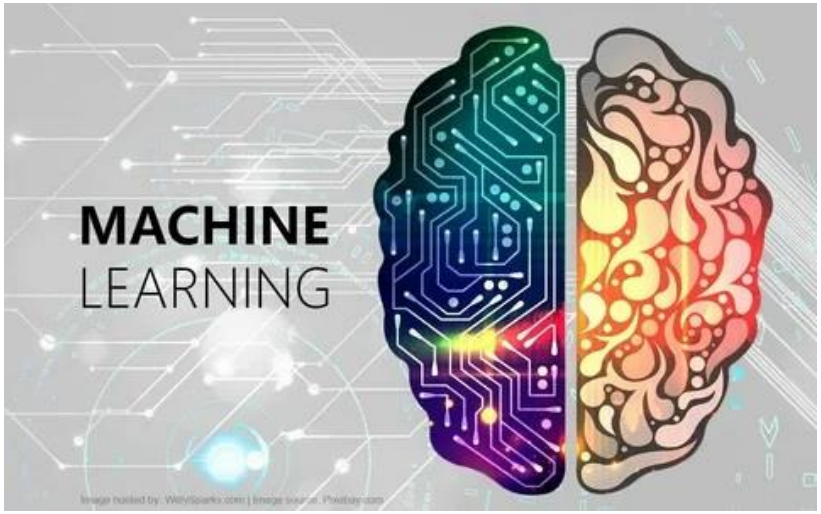
- **Введение**
- Релевантные работы
- Основные определения
- Реализация
- Экспериментальная оценка
- Заключение

# Технология машинного обучения

**Технология машинного обучения (ML) в настоящее время активно внедряется в различных областях, включая**

- **обнаружение аномалий и компьютерных атак,**
- **компьютерное зрение,**
- **распознавание образов,**
- **автономные транспортные средства и другие сложные проблемы.**

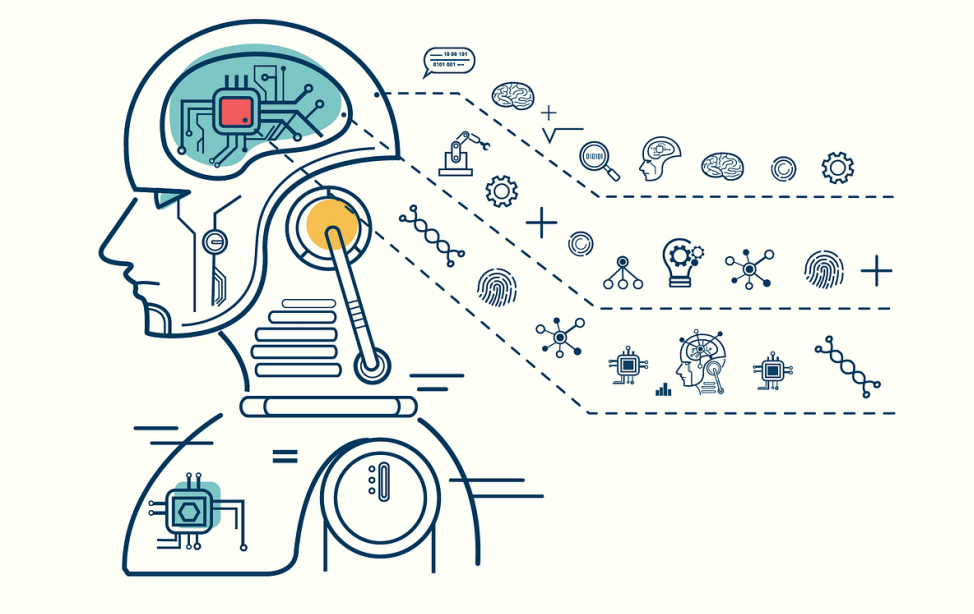
**Однако, несмотря на все преимущества, подходы ML остаются уязвимый к преднамеренным нападениям на обучающие и тестовые наборы данных, а также на настройка параметров из моделей ML.**



# Цель

**1. На основе анализа известных атак на системы ML и методов защиты от них в статье предлагается новый подход к ликвидации последствий нападений в наборах данных изображений, основанный на Neural-Cleanse и JPEG-сжатии методы. Это *основной вклад* из газеты.**

**2. В реализации этого подхода и его экспериментальная оценка рассматриваются, которые показывают значительную увеличение в точность распознавания изображений после кибератаки.**



# Содержание (2)

- Введение
- **Релевантные работы**
- Основные определения
- Реализация
- Экспериментальная оценка
- Заключение

# Релевантные работы

**Основные работы**:

**Воробейчик, Y. и др., “Состязательное машинное обучение. Обобщающие лекции по искусственному интеллекту и машинному обучению,” (2018):**

В данной статье рассматриваются атаки, основанные на примерах состязательности, которые могут обходить защитные механизмы различных моделей машинного обучения, предназначенных для обеспечения безопасности искусственного интеллекта.

**Маркус Comiter. “Атака на искусственный интеллект. Уязвимость в системе безопасности ИИ и что политики могут с этим сделать,” (2019):**

Эта работа содержит обзор основных угроз и вызовов, связанных с безопасностью искусственного интеллекта, а также предлагает рекомендации и решения для политиков и законодателей. Он рассматривает различные варианты использования искусственного интеллекта, которые могут быть атакованы, и описывает потенциальные уязвимости, с которыми они сталкиваются.

# Релевантные работы

**[Чжоу и др., 2023], [Ли и др., 2022], [Ли и др., 2021]:**

- рассматриваются различные типы состязательных атак и их последствия для систем машинного обучения,
- содержит обзор существующих методов и стратегий защиты от таких атак,
- представлены собственные разработки и подходы к защите от состязательных атак, основанные на результатах экспериментов.

**[Чжао и др., 2021] -** охватывает как теоретические аспекты, так и практические примеры и рекомендации по защите от атак на системы ML.

**[Хамайсе и др., 2022], [Ахтар и др., 2021]** обзорные статьи, в которых рассматриваются различные технологии и методы атаки и защиты в области распознавания изображений и компьютерного зрения. Описаны различные механизмы и технологии, используемые злоумышленниками для проведения состязательных атак.

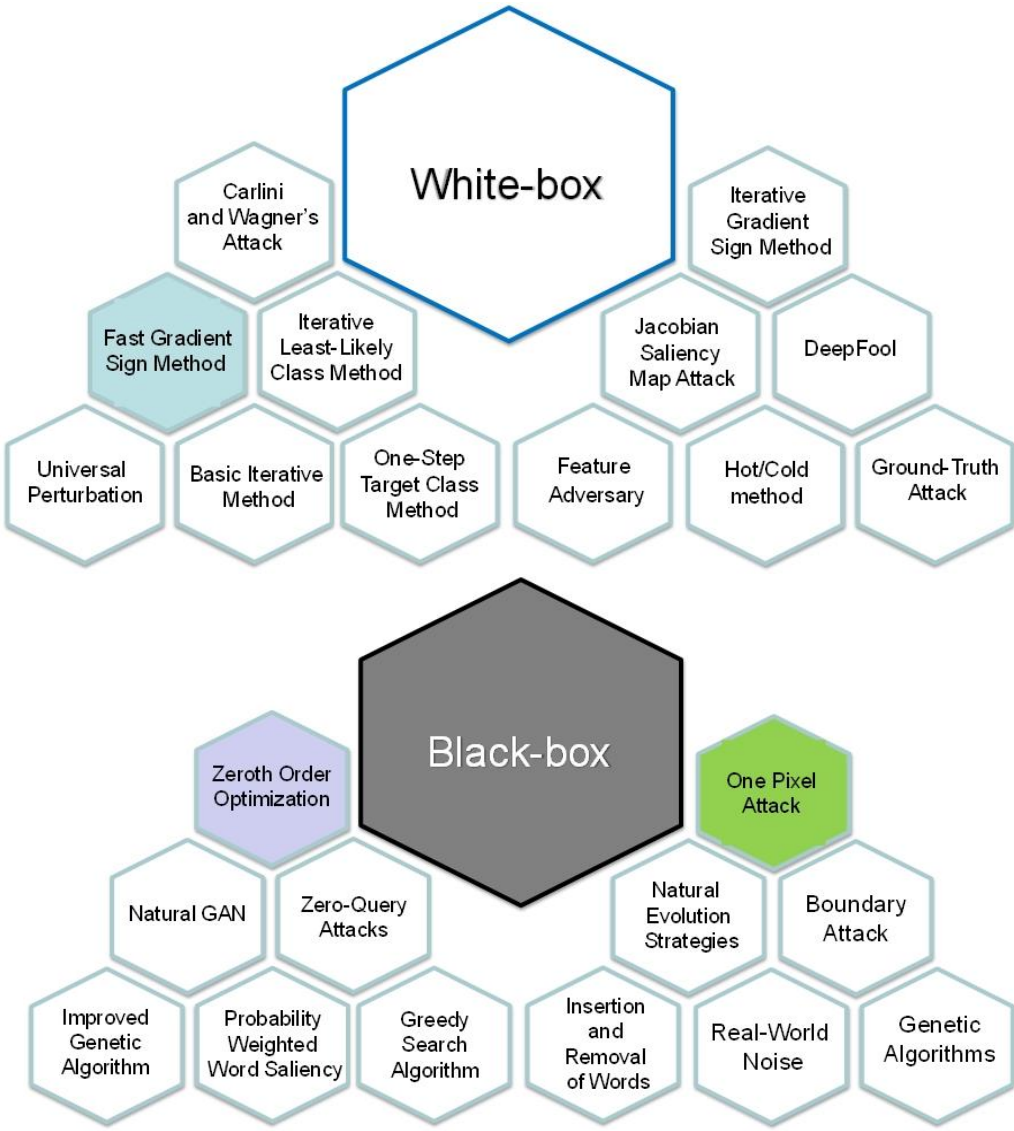
# Релевантные работы **(3/3)**

**Котенко et al., ”Атаки на системы машинного обучения: анализ и подход к защите на основе GAN” (IITI’2023):**

- классификация атак на МО и средств защиты от них,
- метод защиты, основанный на GAN, Neural-Cleance Сжатие в формате Jpeg
- Рассматривается возможность атаки FGSM

В **отличие нашей работы** От **Предыдущая страница** один:

- отсутствие GAN
- рассматриваются 3 атаки: **FGSM, ZOO и OPA**



# Выводы из анализа атак на МО

1. Анализ соответствующих работ приводит к выводу, что подавляющее большинство атаки на системы МО являются состязательными атаками.
2. Состязательные атаки взаимодействуют с искусственными нейронными сетями, искажая входные данные таким образом, чтобы модель выдала неверный результат.
3. Состязательные атаки делятся на две категории ("Белый ящик“ и "Черный ящик") и распределены по трем доменам: Изображение, Аудио и Текст.
4. Самый популярный методы конкурентной атаки являются:

- Быстрый градиентный знаковый метод (FGSM),
- Оптимизация нулевого порядка (ZOO),
- Атака одним пикселем (ОПА).

# Содержание (3)

- Введение
- Релевантные работы
- **Основные определения**
- Реализация
- Экспериментальная оценка
- Заключение

# Основные атаки

**FGSM – это атаки на нейронные сети, которая осуществляет обман обученных моделей. Цель метода FGSM заключается в том, чтобы изменить изображение незначительно таким образом, чтобы обученная модель ошибочно идентифицировала его как другой класс. Для этого используется градиентный спуск, позволяющий найти наиболее чувствительные пиксели на изображении.**

**ZOO - это метод атаки на модели машинного обучения, который обходит ограничение на доступ к градиентам функций потерь и использует только оценки этих функций. Атакующий в этом методе не имеет прямого доступа к внутренним параметрам модели и информации о ее структуре. Вместо этого злоумышленник работает с внешним API модели, отправляя входные данные и получая ответы, но не имея доступа к внутренним параметрам.**

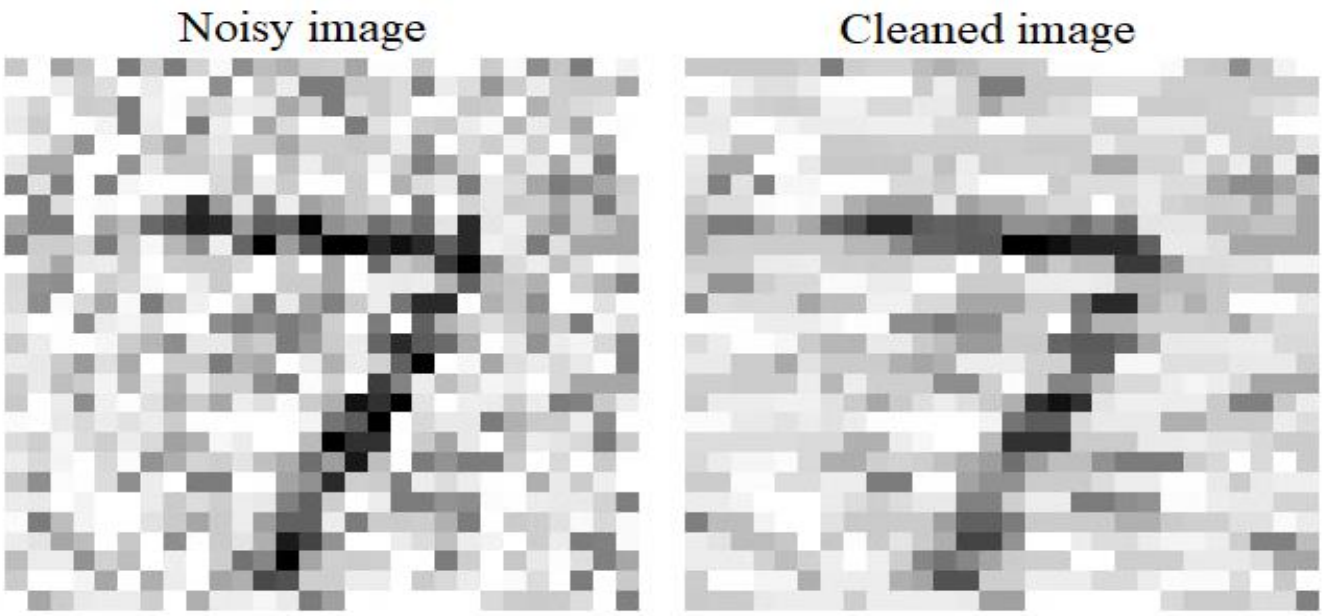
**One Pixel Attack - это вид атаки в области компьютерного зрения и обработки изображений, который изменяет всего один пиксель в изображении, чтобы ввести в заблуждение систему искусственного интеллекта (ИИ) и изменить ее вывод. Этот тип атаки, может быть, использовал для обмана систем ИИ, работающих с изображениями, например, для распознавания объектов или классификации изображений.**

# Сжатие в формате JPEG

**Сжатие в формате JPEG** это метод сжатия изображения, который используется для уменьшения размера изображений без существенной потери качества.

# Neural-Cleanse

**Neural-Cleanse** применяет алгоритм оптимизации для обнаружения и удаления вредоносных примеров. Он использует два основных этапа: Обучение и перечисление. На этапе обучения модель обучается на исходных данных. На этапе перебора она используется для оценики входных данных чтобы определять является ли это вредоносный пример или нет.




Применение **Neural-Cleanse** к зашумленному изображению

# Основная идея

В **идея предлагаемого подхода**:

- на первом этапе в алгоритме возможности состязательных атак максимально искажены. Для этого предлагается использовать Сжатие в формате JPEG процедура;

- на втором этапе, последствиями сжатия Jpeg являются устранен (очищен) используя обработка нейронной сетью реализовано с помощью Технология очищения нервной системы.

# Содержание (4)

- Введение
- Релевантные работы
- Основные определения
- **Реализация**
- Экспериментальная оценка
- Заключение

# Набор данных #1

**MNIST-JPG**, https://github.com/teavanist/MNIST-JPG.

MNIST-JPG состоит из 1000 изображений из чисел с разными вариантами написания. Каждое число представлено 100 различными вариантами написания.

В **тренировочный набор** включенный:

- данные без уязвимостей (65%)
- данные с уязвимостями (35%) в результате атаки.

**Классификаторы:**

- Bagging Classifier,
- Random Forest Classifier,
- Extra Trees Classifier
- Gradient Boosting Classifier

# Набор данных №2

**Набор данных изображений деталей ПК [Классификация],** https://www.kaggle.com/datasets/asaniczka/pc-parts-images-dataset-classification?resource=download. Набор данных состоит из 3279 изображений из компоненты персонального компьютера. Набор данных содержит 14 классов. Каждое изображение имеет определенное разрешение размером 256x256 пиксели в JPG формат.

В **тренировочный набор** включенный:

- надежный данные (70%)
- данные с уязвимостями (30%) в результате нападения.

**Классификаторы:**

- K-Ближайшие соседи Классификатор,
- Случайный Лес Классификатор,
- Наивный Байес Классификатор
- Деревья Принятия решений Классификатор

# Реализация этого подхода

Предложенный подход был реализован в среде PyCharm.

**Библиотеки**:

*import matplotlib.pyplot as plt*

*import numpy as np*

*import os*

*import tensorflow as tf*

*import cleverhans as ch*

*from art.utils import load_dataset*

*from art.attacks.evasion import HopSkipJump*

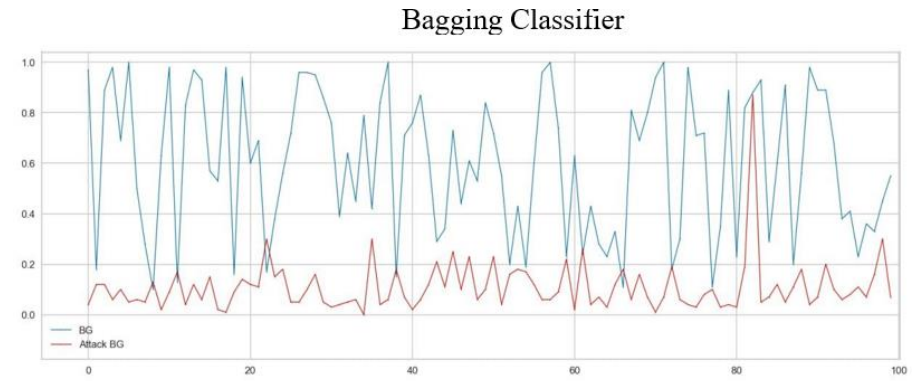
*from art.estimators.classification import SklearnClassifier*

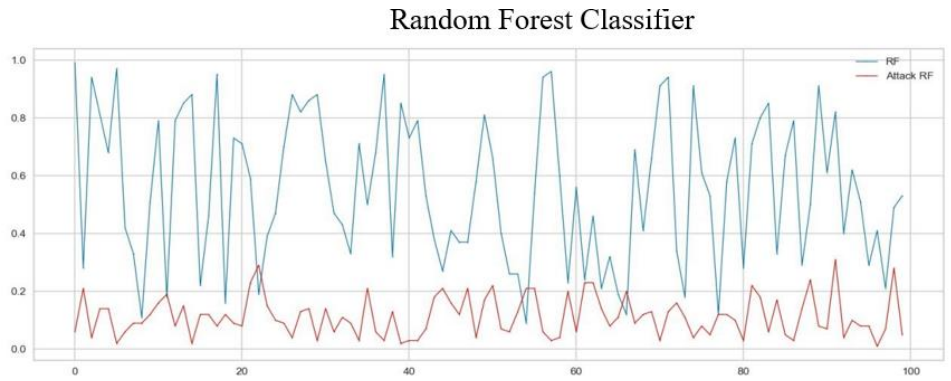
**Построение графиков** было проведено с использованием *Matplotlib*.

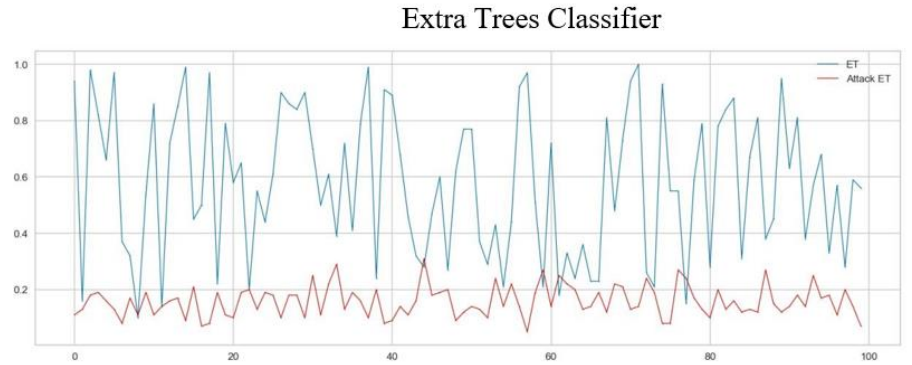
# Содержание (5)

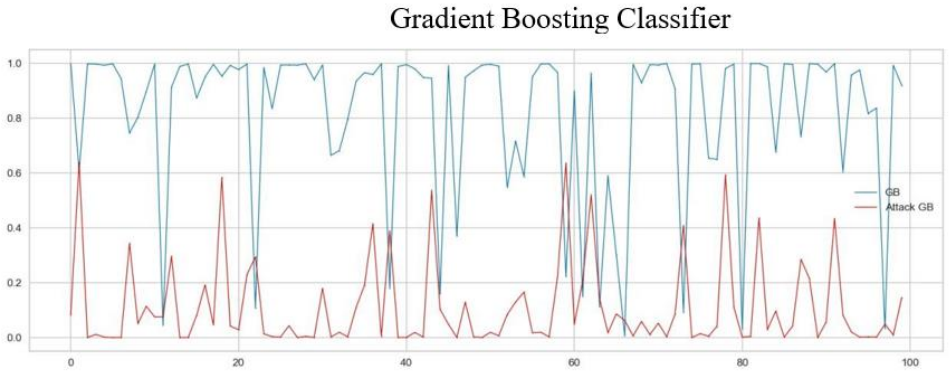
- Введение
- Релевантные работы
- Основные определения
- Реализация
- **Экспериментальная оценка**
- Заключение

# До того , как FGSM в наборе данных MNIST-JPG



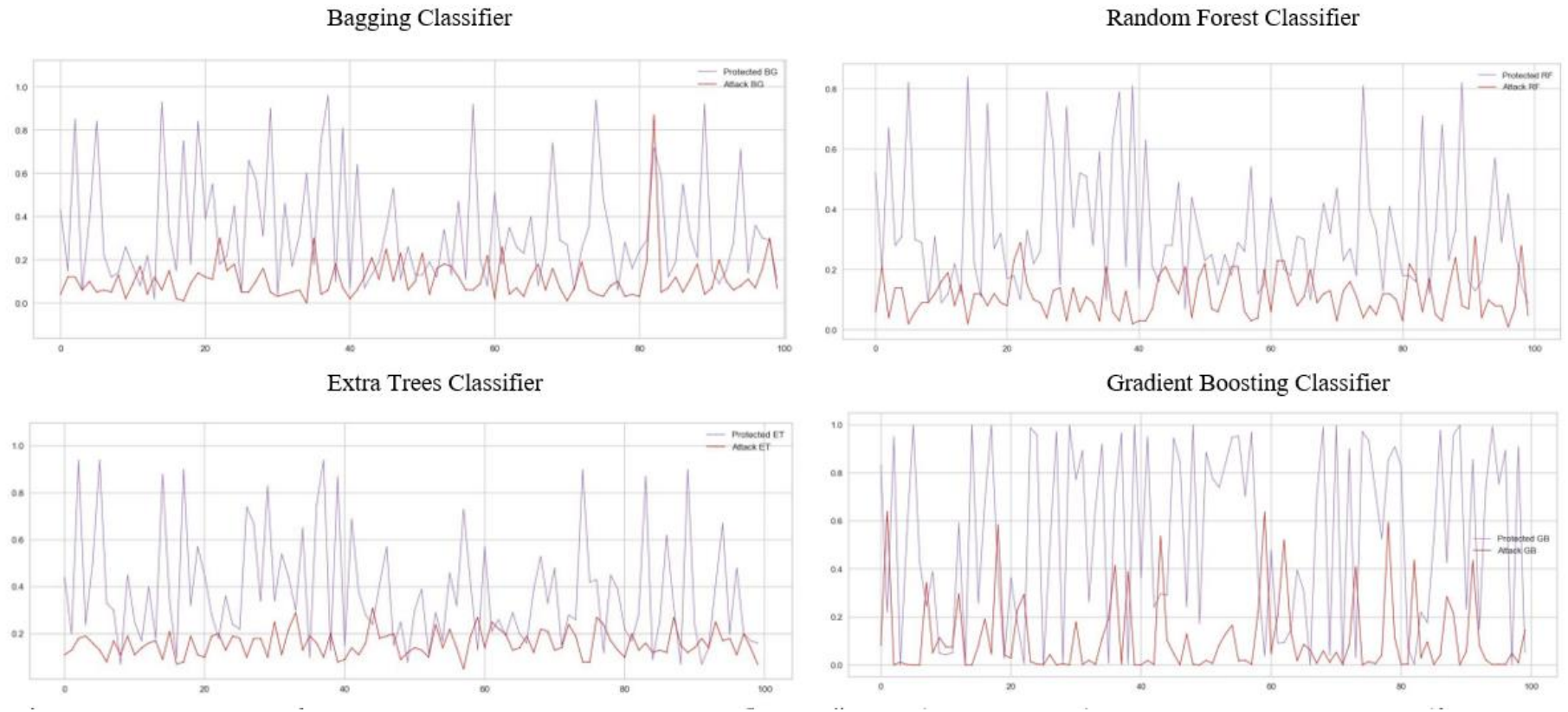




Ось X - номера итераций из классификаторов (*N*)

Ось Y - точность распознавания цифровых изображений (*P*) подается на вход классификатора.

Перед нападением - синие линии. После нападения - красные линии.

# После FGSM в наборе данных MNIST-JPG




Ось X - в номера итераций из классификаторов (*N*)

Ось Y - в точность распознавания цифровых изображений (*P*) подается на вход классификатора.

После нападения - красные линии. После нашего подхода - фиолетовые линии.

# Обсуждение

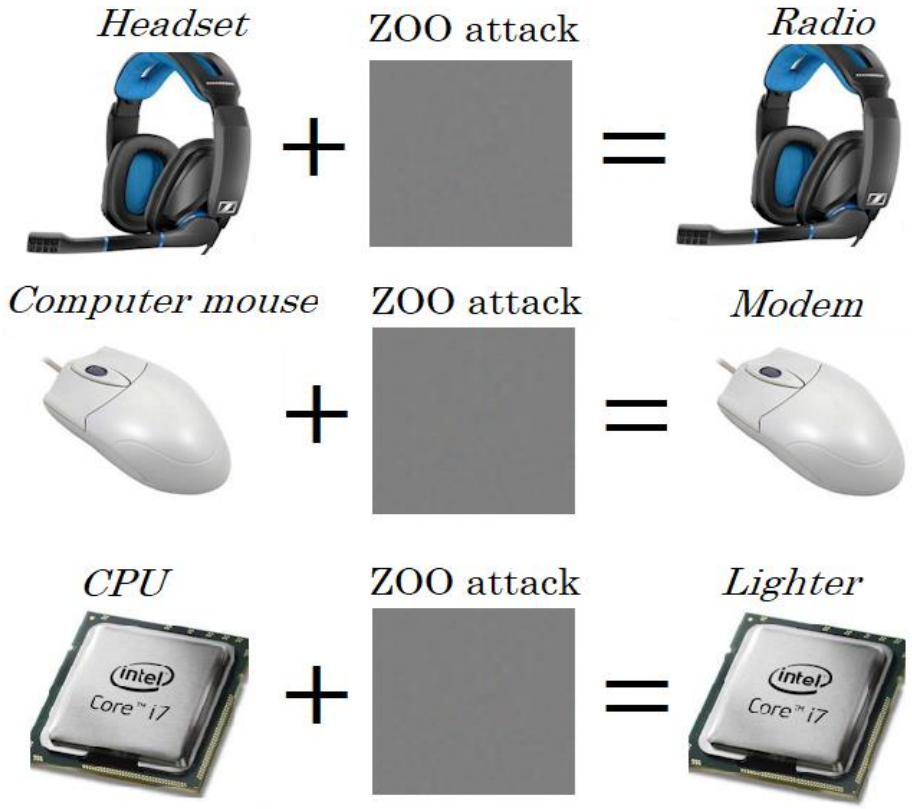
| Classifier | Before the attack | After the attack | After our approach |
|---|---|---|---|
| Bagging | 0.723 | 0.171 | 0.644 |
| Random Forest | 0.756 | 0.158 | 0.672 |
| Extra Trees | 0.794 | 0.208 | 0.689 |
| Gradient Boosting | 0.910 | 0.315 | 0.903 |

Восстановление производительности распознавания изображений было самым сильным для классификатора Gradient Boosting.

Для этого классификатора разница от исходного значения составляет всего 0.007 Другими словами, последствия атака FGSM для этого классификатора были **почти полностью устранена**.

Это объясняется тем фактом, что Gradient Boosting является **ансамблевым** классификатором, который показывает наивысшую эффективность на исходных изображениях.

# Атака ZOO на набор данных «Изображений деталей ПК»

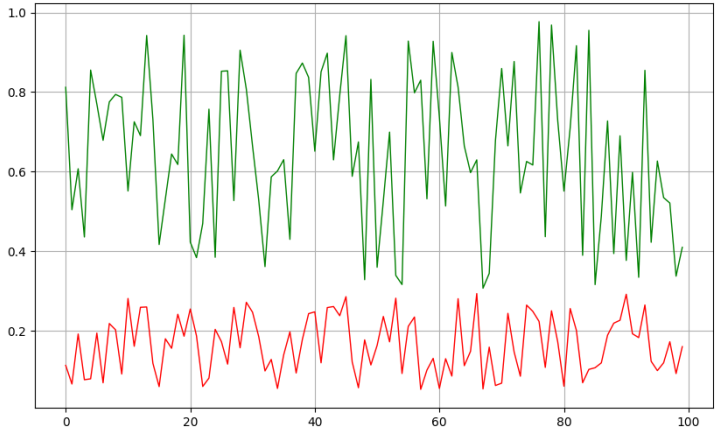



**Атака ZOO** использует *оптимизацию нулевого порядка* для создания состязательных примеров. Это работает с помощью добавления небольших возмущений к исходному входному изображению и наблюдением за тем, как меняются выходные данные модели.

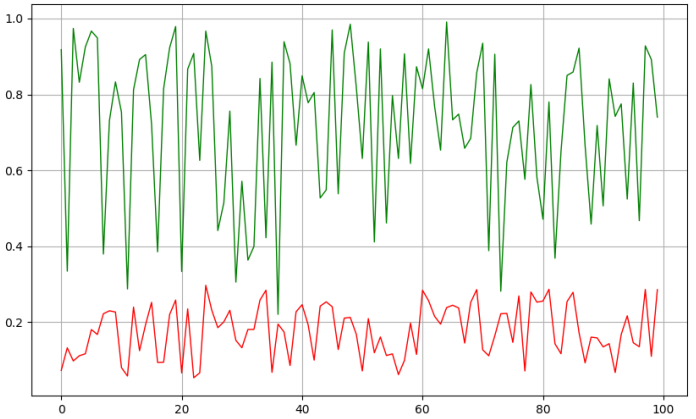
**Примечание:** Далее мы используем добавление Гауссовский и пуассоновский шумы в нашем подходе вместо Сжатие в формате Jpeg.

# До того , как прошла атака ZOO в наборе данных «изображений деталей ПК»

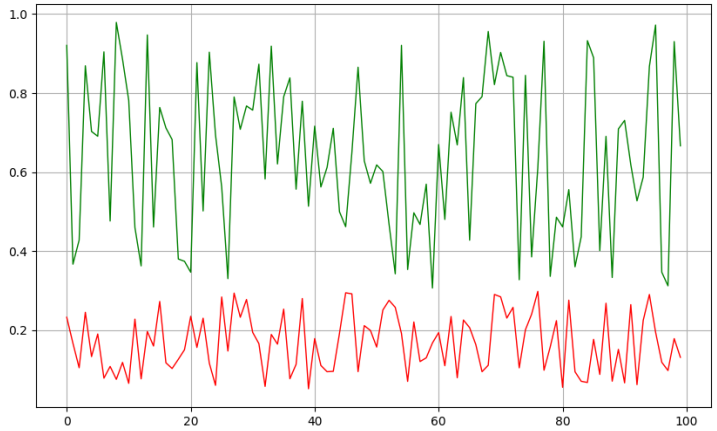
K-Ближайшие соседи



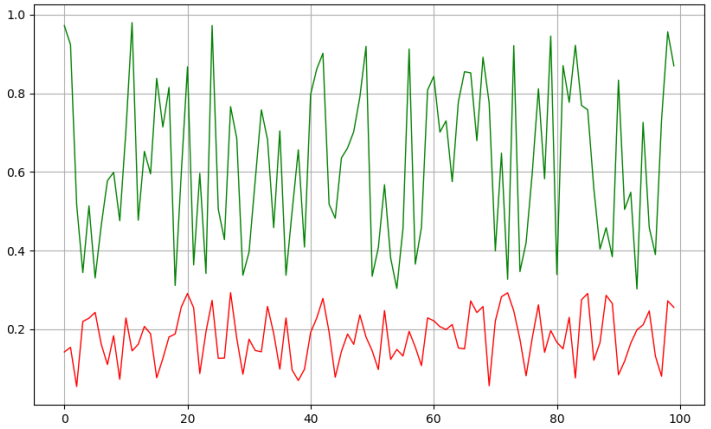

Случайный Лес




Наивный Байес



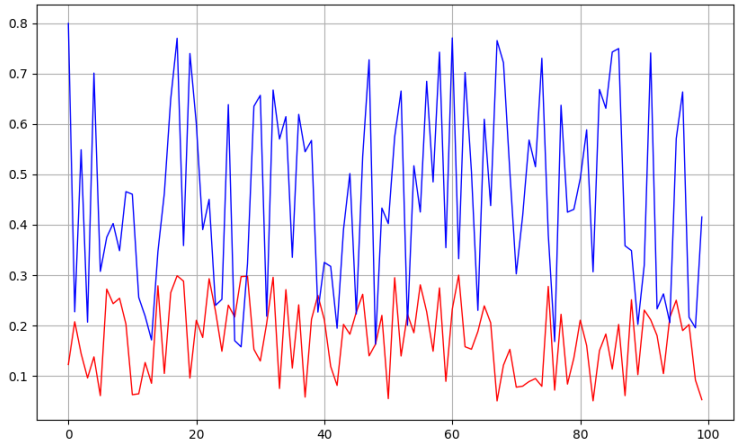

Деревья Принятия решений



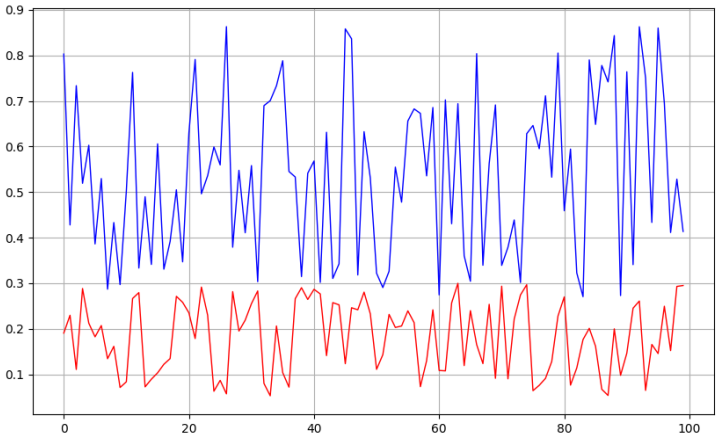

Перед атакой - зеленые линии. После атаки - красные линии.

# После того как прошла атака ZOO в наборе данных «изображений деталей ПК»

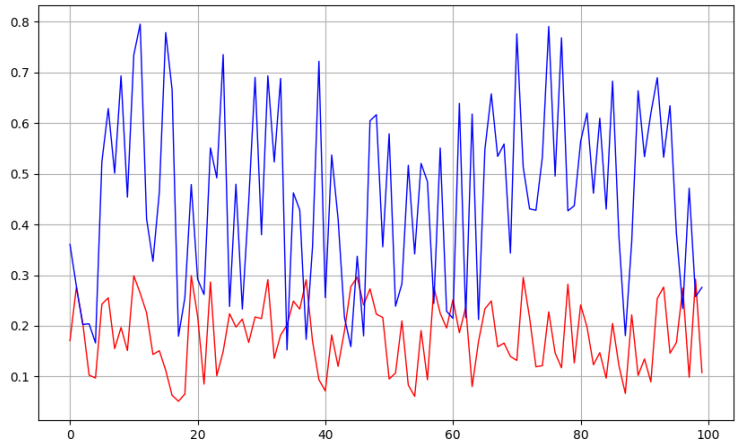
K-Ближайшие соседи



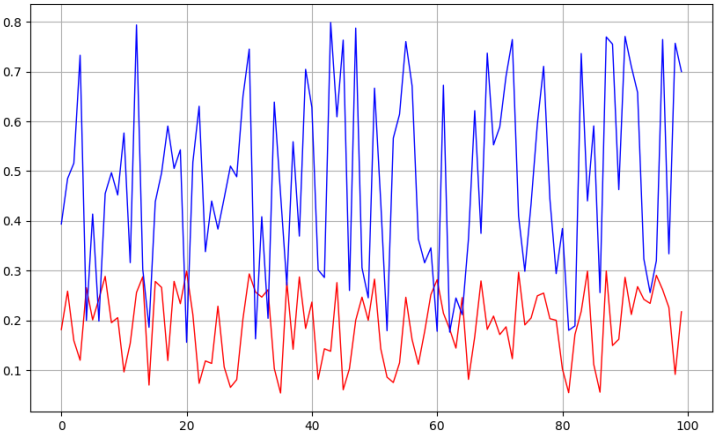
Случайный Лес



Наивный Байес



Деревья Принятия решений



После атаки - красные линии. После нашего подхода – синие линии.

# Обсуждение

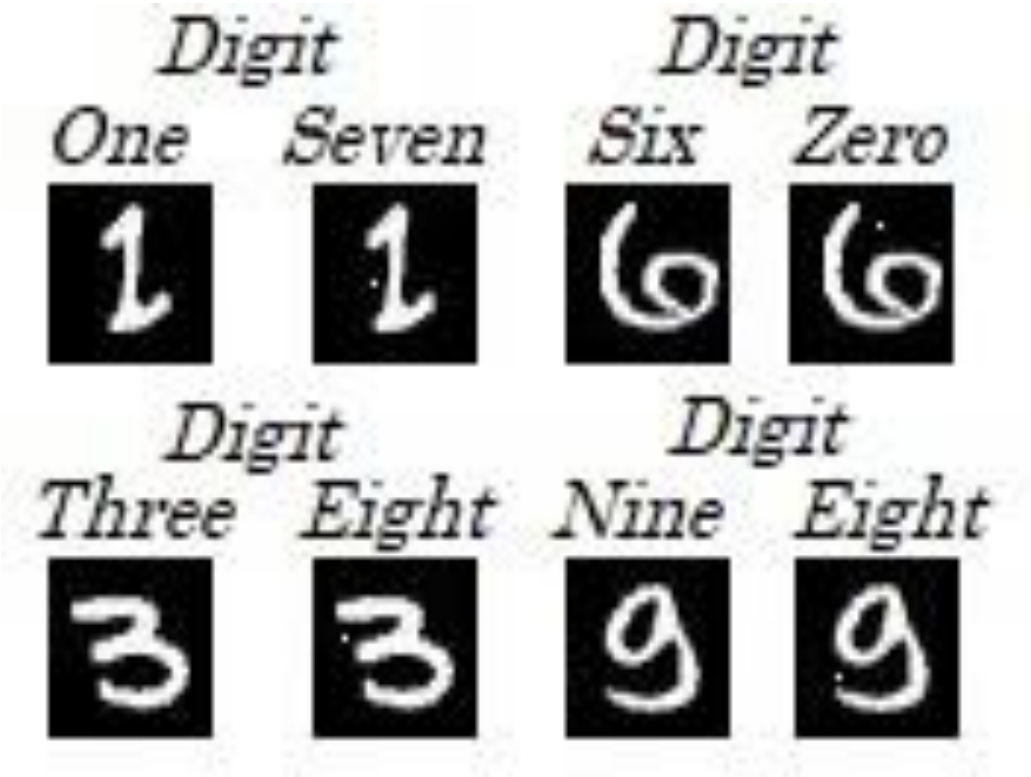
| Classifier | Before the attack | After the attack | After our approach |
|---|---|---|---|
| K-Nearest Neighbors | 0.698 | 0.182 | 0.545 |
| Random Forest | 0.878 | 0.321 | 0.722 |
| Naive Bayes Classifier | 0.789 | 0.209 | 0.643 |
| Decision Trees | 0.756 | 0.244 | 0.658 |

Использование предложенного подхода для устранения последствий атаки ZOO позволяет нам значительно восстановить эффективность распознавания изображений.

Для классификаторов наивного Байеса и деревьев принятия решений, эффективность распознавания была всего лишь 0.146 и 0.098 Меньше чем исходное значение, соответственно.

Наиболее эффективная точность распознавания осталась за случайным лесом и составила 0.722.

# Атака одним пикселем на набор данных MNIST-JPG




**Атака OPA** находит и изменяет один пиксель.

Это приводит к максимальному изменению классов распознавания.

# Влияние OPA на набор данных MNIST-JPG

Средняя точность распознавания до атаки OPA и после OPA для набора данных MNIST-JPG.

| Классификатор | Средняя точность распознавания до атаки | Средняя точность распознавания после атаки |
|---|---|---|
| K-Nearest Neighbors | 0.698 | 0.412 |
| Random Forest | 0.878 | 0.432 |
| Naive Bayes | 0.789 | 0.387 |
| Decision Trees | 0.756 | 0.335 |

# Результаты нашего подхода (1/2)

Анализ средней точности распознавания был проведен после применения подхода для классификатора случайный лес, поскольку он показал высокую точность.

В таблице показан результат применения данного подхода к противодействию атаке одного пикселя для набора данных MNIST-JPG.

| № | Стандартное отклонение гауссовского шума | Среднее значение для шума Пуассона | Средняя точность распознавания после нашего подхода |
|---|---|---|---|
| 1 | 20 | 0,2 | 0,434 |
| 2 | 30 | 0,25 | 0,446 |
| 3 | 40 | 0,30 | 0,587 |
| 4 | 50 | 0,35 | 0,494 |
| 5 | 60 | 0,4 | 0,433 |
| 6 | 70 | 0,5 | 0,326 |

# Результаты нашего подхода (2/2)

Выбор из оптимальный Параметры гауссовского и пуассоновского шума зависит от атака:

| Датасет | Атака | Наилучшее стандартное отклонение шума Гаусса | Наилучшее среднее значения шума Пуассона |
|---|---|---|---|
| PC Parts Images | FGSM | 40 | 0,3 |
| | ZOO | 30 | 0,25 |
| | One Pixel | атака не оказывает существенного эффекта на точность распознавания | |
| MNIST-JPG | FGSM | 60 | 0,4 |
| | ZOO | 50 | 0,25 |
| | One Pixel | 40 | 0,3 |

# Заключение

1. Были проанализированны атаки на системы ML и средства защиты от них.
2. Предложен подход для защиты от враждебных атак основанный на Neural-Cleanse, сжатии в формате Jpeg и добавление шума.
3. Этот подход был изучен на двух наборах данных (MNIST-JPG и детали ПК) и для трех атак (FSGM, ZOO, OPA).
4. Предложенный подход позволяет существенно восстановить эффективность распознавания изображений. Это было подтверждено экспериментами на классификаторах Bagging, Random Forest, Extra Trees and Gradient Boosting (для MNIST-JPG) и K-Ближайшие соседи, Случайный Лес, Наивный Байес и Деревья принятия решений (для IP-адрес ПК).
5. Были выявлены оптимальные значения параметров гауссовского и пуассоновского шума для FGSM, ZOO и One Pixel.

# Контакты

# Спасибо за внимание!

Игорь Саенко

ibsaen@comsec.spb.ru

Владимир Садовников

bladimir1998@mail.ru