Объяснения моделей машинного обучения и состязательные атаки

М.Э. Егоров, И.А Зянчурин, И. Д. Кузьменко, Д.Д. Тарасов, Д.Е. Намиот

***Аннотация*—В работе рассматривается практическое построение состязательных атак (атак уклонения) на модели машинного обучения с использованием объяснений их работы. Несмотря на то, что модели машинного обучения, в общем случае, представляют собой черный ящик, существуют схемы построения объяснений (их приближений), которые позволяют оценить, как именно принимается решение. Даже если наша модель не является решающим деревом, мы можем получить похожее по смыслу объяснение принятия решений в модели. Одним из примеров подобного рода схем является использование SHAP-значений. Подобный подход позволяет формировать атаки в режиме черного ящика. Если известен тренировочный набор данных атакуемой модели или даже часть его, то атакующий может использовать его для тренировки своей модели произвольной архитектуры. Далее для этой модели можно построить объяснения, которые и использовать для формирования атаки. Поскольку состязательные атаки переносимы, то такие атаки можно будет воспроизвести уже н на атакуемой модели. Исходный код для такого рода экспериментов и приводится в данной статье. В качестве атакуемого примера рассматриваются модели классификации сетевого трафика в системе Интернета Вещей.**

***Ключевые слова*—состязательные атаки, IoT, SHAP.**

I. Введение

Модели машинного (глубокого) обучения подвержены состязательным атакам, под которыми понимаются специальные модификации данных на разных этапах стандартного конвейера машинного обучения, призванные изменить работу модели [1]. В данном случае мы говорим об атаках, которые осуществляются на этапе вывода – так называемых атаках уклонения [2].

Существуют разные подходы к их формированию (поиску необходимых модификаций данных), которые зависят, например, от того, что мы знаем об атакуемой модели. Наиболее сложным для атакующего случаем являются так называемые атаки черного ящика, когда атакующий может оперировать только откликом модели на поданные данные. И здесь большим подспорьем в организации атаки является знание тренировочного набора (или хотя бы его части) атакуемой модели. Любая модель после обучения отражает статистические характеристики своего тренировочного набора. Соответственно, зная тренировочный набор, атакующий может подготовить свою собственную модель, которая будет для него уже белым ящиком. На этой теневой модели можно подобрать атаку (модификации исходных данных), которую затем уже перенести на реальную атакуемую модель. К сожалению, состязательные атаки переносимы [3], и подобная практика описывается в атласе MITRE [4]. Примеров подобному уже много [5-8].

Особенно опасно этот подход проявляет себя, когда в промышленной эксплуатации оказываются модели, обученные на публично доступных датасетах. Если известно, что некоторая модель обучена на конкретном публичном датасете, то это серьезно упрощает задачу для атакующих – публичный датасет для обучения теневой модели доступен и им. Отсюда следует общее правило, что информация о тренировочных данных не должна быть публичной. Сами данные могут, естественно, быть открытыми, а вот информация об их использовании в конкретном промышленном проекте – нет.

В данной работе как раз и описывается подобного рода эксперимент. Ниже представлены два проекта, в которых решается следующая задача. На открытом датасете строится модель классификации сетевого трафика (в каждом проекте – своя модель) для определения DOS (DDOS) атак. Для полученной модели строится объяснение ее работы с использованием SHAP, и исходя из данных последнего, формируется представление о необходимых (возможных) модификациях, которые будут “обманывать” построенный классификатор. Оба проекта представлены в исходных кодах и могут быть использованы как основа для последующих разработок.

В качестве исходного датасета использовался набор данных, представленный на сайте Kaggle [9]. Датасет содержит информацию об атаках на сетевую компоненту в IoT. Он небольшой, и обработка не требует больших мощностей. В данных присутствует несколько подкатегорий для каждого типа вторжений в IoT. Набор данных содержит 1048575 экземпляра записей для вторжений и 47 признаков каждого из вторжений. Набор данных может быть использован для подготовки прогностической модели, с помощью которой могут быть обнаружены различные виды интрузивных атак. Данные также подходят для проектирования системы детектирования вторжений (IDS). Можно найти довольно много моделей, обученных на данном датасете [10].

В датасете размечены следующие наиболее частые метки атак:

- DDos
- DoS
- Spoofing

- Mirai (это такой botnet)
- Recon (разведка)

Идея эксперимента заключалась в следующем.

1. Построить классификатор для одной из атак
2. Получить объяснения для модели с использованием SHAP [11, 12]
3. Используя эти объяснения построить атаку (определить модификации данных, которые обходят построенный классификатор)
4. Проверить эту атаку на синтетических данных из пункта 3

Один из интересных моментов – это сравнить объяснения. Если, например, они окажутся совпадающими для разных атак, то, очевидно, способность моделей их различать будет сомнительна.

Отметим также, что состязательные атаки в данном случае являются примерами атак с ограничениями [13]. Не все атрибуты записей лога, очевидно, могут изменяться неограниченно. Мы можем, например, менять время в синтетических данные, но, например, IP адреса и флаги пакетов нельзя менять произвольно.

Оставшаяся часть статьи структурирована следующим образом. В разделе II кратко описываются схемы объяснений для моделей машинного обучения. В разделе III описываются атакуемые классификаторы. Раздел IV содержит описание объяснений и принципы построения состязательных атак.

II. Использование SHAP для интерпретации моделей

SHAP (SHapley Additive exPlanations) — это метод объяснения выходных данных моделей машинного обучения, дающий представление о том, как каждый признак (“фича”) вносит вклад в конкретное предсказание. Он основан на теории игр и использует так называемые значения Шепли (SHAP-values), которые справедливо распределяют предсказание модели среди признаков (определяют вклад каждого признака в результат модели). SHAP помогает понять как отдельные предсказания (локальные объяснения), так и общее поведение модели (глобальные объяснения).

Проще говоря, значения SHAP позволяют разбить прогнозы модели машинного обучения, присваивая каждому признаку (каждой “фиче”) оценку его вклада в конечный результат.

Таким образом, независимо от того, что мы прогнозируем с помощью машинного обучения, значения SHAP сообщают, какие именно признаки управляют этими прогнозами.

Отличительной чертой SHAP является использование методов атрибуции аддитивных признаков, то есть разложение прогноза (результата работы модели) на сумму вкладов каждого признака, точно совпадающих с выходными данными модели.

III. Классификаторы атак

Для построения классификаторов атак на основе данных IoT была выбрана задача определения DoS и DDoS-атак, как одних из наиболее распространенных угроз в области Интернета вещей. Для реализации классификаторов использовалась модель на основе алгоритма случайного леса (Random Forest). Данные для обучения и тестирования модели были взяты из общедоступного датасета IoT Intrusion, размещенного на платформе Kaggle [9]. Этот датасет содержит 1048575 записей, включающих 47 признаков, описывающих различные параметры сетевого трафика (рис.1).

	flow_duration	Header_Length	Protocol Type	Duration	Rate	Srate	Drate	fin_flag_number	syn_flag_number	rst_flag_number	...
0	0.000000	54.00	6.00	64.00	0.329807	0.329807	0.0	1	0	1	...
1	0.000000	57.04	6.33	64.00	4.290556	4.290556	0.0	0	0	0	...
2	0.000000	0.00	1.00	64.00	33.396799	33.396799	0.0	0	0	0	...
3	0.328175	76175.00	17.00	64.00	4642.133010	4642.133010	0.0	0	0	0	...
4	0.117320	101.73	6.11	65.91	6.202211	6.202211	0.0	0	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...
1048570	1.391925	108.00	6.00	64.00	1.437685	1.437685	0.0	0	1	0	...
1048571	0.000000	2.14	46.70	65.91	0.000000	0.000000	0.0	0	0	0	...
1048572	0.132971	30847.00	17.00	64.00	5978.034950	5978.034950	0.0	0	0	0	...
1048573	0.000000	54.00	6.00	64.00	25.672981	25.672981	0.0	1	0	1	...
1048574	128.443556	4264.30	7.10	98.80	13.640648	13.640648	0.0	0	0	0	...

1048575 rows × 47 columns

Рис. 1. Датасет IoT Intrusion.

В ходе предварительной обработки данных были исключены признаки с низкой информативностью и произведена нормализация численных признаков для повышения устойчивости модели к вариациям данных. Данные были разделены на обучающую (80%) и тестовую выборки (20%) с использованием стратифицированного разделения для сохранения пропорций классов.

Модели случайного леса были обучены с использованием библиотеки scikit-learn. Для оценки производительности классификаторов были выбраны следующие метрики: точность (Accuracy), полнота (Recall), точность предсказания положительного класса (Precision) и F1-score. По результатам тестирования, модели показали высокие показатели точности (Accuracy более 99%), что подтверждает их эффективность в идентификации DoS и DDoS-атак.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	173419
1	1.00	1.00	1.00	36296
accuracy			1.00	209715
macro avg	1.00	1.00	1.00	209715
weighted avg	1.00	1.00	1.00	209715

```
[[173416      3]
 [    30  36266]]
```

Рис.2а. Тестовые метрики обученной модели DoS.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	57010
1	1.00	1.00	1.00	152705
accuracy			1.00	209715
macro avg	1.00	1.00	1.00	209715
weighted avg	1.00	1.00	1.00	209715

Рис.2б. Тестовые метрики обученной модели DDoS.

Исходный код классификаторов представлен в открытом доступе на платформе GitHub [14, 15].

IV. Объяснения и атаки

Для интерпретации работы полученного классификатора был применён метод SHAP (SHapley Additive exPlanations). Этот подход позволяет оценить вклад каждого признака в предсказание модели и выявить наиболее значимые атрибуты сетевого трафика, которые оказывают наибольшее влияние на результат классификации.

Проведенный анализ SHAP-значений показал (Рис. 3а), что признаки IAT (разница во времени с предыдущим пакетом), Protocol Type (IP, UDP, TCP, IGMP, ICMP, Unknown), Magnitude (среднее значение длин входящих пакетов в потоке + среднее значение длин исходящих пакетов в потоке), AVG (средняя длина пакета в потоке), ICMP (индикатор того, что протокол сетевого уровня - ICMP) имеют наибольшее влияние на принятие решения моделью о принадлежности трафика к классу DoS-атаки.

Анализ SHAP-значений для DDoS (Рис. 3б) указывает на IAT, Protocol Type, Magnitude как и в DoS, но так же добавляет header_length и tot_sum (суммарное количество пакетов).

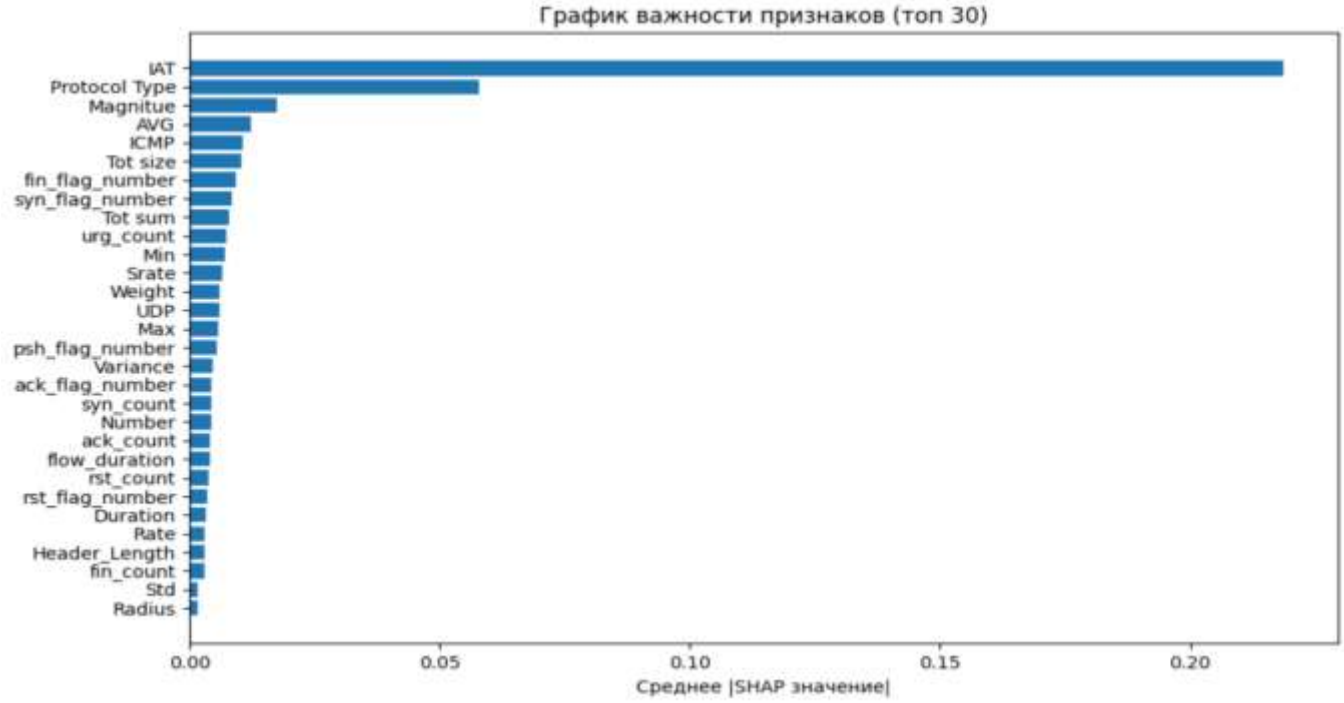


Рис. 3а. Средние SHAP-значения для каждого параметра (DoS).

Далее приведён пример локального объяснения предсказания моделей классификации DoS и DDoS-атак с использованием метода SHAP (Рис. 4а, 4б). Графики отображает вклад признаков в конкретное предсказание модели для одного наблюдения из тестовой выборки. Значения SHAP показывают, насколько каждый признак увеличивает (красный цвет) или уменьшает (синий цвет) итоговое значение выхода модели f(x), относительно среднего значения предсказания E[f(X)].

Наибольшее влияние на решение оказал признак **IAT**, который значительно сместил предсказание модели в сторону негативного класса (нормального трафика), с вкладом около -0.14 для DoS и -0.36 для DDoS. Также значимый отрицательный вклад внесли признаки **UDP**, **AVG**, **ICMP**, **Srate** и **ack_flag_number** для DoS и **header_length, protocol_type, udp, icmp и srate для DDoS**.

С другой стороны, признаки **Protocol Type**, **syn_flag_number**, **urg_count** и **Min** для DoS и **magnitude, tot_sum и avg** для DDoS имели небольшой

положительный вклад, способствуя смещению результата в сторону класса атаки.

Подобные графики позволяют точно определить, какие признаки «отвечают» за итоговое решение модели. На их основе формируется стратегия изменения значений ключевых признаков — например, снижение значения IAT или эмуляция других сетевых характеристик — с целью обхода классификатора. Этот подход позволяет создавать целевые состязательные примеры, практически не отличающиеся от нормального трафика, но классифицируемые моделью как безопасные.

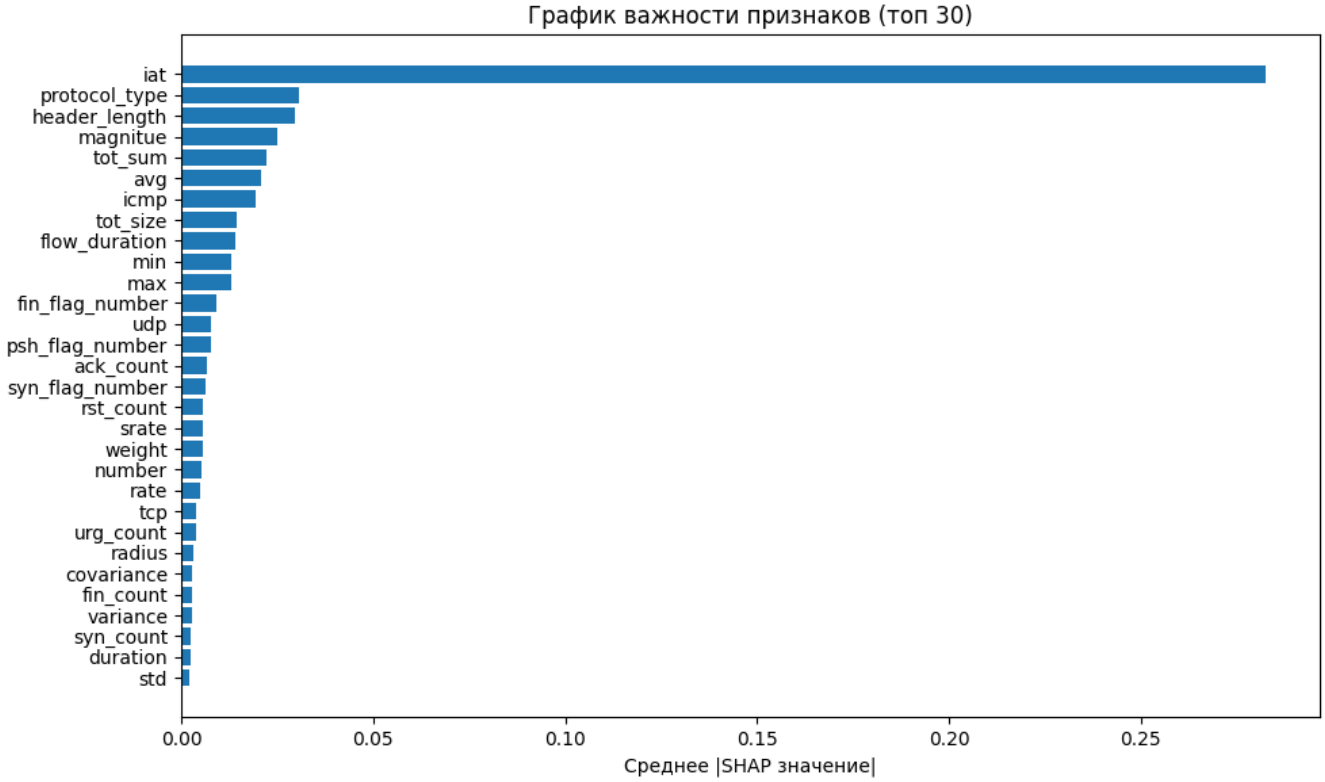


Рис. 3б. Средние SHAP-значения для каждого параметра (DDoS).

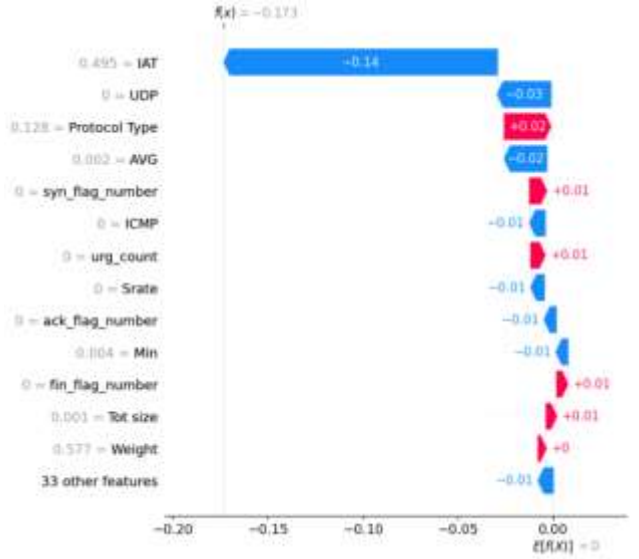


Рис. 4а. SHAP-объяснение предсказания классификатора DoS для одного примера сетевого трафика.

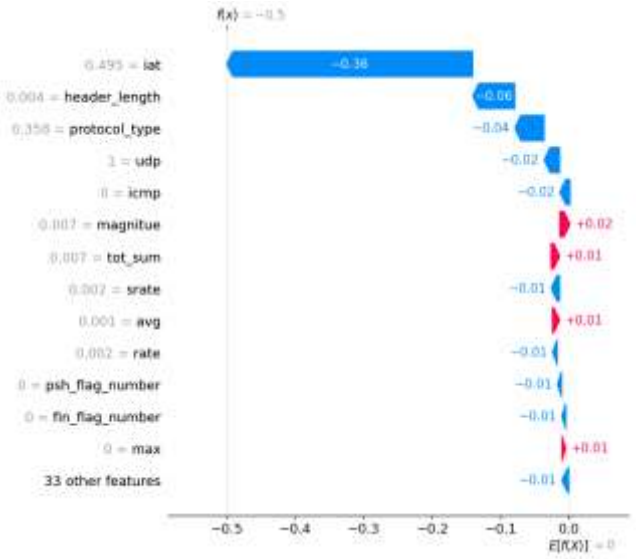


Рис. 4б. SHAP-объяснение предсказания классификатора DDoS для одного примера сетевого трафика.

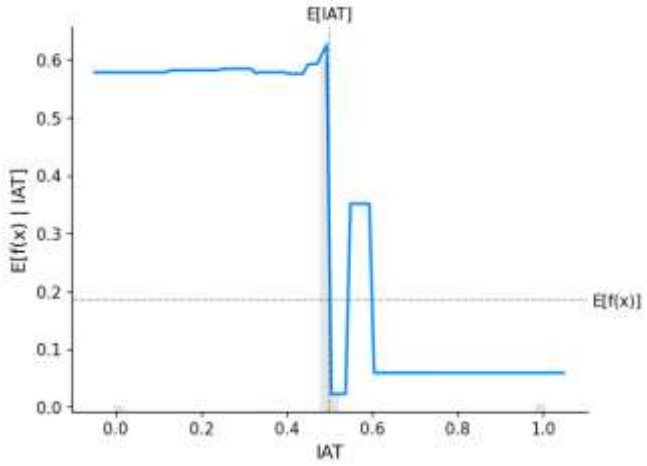


Рис. 5а. Влияние признака IAT на DoS.

График на рис.5а демонстрирует, что значения признака IAT выше 0.5 приводят к резкому снижению предсказания модели. Это означает, что высокие интервалы между пакетами интерпретируются моделью как нетипичные для DoS-атак. Таким образом, увеличение значения IAT может быть использовано для обхода модели при формировании атакующего примера.

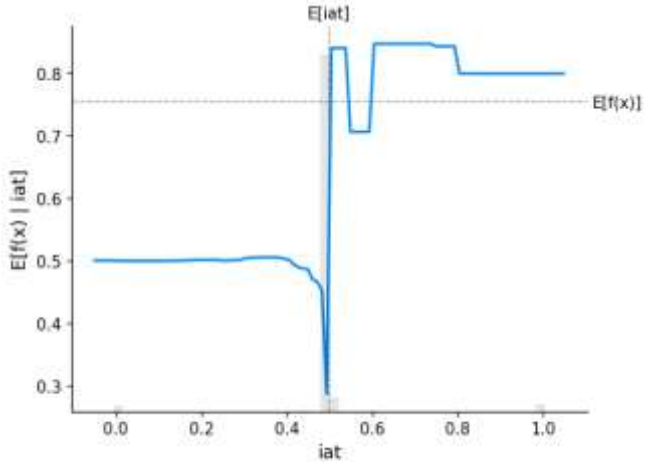


Рис. 5б. Влияние признака IAT на DDoS.

График на рис.5б демонстрирует, что значения признака IAT выше 0.5 приводят к резкому повышению предсказания модели. Это означает, что высокие интервалы между пакетами интерпретируются моделью как типичные для DDoS-атак. Таким образом, уменьшение значения IAT может быть использовано для обхода модели при формировании атакующего примера.

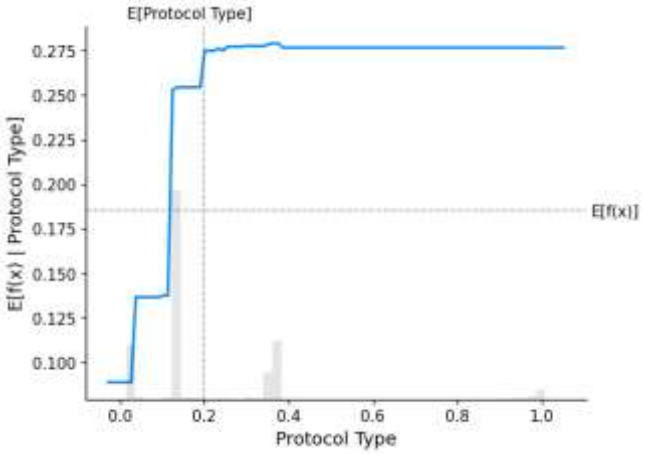


Рис. 6а. Влияние признака Protocol Type на DoS.

На рис.6а показана положительная корреляция между значением признака Protocol Type и предсказанием модели. При переходе от 0 к значению около 0.2 наблюдается резкий рост значения *E[f(x)]*, после чего предсказание стабилизируется на высоком уровне. Это указывает на важность данного признака для классификации атак: определённые протоколы явно ассоциируются у модели с вредоносной активностью.

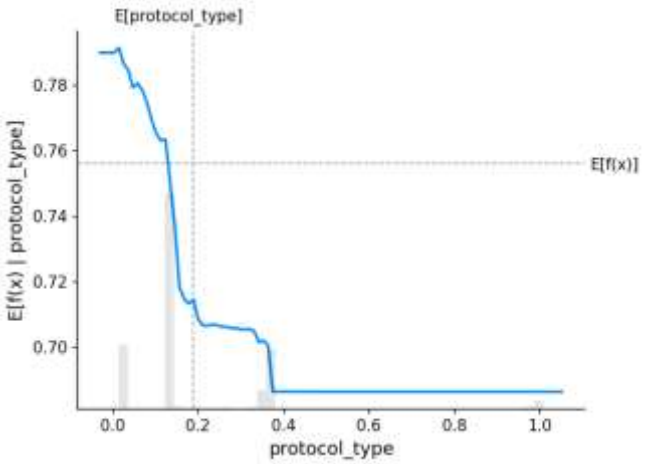


Рис. 6б. Влияние признака Protocol Type на DDoS.

На рис.6б показана отрицательная корреляция между значением признака Protocol Type и предсказанием модели. При переходе от 0 к значению около 0.4 наблюдается снижение значения *E[f(x)]*, после чего предсказание стабилизируется на низком уровне. Это указывает на важность данного признака для классификации атак: определённые протоколы явно ассоциируются у модели с вредоносной активностью.

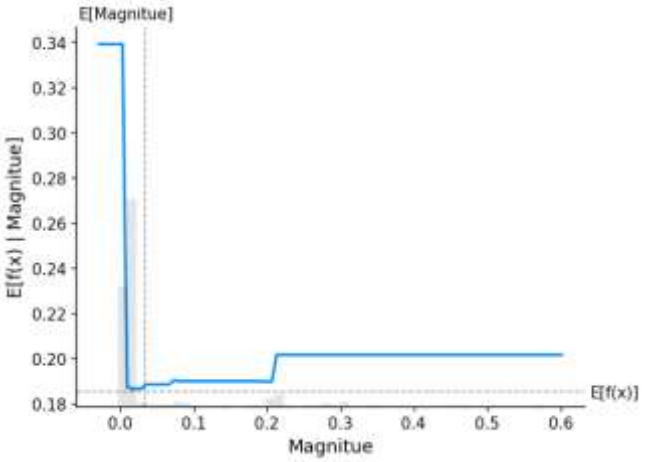


Рис. 7а. Влияние признака Magnitude на DoS.

Признак Magnitude (Рис. 7а) - среднее значение длин входящих пакетов в потоке + среднее значение длин исходящих пакетов в потоке, оказывает сильное влияние при значениях, близких к нулю. Повышение значения признака резко снижает вероятность классификации как атаки. Это подтверждает, что DoS-атаки часто связаны с пиками в интенсивности трафика.

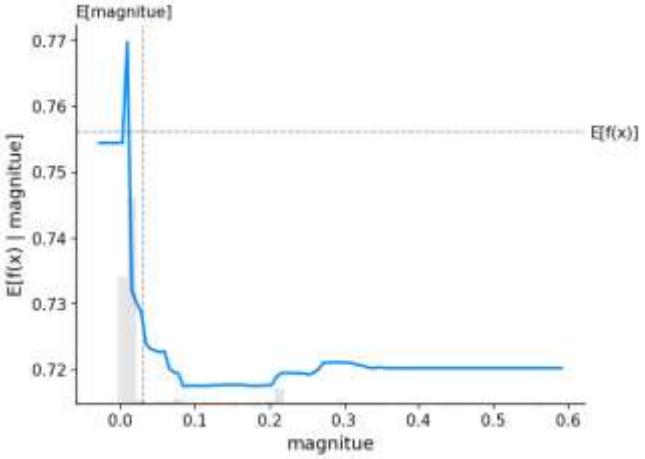


Рис. 7б. Влияние признака Magnitude на DDoS.

Признак Magnitude (Рис. 7б) оказывает сильное влияние при значениях, близких к нулю. Повышение значения признака немного снижает вероятность классификации как атаки. Это подтверждает, что модель DDoS интерпретирует потоки с низкой средней длиной пакетов как потенциально вредоносные.

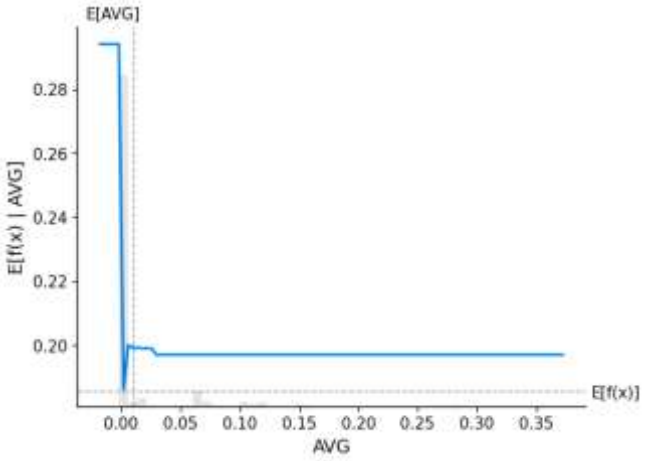


Рис. 8. Влияние признака AVG на DoS.

График на рис. 8 показывает, что даже небольшие значения AVG значительно увеличивают вероятность того, что трафик будет классифицирован как DoS-атака. При превышении значения примерно 0.01 вклад в предсказание существенно снижается, что делает этот признак уязвимым для модификации при построении состязательных примеров.

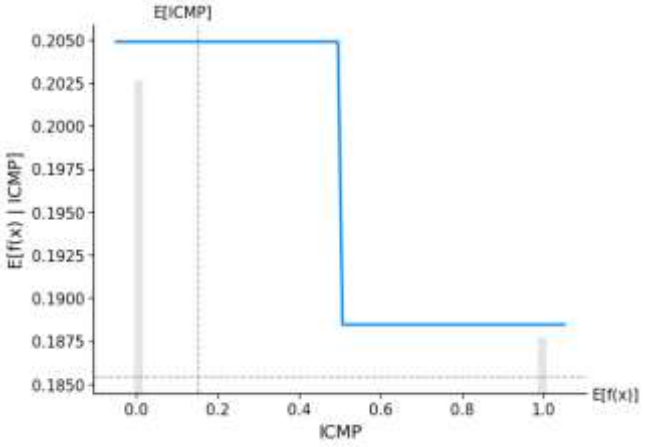


Рисунок 9. Влияние признака ICMP на DoS.

Признак ICMP (Рис. 9) имеет бимодальное распределение (0 или 1), отражающее факт использования протокола ICMP. Видно, что при включённом ICMP (значение 1) модель понижает вероятность отнесения к атаке. Это может быть использовано злоумышленником.

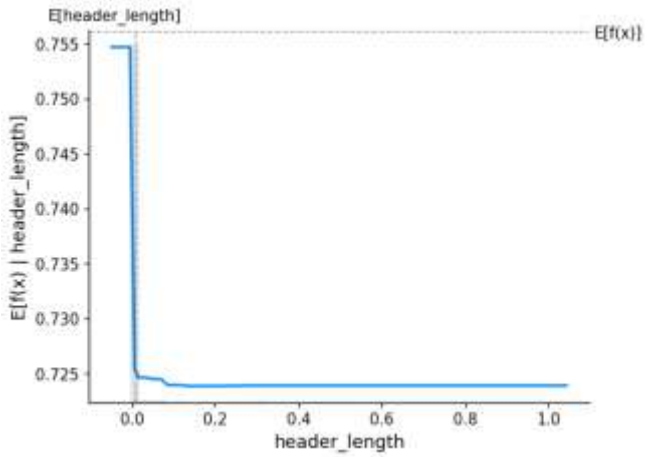


Рисунок 10. Влияние признака Header Length на DDoS.

Признак Header Length (Рис. 10) демонстрирует выраженное влияние на вероятность классификации как атаки при крайне низких значениях. При значениях, близких к нулю, предсказание модели достигает максимума, после чего резкое увеличение длины заголовка приводит к стремительному снижению предсказания. Это указывает на то, что DDoS-атаки в выборке модели чаще ассоциируются с короткими заголовками пакетов.

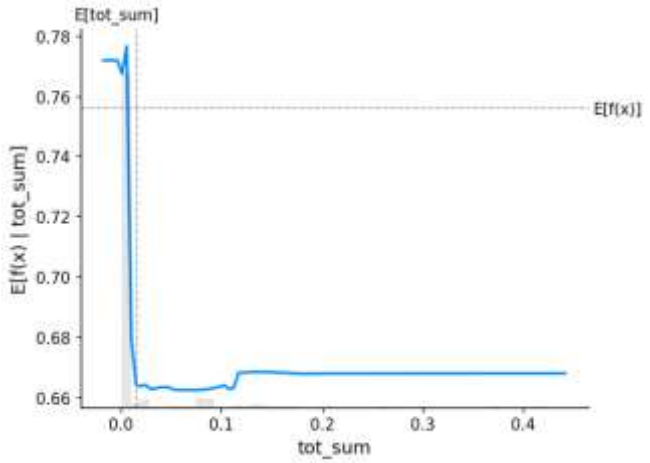


Рисунок 11. Влияние признака Tot Sum на DDoS.

Признак Tot Sum (Рис. 7г), отражающий суммарный объём входящего и исходящего трафика в потоке, оказывает существенное влияние на поведение модели при низких значениях. При близких к нулю значениях вероятность классификации как атаки достигает максимума. С ростом признака наблюдается резкое снижение предсказания. Это указывает на то, что модель интерпретирует низкую суммарную нагрузку в потоке как характерную для DDoS-атаки.

На основе полученных объяснений была сформирована стратегия для построения состязательной атаки.

Для генерации состязательных примеров применён жадный алгоритм, задача которого — минимально модифицировать признаки входного объекта так, чтобы классификатор изменил своё решение. Алгоритм последовательно перебирает признаки и направления изменения, выбирая в каждый момент шаг, максимально увеличивающий вероятность противоположного класса.

В рамках эксперимента было проведено 100 попыток генерации контрафактных записей для объектов исходного класса DoS (1). В 97 из 100 случаев модель была успешно обманута и сменила своё решение на противоположное. Для DDoS так же было проведено 100 попыток, получилось обмануть модель 68 раз из 100.

```
Исходный класс: 1 → Контрфакт предсказан как: 0

Изменённые признаки:
      Оригинал  Контрфакт
IAT  0.494796   0.499796

Исходный класс: 1 → Контрфакт предсказан как: 0

Изменённые признаки:
      Оригинал  Контрфакт
IAT  0.495207   0.500207

Исходный класс: 1 → Контрфакт предсказан как: 0

Изменённые признаки:
      Оригинал  Контрфакт
IAT  0.494663   0.499663

Исходный класс: 1 → Контрфакт предсказан как: 0

Изменённые признаки:
      Оригинал  Контрфакт
IAT  0.494792   0.499792
...

Изменённые признаки:
      Оригинал  Контрфакт
IAT  0.494847   0.499847
```

Рис. 12а. Создание контрфактов для DoS, изменяя параметры жадным алгоритмам.

```
Исходный класс: 1 → Контрфакт предсказан как: 0

Изменённые признаки:
               Оригинал  Контрфакт
header_length  0.000000   0.005000
protocol_type  0.021064   0.026064
tot_sum        0.005593   0.010593
iat            0.495899   0.490899
magnitue       0.000047   0.015047

Исходный класс: 1 → Контрфакт предсказан как: 0

Изменённые признаки:
      Оригинал  Контрфакт
iat  0.495667   0.490667

Исходный класс: 1 → Контрфакт предсказан как: 0

Изменённые признаки:
      Оригинал  Контрфакт
iat  0.497289   0.492289

Исходный класс: 1 → Контрфакт предсказан как: 0
...
protocol_type  0.127660   0.137660
tot_sum        0.007343   0.012343
iat            0.497144   0.492144
magnitue       0.010575   0.000575
```

Рис. 12б. Создание контрфактов для DDoS, изменяя параметры жадным алгоритмам.

Для успешных атак на DoS зачастую достаточно было минимального изменения одного или двух признаков (Рис. 13а). Например, при незначительном увеличении признака IAT с 0.4951 до 0.5001 модель переставала классифицировать запись как DoS-атаку. В других случаях комбинировались малые изменения таких признаков, как AVG и Magnitude, выявленных ранее как значимые на основе SHAP-значений.

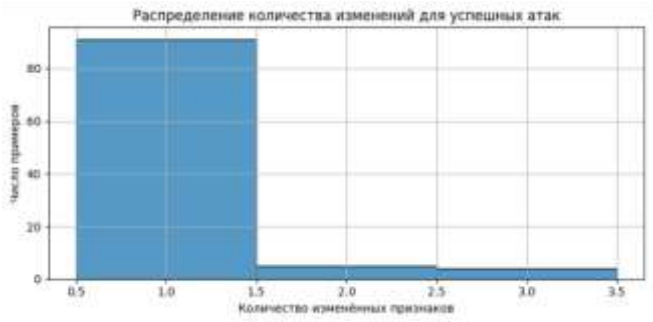


Рис. 13а. Распределение количества изменений параметров для успешных атак на DoS.

Для успешных атак на классификатор DDoS требовалось более сложное воздействие на модель (Рис. 13б). В отличие от DoS, здесь редко удавалось изменить решение модели, варьируя лишь один признак. Распределение показывает, что в большинстве случаев модифицировались сразу несколько признаков — часто четыре или пять. В примерах успешных атак комбинировались изменения признаков iat, tot_sum, header_length, magnitude, и protocol_type, что соответствует их высокому вкладу в предсказание класса атаки, выявленному с помощью SHAP-анализа.

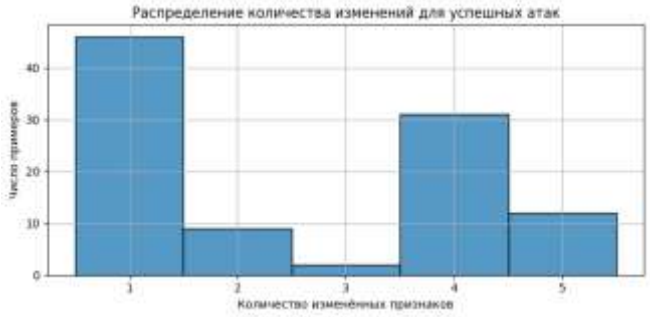


Рис. 13б. Распределение количества изменений параметров для успешных атак на DDoS.

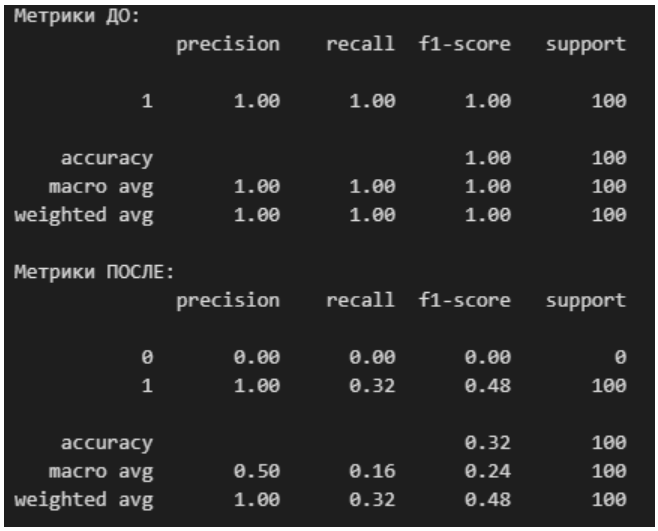
Далее представлены метрики до и после применения атак (Рис. 14а, 14б). До атаки модель демонстрировала 100% точность и полноту на классе атак. После атак — точность упала практически до нуля для DoS: полнота по DoS классу составила всего 3%, а общая точность — 0.03, что подтверждает успешность атак. Для DDoS удалось опустить точность до 0.32.

```
Метрики ДО:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00       100

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100

Метрики ПОСЛЕ:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         0
           1       1.00      0.03      0.06       100

    accuracy                           0.03       100
   macro avg       0.50      0.01      0.03       100
weighted avg       1.00      0.03      0.06       100
```

Рис. 14а. Метрики ДО/ПОСЛЕ совершения атак для DoS.

Эти результаты подтверждают, что даже малые изменения ключевых признаков, выявленных с помощью интерпретации модели, могут быть использованы для построения эффективных атак, демонстрируя уязвимость моделей даже при отсутствии знаний об их архитектуре и параметрах. Более того, они подтверждают гипотезу о переносимости состязательных атак: атака, построенная на основе объяснений для одной модели, может быть успешно применена против другой аналогичной модели.

Исходный код и результаты экспериментов также доступны на GitHub [14, 15].



Метрики ДО:

	precision	recall	f1-score	support
1	1.00	1.00	1.00	100
accuracy			1.00	100
macro avg	1.00	1.00	1.00	100
weighted avg	1.00	1.00	1.00	100

Метрики ПОСЛЕ:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	0
1	1.00	0.32	0.48	100
accuracy			0.32	100
macro avg	0.50	0.16	0.24	100
weighted avg	1.00	0.32	0.48	100

Рис. 14б. Метрики ДО/ПОСЛЕ совершения атак для DDoS.

V. Заключение

В работе представлен модельный пример использования объяснения моделей машинного обучения для построения состязательных атак. Такой подход является одним из самых реалистичных для формирования атак в режиме черного ящика (то есть, без использования какой-либо информации об атакуемой модели). Технически, даже части исходного тренировочного набора данных будет достаточно, чтобы обучить теневую модель, построить для нее объяснения, и сформировать на этой основе состязательную атаку. Также, данная работа может служить примером формирования состязательных атак с ограничениями. Атаки на модели оценки сетевого трафика являются типичным примером.

Благодарности

Статья написана в рамках развития направления «Кибербезопасность» на факультете ВМК МГУ имени М.В. Ломоносова [16]. Представленные проекты были выполнены в рамках учебного курса “Безопасность инфраструктурных технологий” [17]. Работа продолжает серию публикаций по использованию ИИ в кибербезопасности, начатую в работе [18].

Библиография

[1] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22.

[2] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[3] Zhao, Zhengyu, et al. "Towards good practices in evaluating transfer adversarial attacks." arXiv preprint arXiv:2211.09565 (2022).

[4] Navigate threats to AI systems through real-world insights https://atlas.mitre.org/ Retrieved: Jun, 2025

[5] Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

[6] Fidel, Gil, Ron Bitton, and Asaf Shabtai. "When explainability meets adversarial learning: Detecting adversarial examples using shap signatures." 2020 international joint conference on neural networks (IJCNN). IEEE, 2020.

[7] Hickling, Thomas, Nabil Aouf, and Phillippa Spencer. "Robust adversarial attacks detection based on explainable deep reinforcement learning for UAV guidance and planning." IEEE Transactions on Intelligent Vehicles 8.10 (2023): 4381-4394.

[8] Aryal, Kshitiz, et al. "Explainability guided adversarial evasion attacks on malware detectors." 2024 33rd International Conference on Computer Communications and Networks (ICCCN). IEEE, 2024.

[9] IoT Intrusion https://www.kaggle.com/datasets/subhajournal/iotintrusion/data Retrieved: Jun, 2025

[10] CICIoT2023 models https://www.google.com/search?q=CICIoT2023 Retrieved: Jun, 2025

[11] An introduction to explainable AI with Shapley values https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html Retrieved: Jun, 2025

[12] Christoph Molnar Interpretable Machine Learning. A Guide for Making Black Box Models Explainable https://christophm.github.io/interpretable-ml-book/

[13] Grini, Anass, et al. "Constrained Network Adversarial Attacks: Validity, Robustness, and Transferability." arXiv preprint arXiv:2505.01328 (2025).

[14] GitHub 1, 2025. Available: https://github.com/lava-aaa/iot_hw

[15] GitHub 2, 2025. Available:https://github.com/Dark-Avery/DDoS_classifier

[16] Sukhomlin, Vladimir A. "The Concept and Main Characteristics of the Master's Degree Program" Cybersecurity" of the Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University." International Journal of Open Information Technologies 11.7 (2023): 143-148.

[17] Namiot, Dmitry, and Vladimir Sukhomlin. "On cybersecurity of the Internet of Things systems." International Journal of Open Information Technologies 11.2 (2023): 85-97.

[18] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Artificial intelligence and cybersecurity." International Journal of Open Information Technologies 10.9 (2022): 135-147.

Статья получена 15.06.2025

М.Э. Егоров – МГУ имени М.В. Ломоносова (email: 662366@bk.ru)

И.А Зянчурин – МГУ имени М.В. Ломоносова (email: ingv0rr@yandex.ru)

И. Д. Кузьменко – МГУ имени М.В. Ломоносова (email: ilyexakuzmenko@gmail.com)

Д.Д. Тарасов – МГУ имени М.В. Ломоносова (email: s02240560@gse.cs.msu.ru)

Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)

Machine Learning Models Explanations and Adversarial Attacks

Maxim Egorov, Igor Zianchurin, Ilya Kuzmenko, Danil Tarasov, Dmitry Namiot

***Abstract*— The paper considers the practical construction of adversarial attacks (evasion attacks) on machine learning models using explanations of their operation. Despite the fact that machine learning models, in general, are a black box, there are schemes for constructing explanations (their approximations) that allow us to evaluate how exactly a decision is made. Even if our model is not a decision tree, we can obtain a similar explanation for decision making in the model. One example of such schemes is the use of SHAP values. Such an approach allows us to form attacks in black box mode. If the training dataset of the attacked model or even a part of it is known, the attacker can use it to train his model of arbitrary architecture. Then, explanations can be constructed for this model, which can be used to form an attack. Since adversarial attacks are portable, such attacks can be reproduced on the attacked model. The source code for such experiments is given in this paper. Network traffic classification models in the Internet of Things system are considered as an attackable example.**

***Keywords* – adversarial attacks, IoT, SHAP**

REFERENCES

[1] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22.

[2] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[3] Zhao, Zhengyu, et al. "Towards good practices in evaluating transfer adversarial attacks." arXiv preprint arXiv:2211.09565 (2022).

[4] Navigate threats to AI systems through real-world insights https://atlas.mitre.org/ Retrieved: Jun, 2025

[5] Slack, Dylan, et al. "Fooling lime and shap: Adversarial attacks on post hoc explanation methods." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

[6] Fidel, Gil, Ron Bitton, and Asaf Shabtai. "When explainability meets adversarial learning: Detecting adversarial examples using shap signatures." 2020 international joint conference on neural networks (IJCNN). IEEE, 2020.

[7] Hickling, Thomas, Nabil Aouf, and Phillippa Spencer. "Robust adversarial attacks detection based on explainable deep reinforcement learning for UAV guidance and planning." IEEE Transactions on Intelligent Vehicles 8.10 (2023): 4381-4394.

[8] Aryal, Kshitiz, et al. "Explainability guided adversarial evasion attacks on malware detectors." 2024 33rd International Conference on Computer Communications and Networks (ICCCN). IEEE, 2024.

[9] IoT Intrusion https://www.kaggle.com/datasets/subhajournal/iotintrusion/data Retrieved: Jun, 2025

[10] CICIoT2023 models https://www.google.com/search?q=CICIoT2023 Retrieved: Jun, 2025

[11] An introduction to explainable AI with Shapley values https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html Retrieved: Jun, 2025

[12] Christoph Molnar Interpretable Machine Learning. A Guide for Making Black Box Models Explainable https://christophm.github.io/interpretable-ml-book/

[13] Grini, Anass, et al. "Constrained Network Adversarial Attacks: Validity, Robustness, and Transferability." arXiv preprint arXiv:2505.01328 (2025).

[14] GitHub 1, 2025. Available: https://github.com/lava-aaa/iot_hw

[15] GitHub 2, 2025. Available: https://github.com/Dark-Avery/DDoS_classifier

[16] Sukhomlin, Vladimir A. "The Concept and Main Characteristics of the Master's Degree Program" Cybersecurity" of the Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University." International Journal of Open Information Technologies 11.7 (2023): 143-148.

[17] Namiot, Dmitry, and Vladimir Sukhomlin. "On cybersecurity of the Internet of Things systems." International Journal of Open Information Technologies 11.2 (2023): 85-97.

[18] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Artificial intelligence and cybersecurity." International Journal of Open Information Technologies 10.9 (2022): 135-147.