РусКрипто XXVIII НАУЧНО-ПРАКТИЧЕСКАЯ КОНФЕРЕНЦИЯ

Объяснимое обнаружение вторжений: от статистических к большим языковым моделям

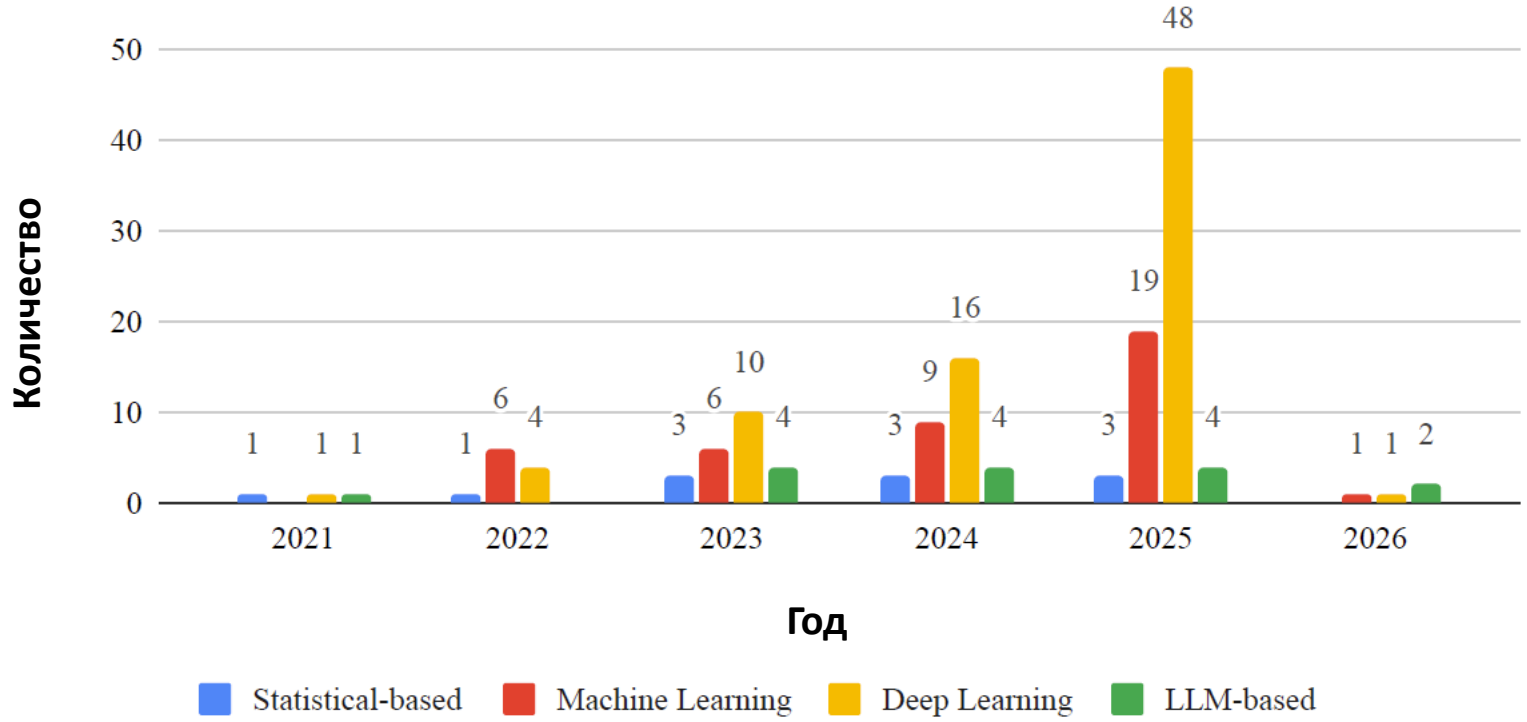
И.В. Котенко
ГНС, д.т.н., проф., Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН),
заслуженный деятель науки Российской Федерации

Содержание

Вопросы исследования

1. ***RQ1:*** *В каких областях X-IDS является необходимым инструментом, и какие уникальные методологические и операционные проблемы возникают?*
2. ***RQ2:*** *Как различные методы XAI сравниваются с точки зрения объяснимости, эффективности обнаружения и результативности в стандартных задачах и наборах данных IDS?*
3. ***RQ3:*** *Каковы аспекты внедрения и развертывания X-IDS, включая использование открытых фреймворков?*
4. ***RQ4:*** *Каковы ключевые проблемы внедрения XAI для IDS?*
5. ***RQ5:*** *Каковы основные уязвимости, риски и этические проблемы при развертывании X-IDS в реальных условиях?*

Распределение типов моделей в рассмотренной литературе по годам



Ранние методы обнаружения вторжений

Метод	Особенности XAI	Достоинства	Недостатки	Текущее использование
Системы, основанные на правилах	Правила, понятные человеку; Прямое сопоставление с образцом; Прозрачная логика принятия решений	Высокая интерпретируемость; Быстрое получение результатов; Не требуются обучающие данные.	Трудоемкое создание правил; Плохая обобщающая способность на неизвестные атаки; Сложности в обработке сложных шаблонов; Высокий FPR	Все еще широко используются при обнаружении на основе сигнатур
Статисти-ческие методы	Базовые профили поведения; решения, основанные на пороговых значениях; анализ распределений	Прозрачные границы принятия решений; Низкие вычислительные затраты; Эффективны для известных моделей атак	Проблемы с зашифрованным трафиком; Высокое количество ложных срабатываний из-за аномалий; Требуется ручная настройка порогового значения; Ограничено одномерным анализом	Интернет вещей /Автомобильные сети; системы реального времени; облегченные COB
Логисти-ческая / Линейная регрессия	Коэффициенты признаков; линейные зависимости; оценка вклада	Интерпретируемая важность признаков; Быстрое обучение и вывод результатов; Вероятностные выходные данные	Предполагает линейные зависимости; Низкая производительность в сложных сетях.	Интернет вещей /Автомобильные сети; системы реального времени; облегченные COB
Анализ главных компонен-тов (PCA)	Линейные комбинации признаков; объясненная дисперсия; факторные нагрузки компонентов	Снижает вычислительную сложность; Сохраняет наиболее значимую дисперсию; Выявляет влиятельные признаки	Предполагает линейные зависимости; Потеря информации	Снижение размерности; этап предварительной обработки

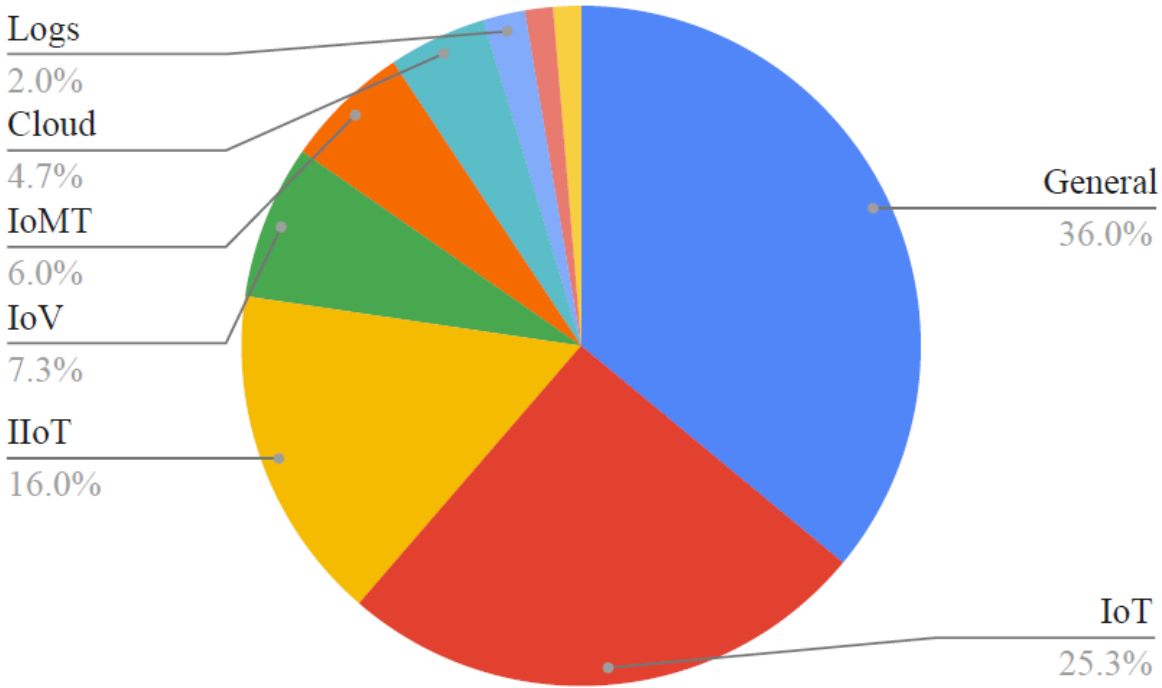
Классическое машинное обучение для обнаружения вторжений

Метод	Особенности XAI	Достоинства	Недостатки	Текущее использование
DT	Внутренняя объяснимость; Визуальные пути принятия решений; Важность признаков; Правила «если-то»	Наиболее широко применяется в исследованиях СОВ; Естественные, понятные человеку правила; Четкая ранжировка важности признаков; Обрабатывает нелинейные зависимости; Минимальные вычислительные затраты	Переобучение на сложных наборах данных; Ограниченная производительность на многомерных данных; Трудности с выявлением сложных взаимодействий признаков.	Широко используется для анализа сетевого трафика, обнаружения вторжений в IoT и сценариев граничных вычислений
SVM	Весовые коэффициенты признаков; Визуализация границ принятия решений; Локально интерпретируемый	Высокая точность бинарной классификации; Устойчивость к масштабированию признаков; Эффективное использование памяти; Быстрое время вывода результатов	Ограничение линейными границами принятия решений; Низкая объяснимость для нелинейных ядер; Трудности с задачами многоклассовой классификации	Сетевой трафик БПЛА/дронов; Анализ зашифрованного трафика Wi-Fi
KNN	Объяснения на основе отдельных экземпляров; Анализ сходства соседей	Интуитивно понятный и простой в использовании; Естественно предоставляет локальные пояснения; Обрабатывает нелинейные закономерности	Высокая вычислительная стоимость на этапе вывода; Чувствительность к масштабированию признаков; Низкая производительность на многомерных данных; Требует больших объемов памяти для больших наборов данных; Требует оптимального выбора k	Ограниченное применение в современных СОВ, в основном для базовых сравнений.
Fuzzy Logic	Вывод на основе нечетких правил	Внутренняя интерпретируемость; Обработка неопределенности; Адаптивное обучение правилам	Сложная архитектура модели; Требует экспертных знаний в предметной области	Используется в крупномасштабных, высокоскоростных сетевых средах

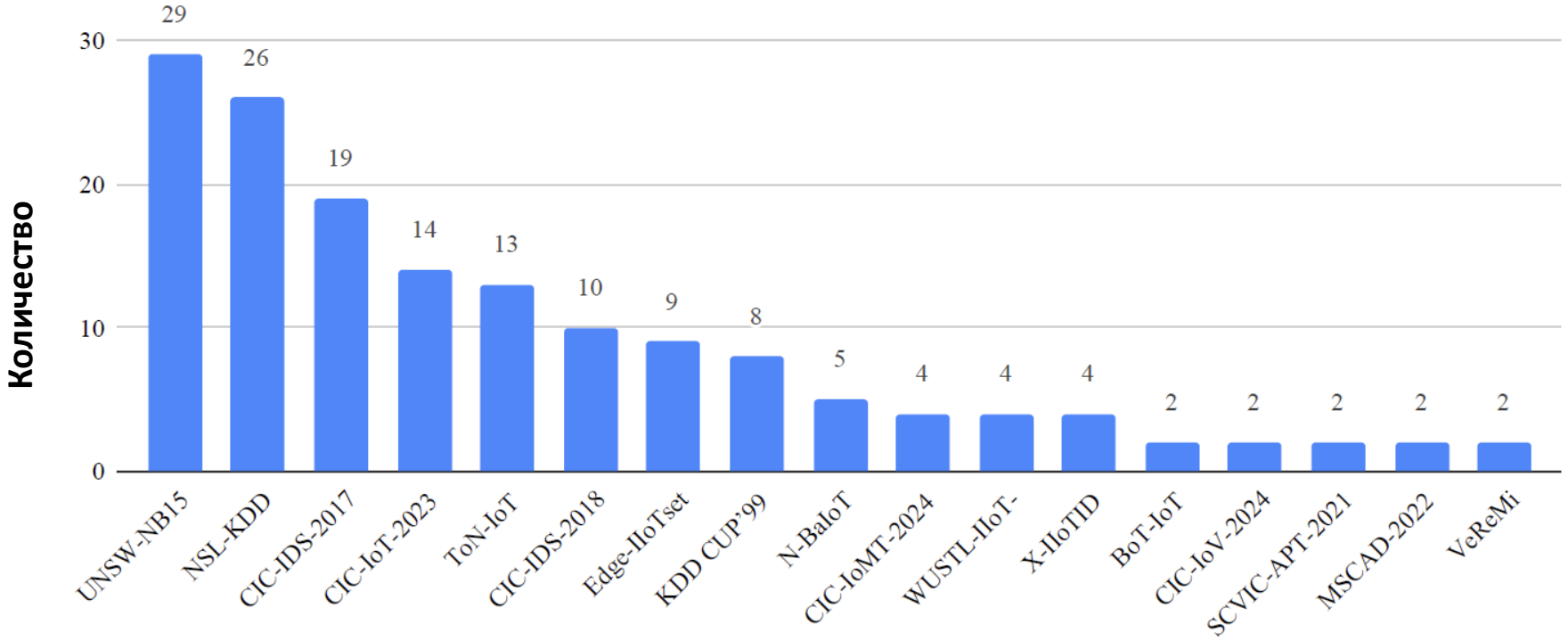
SHAP и LIME в системах обнаружения вторжений

Аспект	SHAP	LIME
Теоретическая основа	Теория игр (значения Шепли)	Локальная линейная аппроксимация
Область объяснения	Глобальная + локальная интерпретируемость	Локальная интерпретируемость (на уровне экземпляра)
Основное применение в СОВ	Ранжирование важности признаков, отладка набора данных, проверка модели	Индивидуальная интерпретация прогнозов, обоснование решений
Масштабируемость	Требует больших вычислительных ресурсов; лучше подходит для отбора признаков	Легковесный подходит для обнаружения вторжений в режиме реального времени
Ключевое преимущество для СОВ	Выявляет системные искажения, обеспечивает обнаружение атак без предварительного обучения (unsupervised) с помощью автокодировщиков	Объясняет конкретные случаи атак аналитикам по безопасности
Лучше всего подходит для	Отладка модели, инженерия признаков, обнаружение новых атак	Оповещения в режиме реального времени, поддержка принятия решений оператором

Распределение публикаций по предметным областям



Распределение публикаций по используемым открытым наборам данных



Оценка производительности с учетом вычислительных ресурсов (1)

Решение	Время	Накладные расходы, связанные с XAI
lFex-IDS [162]	10-15 мс для обнаружения на уровне пакетов и 0,70-0,90 % криптографических накладных расходов на раунд	LIME в 40-60 раз быстрее, чем SHAP
DeepCRNN [114]	Производительность обнаружения оценивается с использованием 5-секундного окна, достигая accuracy 99,50 % с минимальной задержкой, что соответствует требованиям к эффективности блокчейна	Вычисление значений Шапли для каждой характеристики добавляет существенные вычислительные затраты; поэтому объяснения на основе SHAP генерировались только после предсказания модели
E2I3DS [58]	Время отклика тестирования составляет 0,1517 мкс на предсказание, что указывает на высокую производительность обнаружения	–
XDIoT [88]	В качестве показателя эффективности использовалось время выполнения прогнозирования/обнаружения, составляющее 2-4 секунды для прогнозирования потока, что является умеренным показателем	Хотя это не было количественно оценено, в исследовании также отмечалось, что значения Шапли требуют повторных оценок модели и являются вычислительно затратными .
SIoV-IDS [109]	Задержка обработки данных для каждого образца оценивается в 0,128 мс/наблюдение на наборе данных Edge-IIoT и 0,086 мс/наблюдение на наборе данных IoV (попадают в типичные диапазоны генерации пакетов 1-10 мс для CAN и IoT, поддерживая развертывание в реальном времени)	SHAP используется исключительно как инструмент постфактумного анализа, а не в конвейере вывода; поэтому он не влияет на скорость вывода в реальном времени
EADL-IDS [179]	Эксперименты показывают задержку бинарного обнаружения 0,324 мс на туманном уровне (fog layer)	Модуль объяснения на основе LLM добавил задержку в 1,348 мс, увеличив задержку многоклассового обнаружения до 1,672 мс

Краткий обзор основных фреймворков XAI

Метод	Математические основы	Реализация	Совместимость моделей	Поддержка моделей	Выходные данные	Вычислительная сложность	Установка Python
Важность признаков по Scikit-Learn	Снижение неоднородности данных (коэффициент Джини, энтропия, дисперсия); среднее снижение примеси, среднее снижение accuracy	Внутренний атрибут	Зависимый от модели	Модели на основе деревьев	Ранжирование признаков	O(m × n)	pip install scikit-learn
SHAP	Теория игр (значения Шапли)	DeepExplainer	Независимый от модели и зависимый от модели	PyTorch, TensorFlow	Summary, Force	KernelSHAP: O(m × n) LinearSHAP: O(m)	pip install shap
LIME	Локальные суррогатные модели	Таблицы	Независимый от модели	Независимый от модели	Локальные объяснения	O(m × n)	pip install lime
Captum	Множественные (интегрированные градиенты, DeepLIFT)	IG, DeepLIFT	Только модели PyTorch	Только PyTorch	Saliency, IG	Варьируется в зависимости от метода; в целом эффективен	pip install captum
ELI5	Важность перестановок, обертка LIME	Перестановки	scikit-learn, XGBoost, LightGBM	sklearn, XGB, LigthGBM	Весы признаков	От низких до средних	pip install eli5

Заключение (1)

Растущая изощренность кибератак требует **надежных и интерпретируемых систем обнаружения вторжений**. В обзоре было проанализировано **150 рецензируемых статей**, прослеживающих эволюцию методов объяснимого искусственного интеллекта от традиционных статистических методов и классического машинного обучения через глубокое обучение до новых методов обучения с расширенными возможностями. **Ключевые выводы**:

(1) Хотя **модели глубокого обучения** достигают высокой точности обнаружения (>99,00%), **классические методы машинного обучения** часто обеспечивают конкурентоспособные результаты обнаружения, лучшую объяснимость при минимальных вычислительных затратах.

(2) **SHAP и LIME** стали преобладающими методами XAI, но вносят существенную задержку.

(3) **Статистические методы** остаются актуальными для выбора признаков и легковесных реализаций.

(4) **LLM** демонстрируют перспективность для анализа логов и генерации объяснений, но в настоящее время **показывают низкую эффективность при обработке необработанного сетевого трафика без тонкой настройки, специфичной для предметной области**.

Заключение (2)

- Выявлены следующие **критические проблемы**:
  - отсутствие стандартизированных метрик оценки объяснимости,
  - недостаточный мониторинг реализации объяснений после развертывания,
  - ограниченные руководства операторам по внедрению в производство,
  - риски для конфиденциальности, связанные с объяснением, вредоносная эксплуатация интерфейсов XAI и уязвимости LLM остаются недостаточно изученными.
- **Рекомендуется**:
  - создание **стандартизированных систем оценки объяснимости**,
  - внедрение **гибридных архитектур**, сочетающих обнаружение и объяснение,
  - **приоритетное использование методов XAI, обеспечивающих конфиденциальность**, для развертывания в критической инфраструктуре.

Спасибо за внимание!

РусКрипто XXVIII научно-практическая конференция

ГНС, д.т.н., проф. **Котенко Игорь Витальевич,**
ivkote@comsec.spb.ru, http://comsec.spb.ru/kotenko
Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН)

Благодарности

- работа выполнена совместно с Левшун Д.А., Левшун Д.С. И Дун Х. (СПб ФИЦ РАН)
- Данное исследование было поддержано Российским научным фондом и Санкт-Петербургским научным фондом (грант № 25-11-20028, https://rscf.ru/project/25-11-20028/) в СПб ФИЦ РАН.