УДК 004.056

ИНТЕРПРЕТИРУЕМОСТЬ (EXPLANABLE AI) РЕШИНИЙ ИИ ДЛЯ ОБНАРУЖЕНИЯ УГРОЗ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Кудряшов А. Е. (ИТМО)

Научный руководитель - старший научный сотрудник, доцент Гирик А. В. (ИТМО)

Введение. Использование моделей машинного обучения создаёт проблему низкой прозрачности решений[1]. Это затрудняет анализ причин срабатывания системы обнаружения. Невозможность проанализировать факторы оказавшие влияние на решение снижает доверие со стороны специалистов по информационной безопасности, что осложняет как на процесс внедрения на предприятия, так и реагирования на инциденты. Актуальной задачей становится интеграция методов интерпретируемого искусственного интеллекта, позволяющих объяснять выводы моделей и связывать их с событиями безопасности[4].

Основная часть. Архитектура системы интерпретируемого обнаружения вторжений представляет собой программную структуру, состоящую из 4 модулей. В основе системы - сбор событий безопасности «Windows». Собранные данные поступают в модуль преобразования событий в сигнатурные представления на основе метода «log2sig»[3]. Модель формирует векторные описания последовательностей. Далее полученные сигнатуры передаются в модель машинного обучения, для выявления потенциальных вторжений[2]. В случае превышения заданного порога классификации данные передаются в модуль интерпретируемого анализа решений, который объясняет вывод модели. Строиться объяснение методом «SHAP» и отчет, позволяющий количественно оценить вклад каждого признака сигнатуры «log2sig» в итоговый результат классификации. Далее формируется отчёт в формате HTML об инциденте, а так же о ключевых событиях и признаках, повлиявших на принятое решение. Модель обучалась на датасете «CIC-IDS-2017» с платформы «Kaggle», содержащий более двух миллионов строк.

Вывод. В ходе исследования была продемонстрирована возможность эффективной классификации инцидентов с использованием модели машинного обучения. Применение метода «SHAP» позволило обеспечить интерпретируемость решений модели. Показано, что представление последовательностей событий в виде сигнатур с использованием метода «log2sig» позволяет применять алгоритмы машинного обучения к последовательностям фиксированной размерности. Реализована экспериментальная система обнаружения вторжений, ориентированная на анализ журналов безопасности операционной системы «Windows» на конечных точках.

Список использованных источников

1. Molnar, C. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. — 2022. — [Электронный ресурс]. — URL: https://christophm.github.io/interpretable-ml-book/ (дата обращения: 28.09.2025).
2. Veeramachaneni, K., Arnaldo, I., Korrapati, V., et al. AI²: Training a Big Data Machine to Defend // 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity). — 2016. — P. 49–54.

3. Kong K., Liu D., Jin X., Li Z., Geng G. Log2Sig: frequency-aware insider threat detection via multivariate behavioral signal decomposition // arXiv preprint arXiv:2508.05696. – 2025. – Submitted to IEEE TrustCom..
4. Ribeiro, M.T., Singh, S., Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). — 2016. — P. 1135–1144.