

УДК 004.056:004.85

Арванова Саният Мухамедовна, старший преподаватель
ФГБОУ ВО «Кабардино-Балкарский государственный университет
им. Х. М. Бербекова»
Arvanova Saniyat Mukhamedovna

Маршенкулов Абдуллах Аликович, студент
ФГБОУ ВО «Кабардино-Балкарский государственный университет
им. Х. М. Бербекова»
Marshenkulov Abdullah Alikovich

Татроков Тамирлан Асланович, студент
ФГБОУ ВО «Кабардино-Балкарский государственный университет
им. Х. М. Бербекова»
Tatrokov Tamirlan Aslanovich

Битоков Аслан Заурович, студент
ФГБОУ ВО «Кабардино-Балкарский государственный университет
им. Х. М. Бербекова»
Bitokov Aslan Zaurovich

Научный руководитель:
Крымшокалова Джульетта Аbugалиевна
кандидат физико-математических наук, доцент
ФГБОУ ВО «Кабардино-Балкарский государственный университет
им. Х. М. Бербекова»
Krymshokalova Julietta Abugalievna

МУЛЬТИМОДАЛЬНАЯ СИСТЕМА ОБНАРУЖЕНИЯ ВНУТРЕННИХ УГРОЗ НА ОСНОВЕ МАШИННОГО ОБУЧЕНИЯ A MULTIMODAL MACHINE LEARNING SYSTEM FOR INSIDER THREAT DETECTION

Аннотация. Внутренние угрозы трудно обнаруживать по отдельным событиям информационной системы, поскольку легитимные полномочия позволяют сотруднику выполнять потенциально опасные действия без нарушения правил доступа. Проблема исследования заключается в согласовании разнородных свидетельств при редкости подтверждённых инцидентов, изменении поведения пользователей и неполноте наблюдений. Цель работы состоит в разработке архитектуры мультимодального обнаружения, которая объединяет временные характеристики активности, семантику документов и структуру взаимодействий. Основным аналитический результат выражает зависимость точности тревог от распространённости угроз и частоты ложных срабатываний. В условном сценарии при распространённости угроз 0,1%, полноте 80% и частоте ложных срабатываний 1% точность тревог составляет 7,41%. Снижение последнего показателя до 0,1% повышает расчётную точность до 44,47%. Практическая значимость заключается в определении измеримых требований к порогу обнаружения и воспроизводимому испытанию системы.



Abstract. Insider threats are difficult to identify from isolated information system events because legitimate privileges allow employees to perform potentially harmful operations without violating access rules. The research problem concerns combining heterogeneous evidence under rare confirmed incidents, changing user behaviour, and incomplete observations. The objective is to develop a multimodal detection architecture that integrates temporal activity characteristics, document semantics, and interaction structure. The main analytical result relates alert precision to threat prevalence and the false positive rate. In an illustrative scenario with prevalence 0.1%, recall 80%, and false positive rate 1%, alert precision equals 7.41%. Reducing the false positive rate to 0.1% increases calculated precision to 44.47%. The contribution is a set of measurable threshold requirements and reproducible evaluation conditions.

Ключевые слова: Внутренняя угроза; мультимодальное обучение; обнаружение аномалий; поведенческая аналитика; объединение данных; калибровка; ложные срабатывания.

Keywords: Insider threat; multimodal learning; anomaly detection; behavioural analytics; data fusion; calibration; false positives.

Введение. Внутренняя угроза возникает, когда субъект с разрешённым доступом использует информационные ресурсы вопреки интересам организации. Нарушение может выражаться в подготовке утечки, неправомерном копировании либо изменении данных. При этом нетипичная активность сама по себе не доказывает злонамеренности: массовая выгрузка может быть частью служебной задачи, а вход ночью – следствием сменного графика. В работе Yuan и Wu [1] связывает развитие глубокого обучения для обнаружения внутренних угроз с задачами объединения разнородных данных и ограниченной доступностью размеченных примеров. Al-Mhiqani и соавторы [2] систематизируют методы машинного обучения и ограничения наборов данных. Систематизация поведенческих источников у Ко и соавторов [3] показывает, что цифровой след пользователя не сводится к одному журналу событий. Однако увеличение числа источников не гарантирует повышения качества: зависимые признаки могут усиливать одну и ту же ошибку, а отсутствующие данные – смещать итоговую оценку. В настоящей работе мультимодальность понимается как совместная обработка представлений различной природы: последовательностей действий, семантических векторов текста и графовых отношений. А у Baltrušaitis и соавторов [4] задаёт общую таксономию объединения модальностей. Работа Ramachandram и Taylor [5] позволяет различать объединение исходных признаков и объединение решений отдельных моделей.

Цель исследования – предложить архитектуру обнаружения внутренних угроз, в которой объединение свидетельств сопровождается контролем неполноты данных и измерением эксплуатационной нагрузки. Объектом служат наблюдаемые действия пользователей корпоративной информационной системы. Предмет исследования – формирование оценки риска из нескольких модальностей и выбор порога передачи события аналитику. Проверяемая гипотеза состоит в том, что взаимодополняющие признаки могут повышать полноту обнаружения при фиксированном бюджете ложных тревог.

Постановка задачи. Единицей наблюдения выбирается пара «пользователь – временное окно». Продолжительность окна является параметром испытания: короткий интервал уменьшает задержку, но теряет контекст, длинный объединяет связанные операции, однако может скрывать момент начала инцидента. Система оценивает вероятность подтверждения угрозы по доступной информации к концу окна. Истинная метка определяется независимым расследованием либо сценарием исследовательского набора, а не собственным порогом модели. Статья Tuor и соавторов [6] обосновывает интерес к обучению на структурированных потоках событий без полной разметки угроз. Le и соавторы [7]



непосредственно обращаются к гранулярности данных. Эти подходы задают исходную точку для построения контрольных моделей. В предлагаемой постановке необходимо дополнительно учитывать смену должности, отпуск, миграцию сервисов и другие легитимные причины изменения цифрового следа. Система должна выдавать оценку риска вместе с указанием доступных источников и интервала наблюдения. Если текстовый канал недоступен, отсутствие текста не трактуется как отсутствие угрозы. Если изменился сбор журналов, снижение активности не считается подтверждением безопасного поведения. Такое разделение поведенческих признаков и признаков наблюдаемости необходимо для последующего анализа ошибок и воспроизводимого сравнения конфигураций.

Методика исследования

Методика сочетает проектирование архитектуры, аналитическое моделирование и подготовку протокола экспериментальной проверки. Различие между статистической аномалией и нарушением безопасности принципиально: обзор Chandola и соавторов [8] задаёт общую проблематику обнаружения отклонений. Поэтому оценка необычности используется как один из сигналов, который должен проверяться в контексте конкретного сценария угрозы. Входной слой нормализует идентификаторы пользователей, время событий и обозначения ресурсов. Повторные записи удаляются по устойчивому идентификатору, а задержка поступления сохраняется как характеристика качества данных. Нормализация числовых признаков обучается только на обучающем интервале. Для признаков, зависящих от истории, используются исключительно события, зарегистрированные до момента формирования решения. Словари, текстовые преобразования и графовые статистики подчиняются тому же ограничению, иначе сведения из будущего завышают оценку качества.

Поведенческий канал описывает частоты входов, смену устройств, объём обращений к файлам и отклонение времени активности от индивидуального профиля. Для него предлагается использовать Isolation Forest [9] как сравнительно простой источник оценки необычности. **Семантический канал** преобразует разрешённые к анализу тексты в векторные представления; модель внимания [10] служит методологической основой кодировщика. **Графовый канал** описывает связи «пользователь – ресурс» и изменения множества контрагентов. Например, появление доступа к ранее не использовавшейся группе документов оценивается с учётом роли сотрудника, а не только редкости самого события.

Признаки графа вычисляются на историческом срезе; будущие связи исключаются. Для нового сотрудника профиль строится с опорой на сопоставимую рабочую группу, при этом неопределённость явно передаётся в итоговый классификатор. Изменения распределения данных отслеживаются отдельно от тревог безопасности. В предлагаемой системе обновление выполняется после проверки качества на отложенном интервале, поскольку безусловное дообучение может закрепить длительную вредоносную активность как норму.

Каждый канал формирует оценку s_m и индикатор доступности a_m . Для позднего объединения предлагается логистическая модель $p = \sigma(b + \sum_m a_m w_m s_m + \sum_m v_m a_m)$, где σ обозначает логистическую функцию, b – свободный член, w_m – коэффициент канала, v_m – вклад маски доступности. Индекс m принимает три значения. Коэффициенты оцениваются на размеченных данных; ручное назначение весов допустимо только как отдельно обозначенная эвристическая базовая модель. Если все каналы недоступны, результатом становится сообщение о недостаточности наблюдений, а не автоматическое заключение об отсутствии угрозы. Окна, относящиеся к одному инциденту, не должны пересекать границы этих интервалов. Полная система сравнивается с каждым каналом отдельно и с объединением без семантических либо графовых признаков при одинаковом временном разбиении. Для каждой конфигурации измеряются площадь под кривой precision-recall, полнота при заданном бюджете тревог, время обнаружения и число ложных тревог на единицу времени.



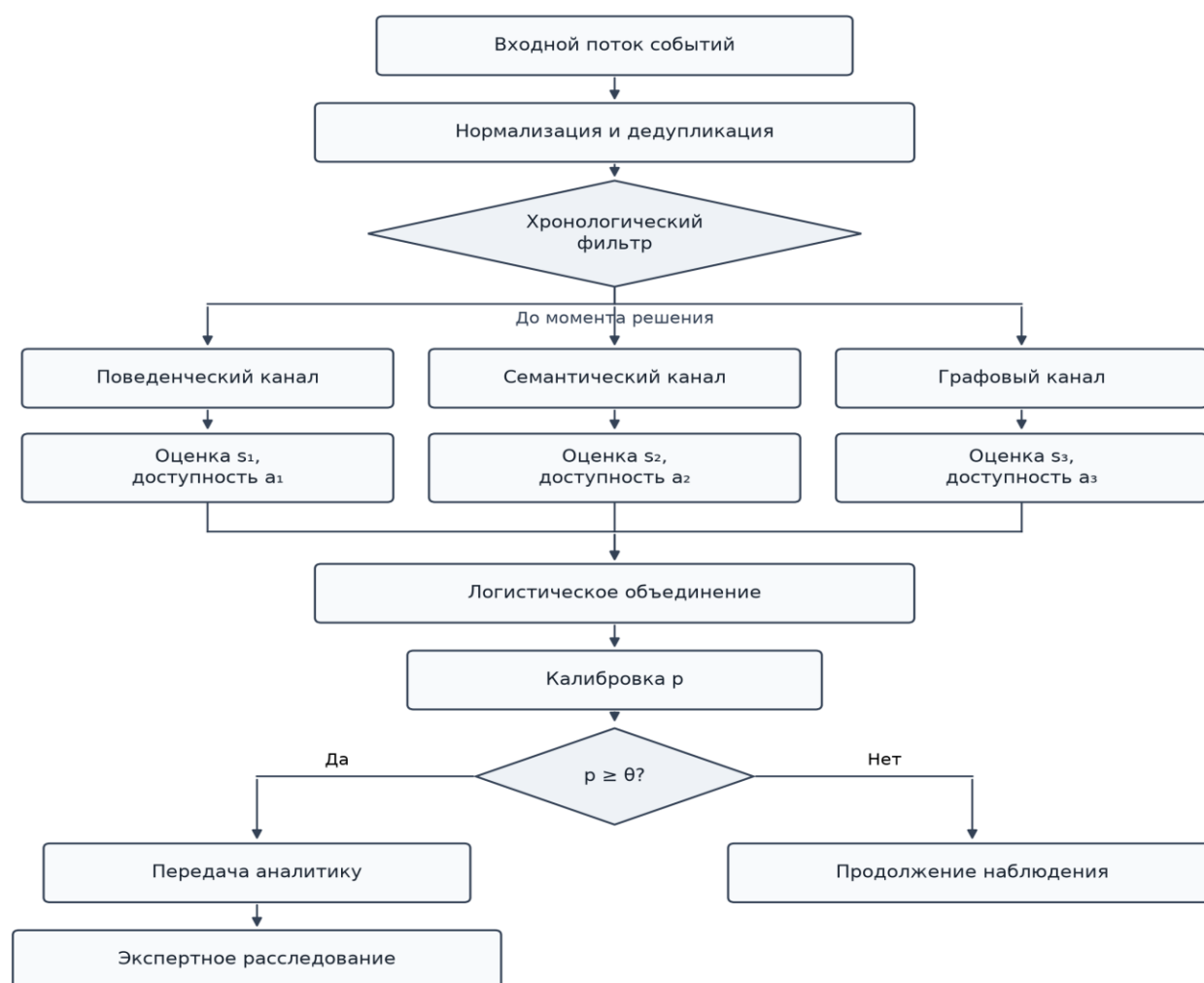


Рисунок 1. Алгоритм мультимодального обнаружения внутренних угроз

Результатом проектирования является архитектура с тремя самостоятельными каналами, маской доступности и единым калибруемым решением. Её существенная особенность заключается в явном учёте неполноты наблюдений: одно и то же значение риска сопровождается сведениями о том, на каких источниках оно основано. Формализованный протокол отделяет подбор порога от окончательной проверки и связывает сравнение моделей с ресурсами экспертного расследования.

Таблица 1

Сравнение подходов к обнаружению внутренних угроз

Метод	Источники данных	Учёт неполноты	Калибровка	Операционные метрики
Пороговые правила	Журналы аудита	Не предусмотрен	Не требуется	Число тревог
Одномодальная аномалия [6]	Последовательность событий	Косвенный	Опционально	Площадь под ROC
Графовое обнаружение	Граф взаимодействий	Косвенный	Опционально	Централизованность узлов
Предлагаемая система	Поведение + текст + граф	Явная маска $a_{m \times m}$	Обязательная	Точность при бюджете В



Аналитический результат характеризует влияние редкости угроз на полезность тревог. Пусть π – доля опасных наблюдений, r – полнота обнаружения, f – частота ложноположительных решений среди безопасных наблюдений. Тогда прогностическая ценность положительного решения, то есть точность тревог, равна $P = r\pi / (r\pi + f(1 - \pi))$. Формула следует из отношения ожидаемого числа истинных тревог к ожидаемому общему числу тревог. Она не предполагает независимости модальностей и применима к итоговому решению любой модели при фиксированном пороге.

Для иллюстрации зададим условный поток $N = 100\,000$ наблюдений, $\pi = 0,001$ и $r = 0,8$. Ожидается 100 опасных наблюдений, из которых система выявит 80, и 99 900 безопасных. При $f = 0,01$ ожидаемое число ложных тревог составляет 999, общее число тревог – 1079, а их точность – $80/1079 = 7,41\%$. Таким образом, высокая полнота сама по себе не обеспечивает приемлемой нагрузки: основная часть направленных на проверку событий окажется безопасной.

При $f = 0,001$ и неизменной полноте ожидаемое число ложных тревог равно 99,9, общее число – 179,9, а точность – 44,47%. Дробное количество означает математическое ожидание для повторяющихся потоков, а не число фактически зарегистрированных случаев. Полученное изменение показывает чувствительность эксплуатационного результата к f . Оно не означает, что предложенная мультимодальная система уже достигла такого снижения: для этого необходим эксперимент с одинаковой полнотой и независимой тестовой выборкой.

Из той же зависимости выводится ограничение для желаемой точности q :

$$f \leq r\pi(1 - q) / (q(1 - \pi)).$$

При $q = 0,5$, $\pi = 0,001$ и $r = 0,8$ допустимая частота ложных срабатываний не превышает приблизительно 0,0008008, или 0,08%. Если организация способна проверять не более B тревог за интервал, дополнительно требуется $N(r\pi + f(1 - \pi)) \leq B$. Например, при условном бюджете $B = 200$ первый сценарий превышает возможности проверки, а второй удовлетворяет ограничению по математическому ожиданию. Фактические всплески нагрузки необходимо исследовать отдельно.

Полезность объединения модальностей должна оцениваться в этих операционных координатах. Если семантический канал повышает площадь под кривой, но увеличивает число повторных сообщений об одном инциденте, выигрыш может не сопровождаться снижением трудозатрат. Поэтому наряду с оконными метриками необходимо измерять число обнаруженных инцидентов и задержку первой полезной тревоги. Правило объединения повторных сообщений фиксируется до сравнения, иначе подавление дублей становится скрытым способом улучшения показателей. Ограничения работы связаны с отсутствием исходного корпуса событий и подтвержденной разметки.

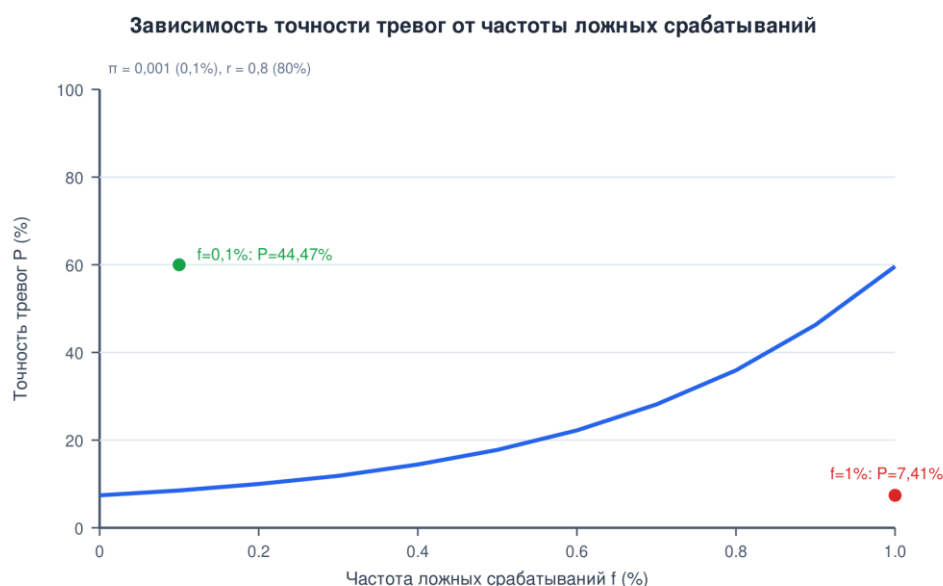


Рисунок 2 – Зависимость точности тревог от частоты ложных срабатываний



Репрезентативность синтетических наборов также требует самостоятельной оценки. Семантический анализ допустим только в пределах установленного доступа к данным; минимизация содержания и разграничение полномочий аналитиков должны входить в проект системы.

Заключение. Определён протокол проверки, исключающий использование будущих событий при обучении и настройке порога. Аналитически показано, что при редких угрозах частота ложных срабатываний определяет как точность тревог, так и ожидаемую нагрузку на расследование. Практический результат состоит в переводе требований к системе в измеримые ограничения на эти показатели.

Список литературы:

1. Yuan S., Wu X. Deep Learning for Insider Threat Detection: Review, Challenges and Opportunities [Электронный ресурс]. – 2020. – arXiv:2005.12433. – URL: <https://arxiv.org/abs/2005.12433> (дата обращения: 29.09.2026).
2. A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations / M. N. Al-Mhiqani, R. Ahmad, Z. Zainal Abidin et al. // Applied Sciences. – 2020. – Vol. 10, no. 15. – Art. 5208. – DOI: 10.3390/app10155208.
3. Insider Threat Detection and Its Future Directions / L. L. Ko, D. M. Divakaran, Y. S. Liao, V. L. L. Thing // International Journal of Security and Networks. – 2017. – Vol. 12, no. 3. – P. 168–187. – DOI: 10.1504/IJSN.2017.084391.
4. Baltrušaitis T., Ahuja C., Morency L.-P. Multimodal Machine Learning: A Survey and Taxonomy // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 41, no. 2. – P. 423–443. – DOI: 10.1109/TPAMI.2018.2798607.
5. Ramachandram D., Taylor G. W. Deep Multimodal Learning: A Survey on Recent Advances and Trends // IEEE Signal Processing Magazine. – 2017. – Vol. 34, no. 6. – P. 96–108. – DOI: 10.1109/MSP.2017.2738401.
6. Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams / A. R. Tuor, S. P. Kaplan, B. J. Hutchinson et al. // The AAAI Workshop on Artificial Intelligence for Cyber Security. – 2017. – WS-17-04. – P. 224–231.
7. Le D. C., Zincir-Heywood N., Heywood M. I. Analyzing Data Granularity Levels for Insider Threat Detection Using Machine Learning // IEEE Transactions on Network and Service Management. – 2020. – Vol. 17, no. 1. – P. 30–44. – DOI: 10.1109/TNSM.2020.2967721.
8. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Computing Surveys. – 2009. – Vol. 41, no. 3. – Art. 15. – P. 1–58. – DOI: 10.1145/1541880.1541882.
9. Liu F. T., Ting K. M., Zhou Z.-H. Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining. – 2008. – P. 413–422. – DOI: 10.1109/ICDM.2008.17.
10. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar et al. // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008

