DOI 10.37882/2223-2966.2025.10.44

АНАЛИЗ ОШИБОК МАШИННОГО ОБУЧЕНИЯ ПРИ ОБНАРУЖЕНИИ SSRF

ERROR ANALYSIS IN MACHINE LEARNING FOR SSRF DETECTION

Gabriela Guadalupe Chavez Quiroz
N. Voinov

Summary. This study analyzes systematic errors in machine learning models for SSRF vulnerability detection. Key findings reveal: (1) confusion between basic and advanced SSRF variants (38 % of errors) associated with HTTP 403 responses and 2800–3200-byte payloads; (2) false positives in legitimate traffic (42 %) triggered by ≥2 redirects or PUT/POST methods; and (3) synthetic dataset limitations (20 %) when processing internal API requests to non-standard ports (8080/8443). The stacking ensemble model achieved optimal performance (96.3 % accuracy), reducing false positives to 1.2 %. SHAP analysis informed three key improvements: multi-level traffic verification, prioritized cloud metadata features (SHAP >0.15), and a new response-to-request size ratio feature. The research highlights the necessity of hybrid datasets combining synthetic and real-world data, particularly for edge cases in classes 4, 5, and 11. Proposed solutions address feature ambiguity while preserving the advantages of automated detection systems.

Keywords: SSRF vulnerabilities (Server-Side Request Forgery), machine learning, vulnerability detection, error analysis, synthetic dataset.

Чавес Кирос Габриэла Гуадалупе
Аспирант, Санкт-Петербургский политехнический университет Петра Великого
chaveskiros.g@edu.spbstu.ru
Воинов Никита Владимирович
Кандидат технических наук, доцент, Санкт-Петербургский политехнический университет Петра Великого
voinov@ics2.ecd.spbstu.ru

Аннотация. Данное исследование анализирует систематические ошибки моделей машинного обучения при обнаружении SSRF-уязвимостей. Основные проблемы включают: (1) путаницу между базовыми и продвинутыми вариантами SSRF (38 % ошибок), связанную с HTTP-кодом 403 и размером ответа 2800–3200 байт; (2) ложные срабатывания в легитимном трафике (42 %) при ≥2 редиректах или HTTP PUT/POST методах; (3) ограничения синтетических данных (20%) для запросов к внутренним API с портами 8080/8443. Ансамблевый метод Stacking показал наилучшие результаты (96.3 % точности), сократив ложные срабатывания до 1.2 %. На основе SHAP-анализа предложены улучшения: многоуровневая верификация трафика, приоритезация метаданных облачных сервисов (SHAP >0.15) и новый признак — отношение размера ответа к запросу. Исследование подчёркивает необходимость комбинирования синтетических и реальных данных, особенно для edge-случаев классов 4, 5 и 11.

Ключевые слова: SSRF-уязвимости, машинное обучение, обнаружение уязвимостей, анализ ошибок, синтетический набор данных.

Введение

Server-Side Request Forgery (SSRF) представляет собой уязвимость, возникающую при некорректной валидации серверных запросов, что позволяет осуществлять несанкционированный доступ к внутренним и внешним ресурсам [1]. Данная уязвимость включена в список OWASP Top 10–2021 [2] и представляет существенную угрозу для облачных инфраструктур. Одним из наиболее масштабных инцидентов, связанных с эксплуатацией SSRF, стал взлом Capital One в 2019 году [3], в результате которого было скомпрометировано свыше 100 миллионов записей пользователей.

В научной литературе предлагаются различные подходы к обнаружению уязвимостей SSRF. Исследование Wessels et al. [4] посвящено методам статического анализа потоков данных в PHP-приложениях, в то время как Wang et al. [5] разработали инструмент SSRFuzz, основанный на фаззинге и позволяющий выявлять ранее неизвестные уязвимости. Параллельно, работы Al-talak [6] и Kulkarni [7] сосредоточены на детектировании активных атак с использованием методов глубокого обучения и специализированных архитектур защиты.

В предыдущем исследовании [8] проведена комплексная оценка алгоритмов машинного обучения, включая Random Forest, XGBoost, LightGBM и ансамблевые методы. Наилучшие результаты показал Stacking Ensemble с точностью 96.3 %, F1-score 0.963 и ROC AUC 0.999. Тем не менее, даже в наиболее эффективных моделях сохраняются систематические ошибки классификации.

Анализ выявил два основных типа ошибок. Во-первых, наблюдается путаница между базовыми и продвинутыми вариантами SSRF, особенно в случаях совпадения характеристик запросов, таких как код состояния HTTP 403 или аналогичные диапазоны размера ответа. Во-вторых, отмечаются случаи ложной классификации легитимного трафика, содержащего атипичные паттерны, например, множественные редиректы или редко используемые HTTP-методы.

Настоящее исследование направлено на детальный анализ остаточных ошибок классификации. Особое внимание уделяется как алгоритмическим ограничениям применяемых методов, так и потенциальным смещениям, обусловленным использованием сбалансированного синтетического набора данных. Результаты исследования легли в основу предложенных усовершенствований, позволяющих сохранить преимущества автоматизированного подхода при значительном снижении частоты ошибочных классификаций.

Методология

Исследование проводилось на трех предварительно обученных моделях машинного обучения: Random Forest, XGBoost и Stacking Ensemble. В работе использовался синтетический набор данных из предыдущего исследования [8], содержащий 15 категорий запросов — от базовых вариантов SSRF до сложных техник эксплуатации в облачных средах.

Методология исследования включала три последовательных этапа. На первом этапе выполнялась количественная оценка ошибок классификации с разбивкой по типам и классам. Второй этап был посвящен анализу интерпретируемости моделей с использованием SHAP-значений и оценки важности признаков. На заключительном этапе проводилась статистическая валидация выявленных закономерностей с применением корреляционного анализа и проверки статистических гипотез.

Для обеспечения надежности результатов на всех этапах исследования применялась стратегия стратифицированной кросс-валидации. Особое внимание уделялось классам с наиболее высокими показателями ошибок классификации. Такой подход позволил минимизировать потенциальные смещения и обеспечить воспроизводимость результатов.

Анализ ошибок

Модели продемонстрировали систематические паттерны ошибочной классификации, которые можно разделить на три основные группы.

Наибольшая доля ошибок (38 %) связана с трудностями различения базовых (Класс 5) и продвинутых (Класс 4) сценариев SSRF, особенно при обработке запросов с HTTP-кодом 403 и размером ответа 2800–3200 байт. Анализ SHAP-значений выявил систематическое завышение важности порта назначения (SHAP 0.07) при недостаточном учете более надежных признаков, таких как метаданные облачных заголовков (SHAP 0.19).

Ложные срабатывания при анализе легитимного трафика (42 % ошибок) преимущественно относились к Классу 0 и были связаны с атипичными характеристиками: цепочками из ≥2 редиректов (встречались в 82 % ошибочных случаев), методами HTTP PUT/POST (на 35 % чаще вызывали ошибки по сравнению с GET) и аномально коротким временем ответа (<200 мс). Использование Stacking Ensemble позволило значительно снизить количество таких ошибок с 2.1 % до 1.2 %.

Оставшиеся 20 % ошибок, затрагивающие преимущественно Классы 4, 5 и 11 (запросы к внутренним API), выявили ограничения синтетического набора данных. Наибольшие сложности возникли при обработке редких комбинаций параметров, таких как использование нестандартных портов (8080, 8443) с определенными HTTP-методами, где искусственная природа данных проявилась наиболее явно. В Таблице 1 представлено детальное распределение ошибок с указанием характерных признаков и затронутых классов, что обеспечивает наглядное представление выявленных закономерностей.

Таблица 1.

Распределение ошибок по категориям

Тип ошибки	% от общего числа ошибок	Ключевые характеристики	Наиболее затронутые классы
Путаница SSRF (базовый/ продвинутый)	38 %	— Коды HTTP 403 — Размер ответа 2800–3200 байт	Класс 4 против Класса 5
Ложные срабатывания (легитимный трафик)	42 %	— ≥2 перенаправления — Методы PUT/POST — Время отклика <200 мс	Класс 0
Ограничения синтетического датасета	20 %	— Редкие комбинации атрибутов — Нестандартные порты (8080, 8443)	Классы 4, 5, 11

Проведенный стратифицированный анализ демонстрирует, что возникающие ошибки носят не случайный, а систематический характер и связаны с тремя ключевыми факторами: (1) частичным совпадением характеристик между функционально схожими классами, (2) наличием статистически аномальных паттернов в легитимном трафике, и (3) ограничениями охвата при генерации синтетических данных. Четкое определение данных закономерностей открывает возможности для разработки адресных корректирующих мер, направленных на устранение ошибок каждой конкретной категории.

Рекомендации

Для решения проблемы ложных срабатываний в легитимном трафике (Класс 0) рекомендуется внедрить многоуровневую систему верификации. Особое внимание следует уделить анализу полных цепочек перена-

правлений и вычислению соотношения времени ответа к размеру запроса. Экспериментальные данные показывают, что пороговое значение этого параметра ниже 0.05 мс/байт обеспечивает точность идентификации нормального трафика на уровне 89 %. Дополнительно следует учитывать частоту использования методов PUT/POST и аномально короткие времена отклика.

Для улучшения классификации сложных SSRF-сценариев (Класс 4) предлагается модификация модели по двум направлениям. Во-первых, необходимо расширить набор анализируемых признаков, включив специфические паттерны перенаправлений, характерные для SSRF. Во-вторых, требуется увеличить значимость метаданных облачных сервисов в процессе принятия решений, установив минимальный порог SHAP-значений для этих признаков на уровне 0.15.

Отдельный комплекс мер направлен на снижение взаимной путаницы между классами. Наиболее эффективным решением представляется введение нового признака — отношения размера ответа к размеру запроса. Статистический анализ подтвердил значимые различия этого показателя между базовыми SSRF (1.8±0.4) и обращениями к внутренним API (3.2±0.7). Реализация данного подхода позволит улучшить точность классификации без существенного увеличения вычислительной нагрузки.

Выводы

Проведенный всесторонний анализ остаточных ошибок выявил их систематический характер, обусловленный двумя ключевыми факторами: внутренней неоднозначностью определенных характеристик запросов и ограничениями используемого синтетического набора данных. Хотя синтетический датасет обеспечивает точный контроль переменных, он создает пробелы в обработке крайних случаев (edge cases), особенно затрагивающих классы с наиболее сложным поведением (Классы 4, 5 и 11).

В качестве приоритетных направлений для улучшения определены: (1) контекстуализированная обработка временных характеристик и перенаправлений, (2) внедрение постклассификационных правил для неоднозначных случаев, и (3) расширение набора данных дополнительными примерами сложных облачных конфигураций. Stacking Ensemble продемонстрировал особую эффективность, сократив долю ложных срабатываний до 1.2 %, что подтверждает преимущество стратегического комбинирования нескольких моделей для компенсации индивидуальных ограничений каждого алгоритма.

Перспективы дальнейших исследований

Основным направлением будущих исследований станет разработка гибридного подхода к формированию datasets, сочетающего синтетически сгенерированные данные с реальными образцами сетевого трафика. Особый акцент будет сделан на тех сценариях, где искусственная природа существующего набора данных не позволяет адекватно отразить сложные поведенческие паттерны, характерные для реальных условий эксплуатации. Такой комбинированный подход сохранит преимущества контролируемой среды синтетических данных, одновременно устраняя их ключевые ограничения в сложных edge-cases.

Дополнительные перспективные направления включают разработку специализированных методов аугментации данных, ориентированных на проблемные классы (4, 5 и 11), где текущая модель демонстрирует наибольшее количество ошибок классификации. Особое внимание будет уделено интеграции современных языковых моделей для семантического анализа заголовков HTTP-запросов в спорных случаях, что позволит существенно повысить точность распознавания сложных SSRF-сценарий. Параллельно планируется исследование новых подходов к обработке временных характеристик и сложных цепочек перенаправлений, которые показали свою значимость в текущем исследовании.

ЛИТЕРАТУРА

1. Krishnaraj N., Madaan C., Awasthi S., Subramani R., Avinash H., Mukim S. Common vulnerabilities in real world web applications // Doors. 2023. Pp. 9–22.
2. OWASP OWASP Top Ten [Электронный ресурс]. 2021. URL: https://owasp.org/Top10/ (дата обращения: 12.03.2025).
3. InsiderSecurity Capital One Data Breach: How SSRF Vulnerability Exposed 100 Million Customer Records [Электронный ресурс]. 2025. URL: https://insidersecurity.co/capitalone-data-breach-how-ssrf-vulnerability-exposed-100-million-customer-records/ (дата обращения: 12.03.2025).
4. Wessels M., van der Kouwe E., Groot J. de, Bos H. SSRF vs. Developers: A Study of SSRF-Defenses in PHP Applications // Proc. of 33rd USENIX Security Symposium (USENIX Security 24). 2024. Pp. 6777–6794.
5. Wang E., Chen L., Zhang M., Li X. Where URLs Become Weapons: Automated Discovery of SSRF Vulnerabilities // Proc. of 2024 IEEE Symposium on Security and Privacy (SP). IEEE, 2024. Pp. 239–257.
6. Al-talak K., Abbass O. Detecting server-side request forgery (SSRF) attack by using deep learning techniques // International Journal of Advanced Computer Science and Applications. 2021. Vol. 12, No 12. Pp. 12.
7. Kulkarni K., Kaur K., Torvi N., Singh A., Annapurna D. Mitigating SSRF Threats: Integrating ML and Network Architecture // Proc. of International Conference on Paradigms of Communication, Computing and Data Analytics. Springer, 2024. Pp. 1–15.
8. Quiroz G.G.C., Voinov N.V., Drobintsev P.D., Zajtsev I.V. Supervised Machine Learning for SSRF Vulnerability Detection // Proc. of XXVIII International Conference on Soft Computing and Measurements (SCM). IEEE, 2025. Pp. 206–209.