Обзор открытых наборов данных для выявления атак на веб-приложения

Е.О. Еремин

***Аннотация* — В настоящее время веб-приложения являются одним из наиболее популярных средств предоставления сервисов конечному пользователю. Защита от атак на веб-приложения с каждым годом только увеличивает свою актуальность. Современные системы класса Web Application Firewall (WAF) зачастую в том или ином виде используют машинное обучение, в том числе глубокое. Исследования последних лет показывают, что системы обнаружения вредоносных HTTP-запросов с использованием классического машинного и глубокого обучения в большинстве случаев превосходят по эффективности системы, основанные исключительно на явно прописанных правилах. Тем не менее, существуют проблемы с воспроизводимостью экспериментов, описываемых в публикациях на данную тематику. В большинстве работ встречаются результаты оценки предлагаемых моделей на основе недоступных публично наборов данных (датасетов) и не рассматривается результат работы модели на нескольких доступных публично наборах данных. В данной работе представлен обзор открытых (публично доступных) наборов данных, на основе которых могут быть обучены и оценены по метрикам модели выявления вредоносных веб-запросов. В статью включены как широко используемые для оценивания моделей машинного обучения наборы данных по этому направлению задач, так и менее известные. Кроме того, обзор может быть полезен при формировании комбинированных наборов данных.**

***Ключевые слова* — брандмауэр веб-приложений, машинное обучение, глубокое обучение, датасет, веб-атака, http-запрос, выявление атак, информационная безопасность, кибербезопасность.**

I. Введение

Несмотря на бурное развитие отрасли кибербезопасности, количество уязвимостей в постоянно растущем множестве веб-приложений не уменьшается [1]. Кроме этого, появляются новые типы уязвимостей, связанных с введением в эксплуатацию новых технологий, форматов и протоколов. Борьба с уязвимостями веб-приложений идет с двух позиций:

1. написание более безопасного кода;
2. использование систем класса Web Application Firewall (WAF).

При этом в случае развертывания на своем веб-сервере сторонних веб-приложений говорить о первом подходе не приходится. Использование же WAF предоставляет возможность до какой-то степени обезопасить потенциально уязвимое приложение от запросов, которые выглядят похожими на эксплуатацию некоторого типа уязвимости. При этом эффективность WAF напрямую зависит от правил, по которым система принимает решение о подозрительности или вредоносности запроса. На сегодняшний день приложения класса WAF можно классифицировать по типу используемых правил детектирования на следующие:

− основанные на явно прописанных конструкциях, в том числе с использованием регулярных выражений;

− основанные на моделях машинного обучения, как классического, так и глубокого;

− гибридные.

В свою очередь модели, на которых основаны системы на машинном обучении, можно классифицировать по типу обучения следующим образом:

− модели, обученные с учителем (supervised), используются для бинарной или мультиклассовой классификации веб-запросов. При бинарной классификации модель относит веб-запрос к «чистым» либо вредоносным, при мультиклассовой же вредоносный веб-запрос должен быть классифицирован конкретным типом уязвимости (sql injection, xss, path traversal, ssti и т.д.). В ряде публикаций этот подход называется сигнатурным (в контексте машинного обучения) [2,3];

− модели, обученные без учителя (unsupervised), используются для выявления аномальных веб-запросов, в том числе атак нулевого дня (zero-day attacks). Такие модели обучаются на легитимных запросах (в том числе могут быть обучены на конкретном веб-приложении), и выявляют те запросы, которые на них не похожи (с использованием таких методов, как k-means, one-class svm, isolation forest и т.д.) Данный подход отличается более высокой (по сравнению с сигнатурными моделями) долей ложнопозитивных сработок (false positive) – случаев, когда легитимный запрос ошибочно относится моделью к вредоносным [4];

−модели, частично обученные с учителем (semi-supervised), используются при наличии не полностью размеченного датасета [4].

Статья получена 18 декабря 2023.
Е. О. Еремин, кандидат технических наук, Infosecurity a Softline company, Россия (e-mail: e.o.eremin@gmail.com).

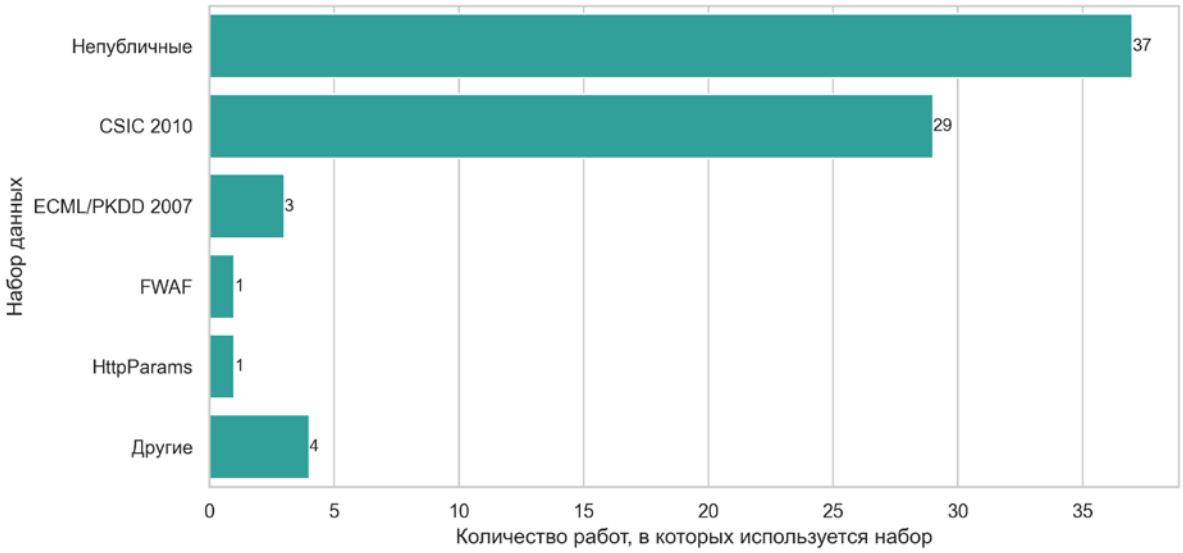

Рис. 1. Статистика использования наборов данных согласно исследованию [2]. Из статистики исключены наборы данных на основе PCAP и Netflow

Наиболее сбалансированным является подход, при котором используются как правила на регулярных выражениях, так и модели машинного обучения, причём как сигнатурные, так и для выявления аномалий.

Однако качество моделей машинного обучения напрямую зависит от качества данных, на которых они обучены и протестированы. По тематике выявления вредоносных http-запросов существуют открытые датасеты, благодаря чему в круг исследователей, которые изучают это направление, может войти любой желающий. Тем не менее, существующие открытые наборы данных не лишены недостатков, и их количество невелико, что открывает отдельное перспективное направление работ [6,7].

Данный обзор сосредоточен на наборах данных, позволяющих оценивать запросы на уровне приложений, т.е. датасеты для Intrusion Detection Systems (IDS) намеренно не рассмотрены, хотя существуют подходы к выявлению веб-атак на основе характеристик сетевого трафика [8].

Дальнейшее содержимое статьи организовано следующим образом: в разделе 2 приведен перечень работ, в которых приведены обзоры датасетов на данную тематику; в разделе 3 рассмотрены общеизвестные размеченные наборы, используемые для оценки метрик разрабатываемых моделей машинного обучения; в разделе 4 рассмотрены менее известные наборы данных, в том числе «сырые» лог-файлы работы веб-серверов; в разделе 5 кратко освещены особенности, сопутствующие созданию комбинированных наборов данных; в разделе 6 приведены выводы и сделано заключение.

II. Обзор работ по тематике

Данные, на основе которых обучаются модели, являются основным залогом положительных результатов при решении прикладных задач с помощью машинного обучения, причем касается это не только информационной безопасности. Существующие наборы данных в области кибербезопасности постепенно устаревают и могут оказаться недостаточными с точки зрения понимания современных поведенческих моделей различных кибератак. Тем не менее, сложилась практика оценки новых state-of-the-art моделей на существующих наборах данных с целью демонстрации значений метрик. Большинство исследователей использует ограниченное число наборов данных для этой цели.

Насколько известно автору, обзоров именно на наборы данных для выявления вредоносных веб-запросов на седьмом уровне модели OSI не проводилось. Тем не менее, существует ряд обзорных работ по датасетам для решения задач кибербезопасности, в том числе IDS, а также системные обзоры статей по машинному и глубокому обучению для выявления атак на веб-приложения [6,7,9,10].

В работе [2] приведена статистика используемых наборов данных в статьях, опубликованных на английском языке с 2010 по 2021 годы. Как видно на рисунке 1, в большинстве случаев для оценки разработанной модели машинного обучения на веб-приложение используется либо непубличный набор, либо CSIC 2010. В работе [2] в статистике так же присутствуют CICIDS 2017 и USNW-NB15, но необходимо отметить, что эти наборы данных предназначены для систем обнаружения и предотвращения вторжения, работающих на сетевом уровне, а не для брандмауэров веб-приложений, поэтому из графика на рисунке 1 эти наборы исключены.

В обзоре материалов по данной тематике [3] также сказано о том, что для обучения моделей машинного обучения в большинстве случаев используются закрытые наборы данных, на которых невозможно воспроизводить эксперименты. При этом для первоначальной оценки модели часто используются открытые наборы данных, которых не так много, они

```
POST http://localhost:8080/tienda1/publico/anadir.jsp HTTP/1.1
User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko)
Pragma: no-cache
Cache-control: no-cache
Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5
Accept-Encoding: x-gzip, x-deflate, gzip, deflate
Accept-Charset: utf-8, utf-8;q=0.5, *;q=0.5
Accept-Language: en
Host: localhost:8080
Cookie: JSESSIONID=AE29AEEBDE479D5E1A18B4108C8E3CE0
Content-Type: application/x-www-form-urlencoded
Connection: close
Content-Length: 146

id=2&nombre=Jam%F3n+Ib%E9rico&precio=85&cantidad=%27%3B+DROP+TABLE+usuarios%3B+SELECT+*+FROM+datos+WHERE+nombre+LIKE+%27%25&B1=A%F1adir+al+carrito

GET http://localhost:8080/tienda1/publico/anadir.jsp?id=2%2F&nombre=Jam%F3n+Ib%E9rico&precio=85&cantidad=49&B1=A%F1adir+al+carrito HTTP/1.1
User-Agent: Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.8 (like Gecko)
Pragma: no-cache
Cache-control: no-cache
Accept: text/xml,application/xml,application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,image/png,*/*;q=0.5
Accept-Encoding: x-gzip, x-deflate, gzip, deflate
Accept-Charset: utf-8, utf-8;q=0.5, *;q=0.5
Accept-Language: en
Host: localhost:8080
Cookie: JSESSIONID=F563B5262843F12ECAE41815ABDEEA54
Connection: close
```

Рис. 2. Пример аномальных запросов из CSIC 2010

зачастую не актуализируются и содержат ограниченное количество типов атак на веб-приложения (в основном, sql injection, cross-site scripting и cross-site request forgery). Кроме того, зачастую в них не отражена сложность реальных веб-приложений, что может ограничивать пользу от эксплуатации моделей в реальных условиях.

Ни один из существующих наборов данных не покрывает целиком все типы атак, предлагаемых в классификации OWASP и таксономии, рассмотренной в работе [1].

Далее рассмотрен ряд публичных наборов данных, в том числе разных типов, которые могут быть использованы при формировании комбинированных датасетов, в том числе синтетических. Построение комбинированного набора данных может помочь обойти часть проблем, которые присущи существующим публичным датасетам.

III. Специальные наборы данных для оценки метрик моделей машинного обучения

Можно выделить датасеты, широко используемые для оценки вновь разрабатываемых моделей:

- CSIC-2010;
- ECML/PKDD 2007.

На момент написания статьи можно отнести в эту же категорию и датасет SR-BH 2020, который опубликован в 2022 году и число ссылок на него растёт.

A. *CSIC 2010*

Набор данных разработан Институтом информационной безопасности CSIC (Consejo Superior de Investigaciones Científicas), содержит трафик для веб-приложений электронной коммерции, а именно 36 000 обычных запросов и 25 065 аномальных запросов [11]. Аномальные запросы включают в себя такие типы атак как SQL-инъекция, переполнение буфера, инъекция возврата/перевода строки (CRLF), межсайтовый скриптинг (XSS), Server Side Include и ряд других. При этом мультиклассовая разметка в датасете отсутствует.

Де-факто этот датасет является стандартом при оценивании вновь предлагаемых моделей в рамках исследований по выявлению атак на веб-приложения с применением машинного обучения. На рисунке 2 представлены примеры аномальных (потенциально вредоносных) GET- и POST-запросов.

B. *ECML/PKDD 2007*

Этот набор данных создан для конференции European Conference on Machine Learning and Knowledge Discovery 2007 в рамках конкурса для интеллектуального

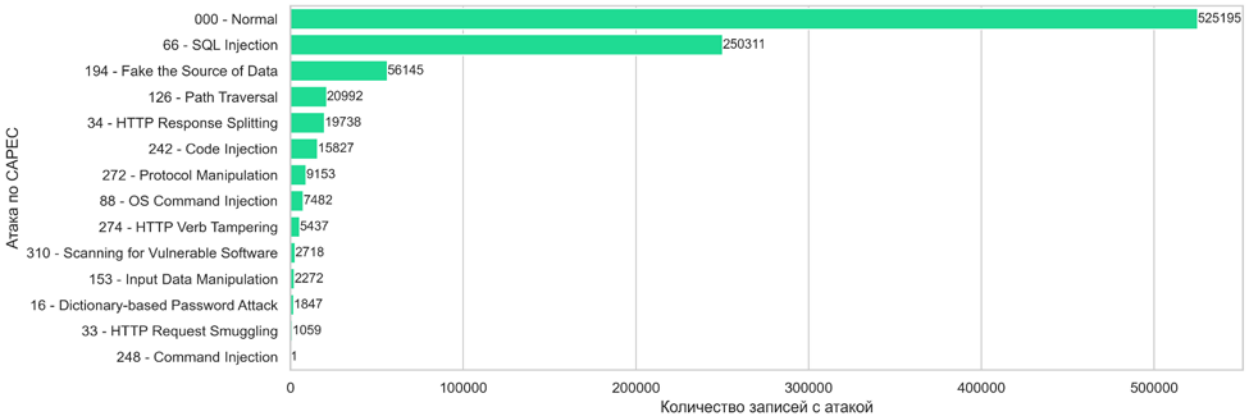

Рис. 3. Распределение типов атак по классификации CAPEC в наборе данных SR BH 2020

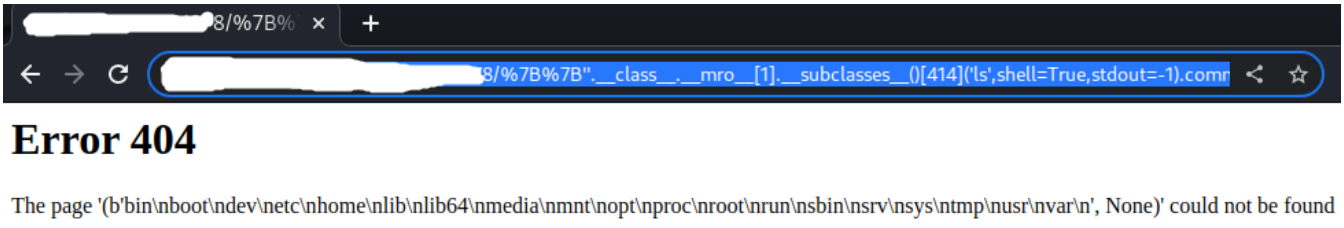

Рис. 4. Пример выполнения запроса, который может быть отнесён к двум классам: Server Side Template Injection и OS Command Injection

выявления сетевых атак [12]. Помимо http-запросов с более длинным списком заголовков (по сравнению с CSIC 2010), присутствует также информация об операционной системе, веб-сервере, функционировании LDAP, СУБД, XPath, а также о типе атаки и о том, находится ли она в контексте применения (сработала или нет). Содержит около 35006 нормальных и 15110 аномальных запросов.

C. SR-BH 2020

Новый набор данных, опубликованный в 2022 году и на который уже ссылается несколько работ по машинному обучению для выявления атак на веб-приложения [13]. Набор размечен автоматически с использованием open-source WAF ModSecurity с набором правил CoreRuleSet от OWASP. На рисунке 3 изображено распределение меток в этом датасете.

Как видно, существует дисбаланс не только между нормальными и вредоносными запросами, но и между разными классами атак. Отсутствуют некоторые классы атак. К достоинствам набора данных можно отнести особенности разметки запросов, когда один запрос может принадлежать к нескольким классам одновременно. На рисунке 4 показан результат выполнения подобного запроса (не из SR BH 2020, SSTI в нём не присутствует), который можно отнести к разным классам атак.

IV. Остальные наборы данных

A. Наборы данных с единичными использованиями

1) FWAF

Набор FWAF [14] представляет собой два файла, содержащих легитимные и нелегитимные «полезные нагрузки» (payloads) в теле http-запроса. Всего 1294531 нормальных и 48126 вредоносных записей. Классификация по типу атаки отсутствует. Примеры вредоносных запросов представлены на рисунке 5. К набору данных приложена базовая модель, обученная методом логистической регрессии на входных признаках, построенных с помощью TF-IDF на n-граммах.

2) Machine-Learning-Web-Application-Firewall-and-Dataset (MLWAF)

В этом наборе данных присутствуют поля «полезной нагрузки» (пользовательский ввод, который может присутствовать в каком-либо http-заголовке запроса к веб-серверу), признак вредности (аномальности) и тип вредоносной нагрузки. Присутствуют такие типы вредоносных «полезных нагрузок» как shell, sql injection и xss. Содержит 93888 легитимных записей и 9278 вредоносных.

Кроме сформированного датасета, в [15] проведено сравнение большого количества моделей, из которых наилучшими показателями обладает модель случайного леса, в качестве признаков использующая вектор TF-IDF на основе n-грамм из 2-ух символов. При этом сравнение моделей проводилось как для векторов

```
/main.php?logout="&del\x0cq31768299&rem\x0c
/\xd0\x97\xd0\xb4\xd0\xbe\xd1\x80\xd0\xbe\xd0\xb2'\xd1\x8f/
/themes/modern/user_style.php?user_colors[bg_color]="</style><script>alert(411136083423)</script>
/nyjgaorz.mscgi?<img src="javascript:alert(cross_site_scripting.nasl);">
/cgi-bin/index.php?op=default&date=200607' union select 1,501184215,1,1,1,1,1,1,1--&blogid=1
/scripts/cfooter.php3
/en-us/dda2qr7j.fts?<script>cross_site_scripting.nasl</script>
/?<meta http-equiv=set-cookie content="testpokn=7494">
/169okeyj.jspa?<meta http-equiv=set-cookie content="testxeoi=3573">
/j734qobz.aspx?<img src="javascript:alert(cross_site_scripting.nasl);">
/examples/jsp/error/forum.php?do=viewtopic&cat=1&topic=1&page=1?<script>foo</script
/post.asp?method=topic&forum_id=1&cat_id=1&type=wvs-xss-magic-string-410025568136
<svg xmlns="http://www.w3.org/2000/svg" onload="alert(document.domain)"/>
GIF89a/*<svg/onload=alert(1)>*/=alert(document.domain)//;
<script src="data:&comma;alert(1)//
"><script src=data:&comma;alert(1)//
<script src="//brutelogic.com.br&sol;1.js&num;
"><script src=//brutelogic.com.br&sol;1.js&num;
<link rel=import href="data:text/html&comma;&lt;script&gt;alert(1)&lt;&sol;script&gt;
"><link rel=import href=data:text/html&comma;&lt;script&gt;alert(1)&lt;&sol;script&gt;
<base href=//0>
<script/src="data:&comma;eval(atob(location.hash.slice(1)))//#alert(1)
<body onload=alert(1)>
<body onpageshow=alert(1)>
<body onfocus=alert(1)>
<body onhashchange=alert(1)><a href=#x>click this!#x
```

Рис. 5. Примеры вредоносных запросов из набора данных FWAF

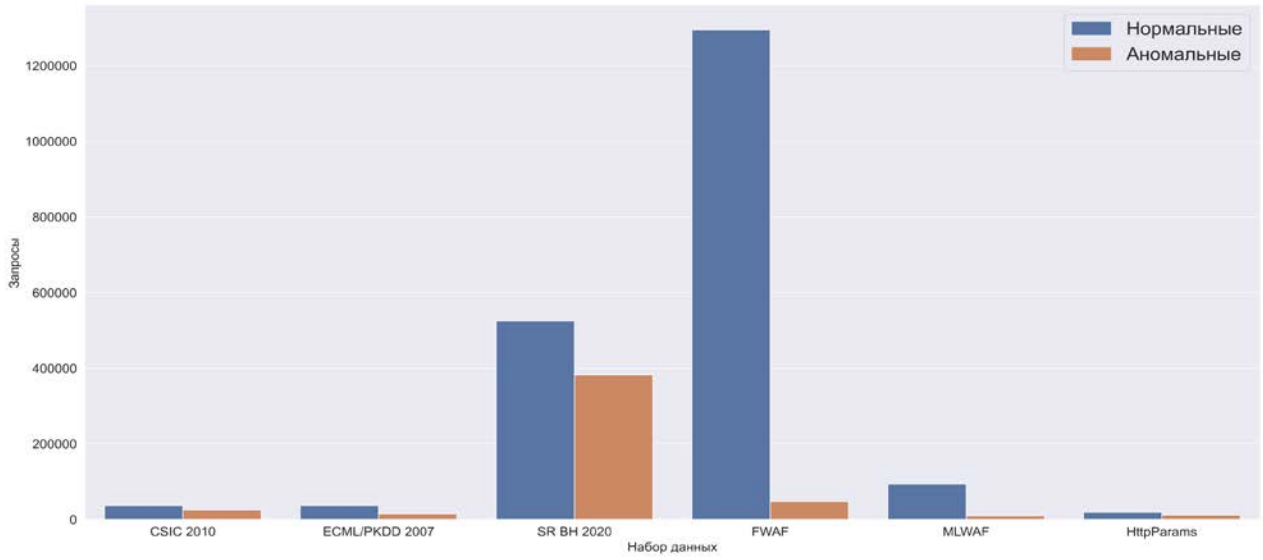

Рис. 6. Сравнительный размер наборов данных, имеющих разметку

признаков, построенных по методу bag-of-words (на основе n-грамм), так и для набора специально выбранных признаков:

- длина строки запроса;
- количество непечатаемых символов в запросе;
- количество знаков препинания в запросе;
- минимальное значение среди байтов в запросе;
- максимальное значение среди байтов в запросе;
- среднее значение среди байтов в запросе;
- стандартное отклонение значений байтов в запросе;
- количество различных байтов в запросе;
- количество ключевых слов SQL в запросе;
- количество ключевых слов javascript в запросе.

3) HttpParams

Структура набора HTTPParams [16] схожа с MLWAF, но дополнительно содержит поле длины «полезной нагрузки». Присутствуют такие типы вредоносных нагрузок как command injection, sql injection, path traversal и xss. Содержит 19304 нормальных и 11763 вредоносных записей. Нормальные записи подобраны из CSIC 2010, вредоносные сгенерированы с использованием ряда утилит, а также базы данных для фаззинга fuzzdb [17].

4) Vulnbank – PTSecurity

Этот набор данных [18] построен из запросов к демонстрационному уязвимому веб-приложению. Начало и окончание каждого запроса помечено маркерами ST@RT и END, а маркером INFO отмечена дата. Содержит 21382 легитимных запросов и 1096 аномальных. В аномальных запросах нелегитимные значения содержатся не только в URL, но и в таких полях, как X-Forwarded-For, Pragma и др. Кроме набора данных, представлена модель Seq2seq для выявления аномалий, построенная с помощью tensorflow и обученная на полных http-запросах к веб-серверу.

5) Boss of the soc v1

Набор данных с соревнования 2015 года по выявлению и реагированию на инциденты от компании Splunk [19]. Набор содержит логи от большого количества источников, включая разные журналы Windows, а также веб-сервера IIS. Набор не размечен, имеет размер 22616 записей, содержит несколько типов атак. Польза от такого набора данных неочевидна, но данные из него теоретически могут быть использованы в собственном комбинированном датасете.

B. «Сырые» логи веб-серверов

Сайт Secrepo.com [20] содержит ссылки на ресурсы с лог-файлами, имеющими отношение к информационной безопасности. Кроме прочих, доступны также лог-файлы веб-сервера самого сайта. Лог-файлы не размечены.

C. Подборки данных для фаззинга веб-приложений

В некоторых упомянутых ранее наборах данных при формировании аномальных запросов использовались «полезные нагрузки», соответствующие различным типам атак, в том числе с ресурсов для фаззинга приложений, а именно PayloadAllTheThings [21] и fuzzdb [17]. Данные ресурсы содержат файлы для таких инструментов, как Burp Intruder и аналогичных. «Полезные нагрузки» классифицированы по категориям атак, что упрощает построение собственного размеченного датасета.

D. Подборки правил детектирования WAF и SIEM

1) Core rule set

Правила для open source WAF ModSecurity [22], являющегося библиотекой, которая может быть подключена к основным популярным веб-серверам. Сам по себе набор правил содержит в себе списки и регулярные выражения. Теоретически списки могут

быть использованы для обучения с учителем с метками вредоносных запросов.

2) Sigma rules для web

Для систем класса SIEM (Security information and event management) открытым стандартом описания правил корреляции де-факто является SIGMA [23]. На основе правил корреляции для веб-серверов осуществляется поиск подстрок в каждой записи access- и error- логов для Apache http server, Apache tomcat, nginx, а также в едином файле логов Microsoft IIS. Кроме того, присутствуют индикаторы конкретных CVE для популярных веб-приложений. Данные могут быть использованы в собственном комбинированном датасете.

V. Особенности формирования комбинированных наборов данных

Существующие наборы данных имеют дисбаланс как по корректным/вредоносным запросам, так и по разным типам атак среди вредоносных. Сравнительные размеры размеченных наборов данных представлены на рисунке 6. Обучение классификатора по несбалансированному набору данных может давать на валидационной выборке гораздо более высокие значения метрик, чем при реальной эксплуатации модели (inference). Выходом может послужить создание комбинированного датасета, в том числе с частичным синтетическим формированием.

Однако при формировании комбинированного набора данных на основе общедоступных возникают определенные трудности:

– существующие наборы данных имеют различную структуру, какие-то построены на основе полных http-запросов и ответов, какие-то на основе access-логов веб-серверов, а другие просто имеют одно поле с «полезной нагрузкой», которая может присутствовать в разных заголовках http-запроса;

– не все существующие наборы данных имеют разметку, среди имеющих разметку не все разделяют различные типы атак, а среди разделяющих - типы атак размечены по разным классификациям, причем иногда не придерживаясь формальных названий атак в таксономии.

Тем не менее, эти трудности принципиально решаемы и связаны скорее с необходимостью использования больших трудозатрат.

Создание комбинированного набора с единой структурой может быть решено либо путём дополнения недостающих полей до структуры более полных датасетов, либо сужением более полных датасетов до структуры с одним полем корректной/вредоносной «полезной нагрузки». Очевидно, при втором варианте теряется часть информации, на основе которой модель может чему-то научиться, поэтому предпочтительным является скорее первый подход. Дополнение недостающих полей может быть выполнено синтетически, либо полным повторением эксперимента по сбору исходных датасетов с развертыванием стенда веб-сервера и сбором http-запросов и ответов. В некоторых случаях повторение эксперимента невозможно в виду недоступности исходных текстов веб-приложений, на которых они проводились.

Трудности с разметкой практически целиком обусловлены временными трудозатратами эксперта, проводящего разметку/переразметку. Категории атак по разным классификациям теоретически возможно переразметить по одной наиболее полной классификации. Как отмечалось ранее, большой разбор различных таксономий и классификаций веб-уязвимостей проведён в работе [1]. В некоторых случаях это потребует полной переразметки датасета.

VI. Заключение

В статье приведен обзор как широко используемых открытых наборов данных для выявления вредоносных веб-запросов, так и менее популярных. Кратко рассмотрены характеристики представленных наборов данных. По результатам обзора можно сделать вывод, что на текущий момент существует дефицит больших и качественных наборов данных с разметкой. Ряд других публикаций также подтверждают этот вывод.

В большинстве работ по данной тематике используются непубличные датасеты, а также широко распространенные CSIC 2010 и ECML/PKDD 2007, уже значительно устаревшие и не содержащие актуальных вредоносных «полезных нагрузок». В результате практически невозможно воспроизвести оценки вновь разрабатываемых state-of-the-art моделей.

Тем не менее, для выхода из сложившейся ситуации при разработке новых моделей могут быть:

– использованы подходы к кросс-валидации по нескольким публичным датасетам;

– собраны комбинированные наборы на основе существующих;

– созданы новые наборы по примеру SR-BH 2020, в котором разметка произведена с помощью CoreRuleSet ModSecurity.

Библиография

[1] Y.Sadqi, and Y.Maleh. (2022). A systematic review and taxonomy of web applications threats. Information Security Journal: A Global Perspective, Taylor & Francis, 31, 1-27, DOI: 10.1080/19393555.2020.1853855

[2] R.L.Alaoui, and E.H.Nfaoui. (2022). Deep Learning for Vulnerability and Attack Detection on Web Applications: A Systematic Literature Review. Future Internet, 14(4), 118, DOI:10.3390/fi14040118

[3] Hassan I. Halim, Mohamed Kholief, Fahima Maghraby et al. Deep Learning Methods in Web Intrusion Detection: A Systematic Review, 14 November 2022, PREPRINT (Version 1) available at Research Square. DOI:10.21203/rs.3.rs-2214647/v1

[4] Toprak, S. & Yavuz, A.G. (2022). Web application firewall based on anomaly detection using deep learning. Acta Infologica. Advance online publication. DOI:10.26650/acin.1039042

[5] B.A.Tama, and S.Lim. (2021). Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation. Computer Science Review, 39, 100357. DOI:10.1016/j.cosrev.2020.100357

[6] Catillo, M., Pecchia, A., Rak, M., and Villano, U. (2021). Demystifying the role of public intrusion datasets: a replication study of DoS network traffic data. Computers & Security, Elsevier, 108, 102341. DOI:10.1016/j.cose.2021.102341

[7] Alshaibi, A.; Al-Ani, M.; Al-Azzawi, A.; Konev, A.; Shelupanov, A. The Comparison of Cybersecurity Datasets. Data 2022, 7, 22. DOI: 10.3390/data7020022

[8] Гетьман А.И., Горюнов М.Н., Мацкевич А.Г., Рыболовлев Д.А. Сравнение системы обнаружения вторжений на основе машинного обучения с сигнатурными средствами защиты

информации. Труды ИСП РАН, том 34, вып. 5, 2022 г., стр. 111-126. DOI: 10.15514/ISPRAS-2022-34(5)-7

[9] Kaniski, Matija & Dobša, Jasminka & Kermek, Dragutin. (2023). Deep Learning within the Web Application Security Scope - Literature Review. DOI:10.23919/MIPRO57284.2023.10159847.

[10] Ерохин С.Д., Журавлев А.П. Сравнительный анализ открытых наборов данных для использования технологий искусственного интеллекта при решении задач информационной безопасности // Системы синхронизации, формирования и обработки сигналов, 2020. Т.3. №3. С. 12-19.

[11] Gim´enez, C. T., Villegas, A. P., and Mara˜n´on, G. A. (2010). HTTP data set CSIC 2010. Information Security Institute of CSIC (Spanish Research National Council), 64, https://www.isi.csic.es/dataset/ Retrieved: 17.12.2023

[12] Chedy Ra¨ıssi, Johan Brissaud, G´erard Dray, Pascal Poncelet, Mathieu Roche, et al, Web Analyzing Traffic Challenge: Description and Results. ECML PKDD 2007 Discovery Challenge, 2007, Warsaw, Poland

[13] Riera, T. S., Higuera, J. B., Higuera, J. B., Herraiz, J. M., and Montalvo, J. S. (2022). A new multi-label dataset for Web attacks CAPEC classification using machine learning techniques. Computers & Security, Elsevier, 120, 102788. DOI: 10.1016/j.cose.2022.102788.

[14] FWAF Machine Learning driven Web Firewall https://github.com/faizann24/Fwaf-Machine-Learning-driven-Web-Application-Firewall Retrieved: 17.12.2023

[15] Machine Learning Web Application Firewall and Dataset https://github.com/grananqvist/Machine-Learning-Web-Application-Firewall-and-Dataset Retrieved: 17.12.2023

[16] HttpParams dataset https://github.com/Morzeux/HttpParamsDataset Retrieved: 17.12.2023

[17] FuzzDB https://github.com/fuzzdb-project/fuzzdb Retrieved: 17.12.2023

[18] Vulnbank dataset https://github.com/PositiveTechnologies/seq2seq-web-attack-detection Retrieved: 17.12.2023

[19] Boss of the SOC v1 dataset https://github.com/splunk/botsv1 Retrieved: 17.12.2023

[20] Samples of security related data https://www.secrepo.com/ Retrieved: 17.12.2023

[21] PayloadAllTheThings https://github.com/swisskyrepo/PayloadsAllTheThings Retrieved: 17.12.2023

[22] OWASP ModSecurity Core Rule Set (CRS) https://github.com/coreruleset/coreruleset Retrieved: 17.12.2023

[23] Sigma rules for SIEM https://github.com/SigmaHQ Retrieved: 17.12.2023

Overview of public datasets for web application attack detecting

Evgeny Eremin

Abstract. **Currently, web applications are one of the most popular means of providing services to the end user. Protection against attacks on web applications only increases its relevance every year. Modern Web Application Firewalls (WAF) often use machine learning in one form or another. Recent research shows that systems for detecting malicious HTTP requests using classical machine learning and deep learning in most cases outperform systems based on explicitly prescribed rules in terms of efficiency. Nevertheless, there are problems with reproducibility of experiments described in publications on this topic. In most works, there are results of evaluating proposed models based on publicly unavailable datasets; the result of the model's work on several publicly available datasets is not considered. This paper provides an overview of open (publicly available) datasets on the basis of which models for detecting malicious web requests can be trained and evaluated by metrics. This overview includes both widely used datasets for benchmarks and less well-known ones. In addition, the overview can be useful when forming combined datasets.**

Keywords: **web application firewall, machine learning, deep learning, dataset, web attack, http request, attack detection, information security, cybersecurity.**

REFERENCES

[1] Y.Sadqi, and Y.Maleh. (2022). A systematic review and taxonomy of web applications threats. Information Security Journal: A Global Perspective, Taylor & Francis, 31, 1-27, DOI: 10.1080/19393555.2020.1853855

[2] R.L.Alaoui, and E.H.Nfaoui. (2022). Deep Learning for Vulnerability and Attack Detection on Web Applications: A Systematic Literature Review. Future Internet, 14(4), 118, DOI:10.3390/fi14040118

[3] Hassan I. Halim, Mohamed Kholief, Fahima Maghraby et al. Deep Learning Methods in Web Intrusion Detection: A Systematic Review, 14 November 2022, PREPRINT (Version 1) available at Research Square. DOI:10.21203/rs.3.rs-2214647/v1

[4] Toprak, S. & Yavuz, A.G. (2022). Web application firewall based on anomaly detection using deep learning. Acta Infologica. Advance online publication. DOI:10.26650/acin.1039042

[5] B.A.Tama, and S.Lim. (2021). Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation. Computer Science Review, 39, 100357. DOI:10.1016/j.cosrev.2020.100357

[6] Catillo, M., Pecchia, A., Rak, M., and Villano, U. (2021). Demystifying the role of public intrusion datasets: a replication study of DoS network traffic data. Computers & Security, Elsevier, 108, 102341. DOI:10.1016/j.cose.2021.102341

[7] Alshaibi, A.; Al-Ani, M.; Al-Azzawi, A.; Konev, A.; Shelupanov, A. The Comparison of Cybersecurity Datasets. Data 2022, 7, 22. DOI: 10.3390/data7020022

[8] Get'man A.I., Goryunov M.N., Mackevich A.G., Rybolovlev D.A. Sravnenie sistemy obnaruzheniya vtorzhenij na osnove mashinnogo obucheniya s signaturnymi sredstvami zashchity informacii. Trudy ISP RAN, tom 34, vyp. 5, 2022 g., str. 111-126. DOI: 10.15514/ISPRAS-2022-34(5)-7

[9] Kaniski, Matija & Dobša, Jasminka & Kermek, Dragutin. (2023). Deep Learning within the Web Application Security Scope - Literature Review. DOI:10.23919/MIPRO57284.2023.10159847.

[10] Erohin S.D., Zhuravlev A.P. Sravnitel'nyj analiz otkrytyh naborov dannyh dlya ispol'zovaniya tekhnologij iskusstvennogo intellekta pri reshenii zadach informacionnoj bezopasnosti // Sistemy sinhronizacii, formirovaniya i obrabotki signalov, 2020. T.3. №3. S. 12-19.

[11] Gim´enez, C. T., Villegas, A. P., and Mara˜n´on, G. A. (2010). HTTP data set CSIC 2010. Information Security Institute of CSIC (Spanish Research National Council), 64, https://www.isi.csic.es/dataset/ Retrieved: 17.12.2023

[12] Chedy Ra¨ıssi, Johan Brissaud, G´erard Dray, Pascal Poncelet, Mathieu Roche, et al, Web Analyzing Traffic Challenge: Description and Results. ECML PKDD 2007 Discovery Challenge, 2007, Warsaw, Poland

[13] Riera, T. S., Higuera, J. B., Higuera, J. B., Herraiz, J. M., and Montalvo, J. S. (2022). A new multi-label dataset for Web attacks CAPEC classification using machine learning techniques. Computers & Security, Elsevier, 120, 102788. DOI: 10.1016/j.cose.2022.102788.

[14] FWAF Machine Learning driven Web Firewall https://github.com/faizann24/Fwaf-Machine-Learning-driven-Web-Application-Firewall Retrieved: 17.12.2023

[15] Machine Learning Web Application Firewall and Dataset https://github.com/grananqvist/Machine-Learning-Web-Application-Firewall-and-Dataset Retrieved: 17.12.2023

[16] HttpParams dataset https://github.com/Morzeux/HttpParamsDataset Retrieved: 17.12.2023

[17] FuzzDB https://github.com/fuzzdb-project/fuzzdb Retrieved: 17.12.2023

[18] Vulnbank dataset https://github.com/PositiveTechnologies/seq2seq-web-attack-detection Retrieved: 17.12.2023

[19] Boss of the SOC v1 dataset https://github.com/splunk/botsv1 Retrieved: 17.12.2023

[20] Samples of security related data https://www.secrepo.com/ Retrieved: 17.12.2023

[21] PayloadAllTheThings https://github.com/swisskyrepo/PayloadsAllTheThings Retrieved: 17.12.2023

[22] OWASP ModSecurity Core Rule Set (CRS) https://github.com/coreruleset/coreruleset Retrieved: 17.12.2023

[23] Sigma rules for SIEM https://github.com/SigmaHQ Retrieved: 17.12.2023