Information Technology Innovations and Applied Communication

Vol. 3, No. 3, Autumn, 2025, Serial No.7,pp.55-69

Detecting SQL Injection in database systems is based on the features Extraction

Seyyed Mostafa Taghavimanesh[1], Majid Ghayouri-Sales[*2]

* Assistant Professor, Imam Hossein Comprehensive University.

(**Received:** 23 September 2025, **Revised:** 09 November 2025, **Accepted:** 17 November 2025, **Published:** 27 November 2025)

Abstract

In the past years, the Internet has been widely used for a variety of applications. Although the systems have become more complex, the attacks have also become more complex and widespread, and traditional methods such as intrusion detection and prevention systems are not able to prevent all attacks. Intrusion detection using machine learning is becoming a widespread logic in recent research that can be a suitable alternative to traditional methods. This article aims to detect SQL injection attacks by focusing on extracting effective features from several appropriate data sets, and tries to recognize these types of attacks more effectively. In this context, using machine learning algorithms, including logistic regression and random forest, he has performed various experiments on three data sets and tested manual and automatic feature selection processes with the relevant algorithms. It has also evaluated whether some features are more important than others and displayed the results. From the results obtained in this research, it was found that feature extraction from data sets with high specialized knowledge in this field and selection of optimized basic tables to determine features leads to high accuracy results and we reached 99.57% accuracy.

Keywords: SQL injection attacks, Database, Effective features, influence, machine learning.

* Corresponding Author E-mail: ghayoori@ihu.ac.ir

«نوآوری‌های فناوری اطلاعات و ارتباطات کاربردی»
سال سوم، شماره ۳، فصل پاییز ۱۴۰۴؛ ص ۶۹-۵۵.

مروری- ترویجی

تشخیص نفوذ تزریق کد در سیستم‌های پایگاه‌داده مبتنی بر استخراج ویژگی‌های مؤثر

سیدمصطفی تقوی‌منش[۱]، مجید غیوری‌ثالث[۲*]

۱-کارشناسی ارشد ؛ ۲- استادیار، دانشگاه جامع امام حسین(ع)، تهران ، ایران

(**دریافت:** ۱ مهر۱٤۰٤، **بازنگری:** ۱۸ آبان ۱٤۰٤، **پذیرش:** ۲٦ آبان۱٤۰٤ ، **انتشار:** ۰٦ آذر ۱٤۰٤)

چکیده

در سال‌های گذشته اینترنت به‌صورت گسترده‌ای برای انواع کاربردها مورداستفاده قرار گرفته است. هرچند که سیستم‌ها پیچیده‌تر شده‌اند، حملات هم به‌تبع آن پیچیده‌تر و گسترده‌تر شده‌اند و روش‌های سنتی مثل سیستم تشخیص نفوذ[۱] و سیستم جلوگیری از نفوذ[۲] قادر به جلوگیری از همه حملات نیستند. تشخیص نفوذ با استفاده از یادگیری ماشین در حال تبدیل‌شدن به یک منطق گسترده در تحقیقات اخیر شده که می‌تواند جایگزین مناسبی برای روش‌های سنتی باشد. این مقاله باهدف تشخیص حملات تزریق کد[۳] با تمرکز بر روی استخراج ویژگی‌های مؤثر از چند مجموعه‌داده مناسب، سعی در شناخت مؤثرتر این‌گونه حملات را دارد. در این زمینه با استفاده از الگوریتم‌های یادگیری ماشین ازجمله مدل آماری رگرسیون متغیرهای وابسته[۴] و درخت تصمیم‌گیری[۵] آزمایش‌های مختلفی را بر روی سه مجموعه‌داده انجام داده و فرایندهای انتخاب ویژگی دستی و خودکار را با الگوریتم‌های مربوطه آزمایش کرده است. همچنین ارزیابی در مورد اینکه آیا برخی از ویژگی‌ها نسبت به برخی دیگر دارای اهمیت بالاتری هستند را به عمل‌آورده و نتایج حاصله را نمایش داده است. از خروجی‌های به‌دست‌آمده در این تحقیق، مشخص گردید که استخراج ویژگی از مجموعه‌داده‌ها بادانش تخصصی بالا در این حوزه و انتخاب جداول پایه‌ای بهینه‌سازی شده برای تعیین ویژگی‌ها، منجر به تولید نتایج بادقت بالا می‌گردد و به‌دقت ۹۹،۵۷ درصد رسیدیم.

کلیدواژه‌ها: حملات تزریق SQL، پایگاه داده، ویژگی‌های موثر، نفوذ، یادگیری ماشین.

* رایانامه نویسنده مسئول: E-mail: ghayoori@ihu.ac.ir

[1] Intrusion Detection System(IDS)
[2] Intrusion Prevention System(IPS)
[3] SQL Injection
[4] Logistic Regression
[5] Random Forest

۱- مقدمه

بسیاری از سازمان‌ها بانک‌های اطلاعاتی را به‌عنوان فناوری اصلی مدیریت داده‌ها به‌منظور ذخیره و دسترسی به داده‌های خود به کار می‌برند. این اطلاعات اغلب به‌عنوان یک دارایی باارزش در سازمان‌ها در نظر گرفته‌شده است. این یک موضوع جدی برای محافظت از این اطلاعات در برابر مزاحمان است. ازاین‌رو، داده‌ها در پایگاه‌داده‌ها باید از دسترسی‌های غیرمجاز و تغییرات محافظت شود. امروزه سازوکارهای سنتی امنیتی پایگاه‌داده از قبیل رمزنگاری، کنترل دسترسی و تصدیق اصالت تا حدودی سطح امنیت پایگاه‌داده را فراهم می‌کنند، اما برای حفاظت از اطلاعات حساس در مقابل حملات امنیتی نوآورانه کافی نیستند. بااین‌وجود سامانه‌های تشخیص نفوذ در پایگاه‌داده امری ضروری به نظر می‌رسد. سیستم تشخیص نفوذ در پایگاه‌داده برای تشخیص حملات درجایی که فقط سازوکارهای سنتی پیشگیری هستند و یا ممکن است به‌راحتی دور بخورند می‌توانند مورداستفاده قرار گیرد. تشخیص نفوذ به دو روش مبتنی بر امضا و تشخیص ناهنجاری انجام می‌گیرد.

چالش اصلی در روش‌های مبتنی بر امضا این است که به دنبال الگوهای خاصی در رشته یا درخواست ارسالی به سمت پایگاه‌داده است و همیشه درصد قابل‌توجهی از حملات به‌صورت ناشناخته باقی می‌ماند. روش تشخیص ناهنجاری به دنبال رفتارهای غیرطبیعی در ترافیک ارسالی هست که در این روش هم هشدارهای مثبت کاذب و منفی کاذبی وجود دارند که برای به‌حداقل‌رساندن آن‌ها بایستی به دنبال انجام کارهای مؤثرتری بود [۱].

استفاده از یادگیری ماشین[1] برای تشخیص نفوذ تبدیل به یک منطق گسترده در تحقیقات اخیر شده است، فناوری‌های پیشرفته مانند یادگیری ماشین را می‌توان به سیستم‌های تشخیص نفوذ برای مقابله با مشکلات جدید و با تهدیدات فعلی اعمال کرد. باتوجه‌به ماهیت الگوریتم‌های یادگیری ماشین، این الگوریتم‌ها با کار بیشتری که در مرحله پیش‌پردازش انجام می‌شود، می‌توانند خودشان را بر اساس حملات و ترافیک مخرب بهبود بخشند که منجر به شناسایی اشکال جدید حملات شده و همچنین باعث افزایش سطح دقت در شناسایی حملات که قبلاً دیده نشده‌اند می‌شوند [۲].

گسترش استفاده از اینترنت و وب در امور روزمره، باعث افزایش سرویس‌های وب در زمینه‌های مختلف شده است. این سرویس‌ها عموماً با داده‌هایی سروکار دارند که برای سازمان‌دهی و سهولت بازیابی آن‌ها در پایگاه‌داده‌ها ذخیره می‌شوند. ارتباط با پایگاه‌داده‌ها به‌وسیله زبانی به نام زبان جستجوی ساختاریافته[2] انجام می‌شود. زبان جستجوی ساختاریافته برای ذخیره، تغییر و بازیابی اطلاعات از جداول موجود در پایگاه‌داده‌ها بکار می‌رود. ازآنجاکه زبان جستجوی ساختاریافته با اطلاعات موجود در وبگاه‌ها ارتباط دارد، موردتوجه و سوءاستفاده نفوذگرها قرار می‌گیرد و باعث به‌وجودآمدن تهدیدی به نام تزریق کد می‌شود. تزریق کد زمانی اتفاق می‌افتد که برنامه کاربردی تحت وب که به زبان‌هایی مانند صفحات وب پویا[3]، جاوا[4]، زبان فرامتن[5] و... نوشته‌شده است، بر روی ورودی‌های کاربر، اعتبارسنجی[6] مناسبی نداشته باشد.

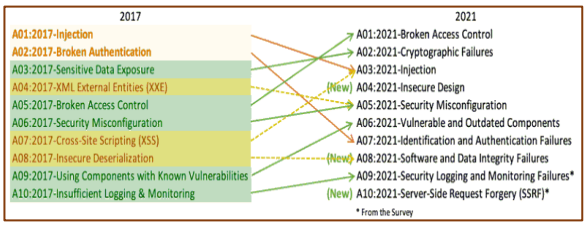
عدم اعتبارسنجی ورودی‌هایی که در تولید یک پرس‌وجو[7] بکار گرفته می‌شوند، باعث این آسیب‌پذیری می‌شود. ازآنجاکه این پرس‌وجوها توسط پایگاه‌های داده برای ذخیره و بازیابی اطلاعات اجرا خواهند شد، نفوذگرها با اعمال ورودی‌های ماهرانه، سعی در اجرای نادرست برنامه و تحقق حملاتی مانند دورزدن سازوکارهای احراز اصالت، افشا و یا تغییر اطلاعات و ازکارانداختن سرویس می‌کنند. در اینجا، حملات بانک اطلاعاتی اشاره به رفتارهای مخربی است که ممکن است در درخواست‌های صادرشده به سمت پایگاه‌داده سیستم ظاهر شوند.

گسترش اینترنت باعث شده است که درخواست‌های وب بخشی از زندگی روزمره باشد. درنتیجه، تعداد حوادثی که از آسیب‌پذیری‌های وب استفاده می‌کنند افزایش می‌یابد. درصد زیادی از این موارد، حملات تزریق کد است که یک تهدید جدی امنیتی برای پایگاه‌های داده با اطلاعات بالقوه حساس است؛ بنابراین، تحقیقات زیادی برای شناسایی و جلوگیری از این حملات انجام‌شده است و منجر به کاهش حملات تزریق کد شده است. بااین‌حال، هنوز روش‌هایی برای جلوگیری از آن‌ها وجود دارد و این روش‌ها در برنامه‌های واقعی وب بسیار پیچیده است[۳].

استخراج ویژگی[8] فرایندی است که با داشتن دانش تخصصی در داده‌ها باهدف استخراج و طبقه‌بندی الگوهای خاص به کار می‌رود. تشخیص و طبقه‌بندی حملات تزریق کد از یک‌سری حملات با الگوهای مشخص پیروی می‌کند. بااین‌حال، نحو و زبان جستجوی ساختاریافته اجازه می‌دهد تا تنوع زیاد و ترکیبی از کلمات که تشخیص آن با استفاده از امضاهای استاتیک سنتی دشوار است را آسان کند. یک رویکرد معقول برای مقابله با این مشکل استفاده از روش تشخیص غیرعادی و بهره‌برداری از ویژگی‌های بارز حملات تزریق کد از طریق استخراج ویژگی‌های جدید است [۴].

[1] Machine Learning
[2] Structured Query Language (SQL)
[3] Active Server Page
[4] Just Another Virtual Accelerator(Java)
[5] Hypertext Preprocessor(PHP)
[6] Input Validation
[7] Query
[8] Feature Extraction

مسئله آسیب‌پذیری برنامه‌های تحت وب و کاربردی در برابر حملات تزریق کد، گستردگی این نوع حمله و مرغوبیت و سادگی استفاده از آن است و در ورودی‌های ارسالی از سمت سرویس‌گیرنده‌ها[1] به سرویس‌دهنده[2] است و بانک اطلاعاتی روی سرویس‌دهنده‌ها درگیر این نوع حمله هستند. همچنین در چند سال اخیر از ۱۰ حملات برتر در این حوزه، جزء رده‌های اول تا سوم از نوع حملات پایگاه داده‌ای بوده و هیچ‌گاه به طور کامل کنار نرفته. شکل (۱) وضعیت حملات سمت وب را از سال ۲۰۱۷ تا ۲۰۲۱ نشان می‌دهد.



شکل (۱). وضعیت حملات تزریق کد در چند سال اخیر [۵]

صاحبان داده در سرویس‌دهنده‌های پایگاه داده‌ای متضرر این نوع حملات هستند. مخاطبین مسئله مراکز علمی و تحقیقاتی، دانشگاهی و همچنین شرکت‌های تولیدکننده سیستم‌های تشخیص نفوذ بانک‌های اطلاعاتی سخت‌افزاری و نرم‌افزاری هستند. اگر مسئله حل نشود، باتوجه‌به اینکه در چند سال اخیر جزء دغدغه‌های مدیران سازمان‌ها و برنامه‌نویسان است. این آسیب‌پذیری هنوز هم در اکثر برنامه‌های کاربردی وجود دارد و به‌عنوان یک چالش باقی خواهد ماند. هنوز هم دیواره‌های آتش[3] در برابر این حمله دور می‌خورند و از فیلترهای سمت سرویس‌گیرنده یا سرویس‌دهنده رد می‌شوند و راهکار جامعی به‌طورقطع وجود ندارد. هدف، جلوگیری از حملات تزریق کد و حفظ اطلاعات حساس روی پایگاه‌داده‌ها در سرویس‌دهنده‌ها است. این مسئله با شناخت کامل از انواع تزریق‌ها، شناخت قالب ورودی‌های ارسالی به سرویس‌دهنده‌های وب و بازرسی دقیق دستورات ورودی و تجزیه آن و استخراج ویژگی‌های مؤثر در درخواست‌های ارسالی به پایگاه‌داده حل می‌گردد.

سیستم تشخیص نفوذ وظیفه شناسایی و تشخیص هرگونه استفاده غیرمجاز به سیستم، سوءاستفاده یا آسیب‌رسانی توسط هر دودسته کاربران داخلی و خارجی را بر عهده‌دارند. تشخیص و جلوگیری از نفوذ امروزه به‌عنوان یکی از سازوکارهای اصلی در برآوردن امنیت شبکه‌ها و سیستم‌های رایانه‌ای مطرح است و عموماً در کنار دیواره‌های آتش و به‌صورت مکمل امنیتی برای آن‌ها مورداستفاده قرار می‌گیرند.

[1] Clients
[2] Servers
[3] Firewalls

سامانه‌های تشخیص نفوذ به‌صورت سامانه‌های نرم‌افزاری و سخت‌افزاری ایجادشده و هرکدام مزایا و معایب خاص خود رادارند. سرعت و دقت از مزایای سیستم‌های سخت‌افزاری است و عدم شکست امنیتی آن‌ها توسط نفوذگران، قابلیت دیگر این‌گونه سیستم‌ها است؛ اما استفاده آسان از نرم‌افزار، قابلیت سازگاری در شرایط نرم‌افزاری و تفاوت سیستم‌های عامل مختلف، عمومیت بیشتری را به سامانه‌های نرم‌افزاری می‌دهد و عموماً این‌گونه سیستم‌ها انتخاب مناسب‌تری هستند. به‌طورکلی سه عملکرد اصلی سیستم تشخیص نفوذ عبارت است از: نظارت و ارزیابی، کشف و واکنش. بر همین اساس هر سیستم تشخیص نفوذ را می‌توان بر اساس روش‌های تشخیص نفوذ، معماری و انواع پاسخ به نفوذ دسته‌بندی کرد. قسمت دیگری نیز به نام سیستم جلوگیری از نفوذ وجود دارد که داخل هسته آن می‌توان سیستم تشخیص نفوذ را یافت. درواقع سیستم تشخیص نفوذ تنها ورود افراد غیرقانونی به شبکه را شناسایی می‌کنند؛ اما سیستم‌های جلوگیری از نفوذ جلوی ورودهای غیرقانونی به شبکه را می‌گیرند [۶].

۲- تاریخچه

تزریق کد برای اولین‌بار در سال ۱۹۹۸ توسط فردی به نام رین فارست پاپی[4] در مجله اینترنتی به نام فارک منتشر شد. این مقاله بانام آسیب‌پذیری فناوری وب[5] منتشر شد و به چندین سناریوی حمله به پایگاه‌داده اشاره کرده بود. در این مقاله در مورد برنامه‌های پزشکی و صفحات وب پویا بحث شده بود که بر روی سرویس‌دهنده وب در حال اجرا هستند و از پایگاه‌داده مایکروسافت استفاده می‌کنند. در آن زمان نام تزریق کد شناخته‌شده نبود. در سال ۱۹۹۹ مدتی بعدازاین مقاله، مقاله‌ی دیگری در همین مورد توسط آلیر[6] ارائه شد. چندی بعد بازهم آقای پاپی به انتشار مقاله‌ای دیگر در این زمینه پرداخت که در آن به نحوه سوءاستفاده از این حمله پرداخته بود. در این مقاله آقای پاپی توانسته بود به وبگاه پکت استروم[7] که یک وبگاه مرتبط بانفوذ گرها است به‌وسیله رخنه در یک برنامه پرل[8] موسوم به رشته‌ها نفوذ کند و فرمان‌های دلخواه خود را بر روی پایگاه‌داده اجرا کرده و خود را به‌عنوان مدیر سیستم معرفی نماید. در سال ۲۰۰۱ فردی بانام دیوید لیچفید[9] مقاله دیگری ارائه داد و پس از آن دیگران سعی در کامل‌کردن انواع حملات کردند. تزریق کد، 10% از کل جرائم رایانه‌ای را از سال ۲۰۰۲

[4] Rain Forest Puppy
[5] NT Web Technology Vulnerabilities
[6] Allaire
[7] packetstorm
[8] perl
[9] David Litchfield

تابه‌حال، تشکیل می‌دهد ۶. مؤسسه ملی فناوری و استانداردها[1] در پایگاه‌داده آسیب‌پذیری‌های خود، از ابتدای سال ۲۰۰۷ تا ماه می سال ۲۰۰۹ میلادی، بالغبر ۲۰۰۰ مورد آسیب‌پذیری تزریق کد را در برنامه‌های مهم کاربردی تحت وب گزارش کرده است. در بررسی به‌عمل‌آمده توسط موتورهای جستجو نشان می‌دهد که حدود 11% از وبگاه‌های مورد جستجو، مستعد بروز حمله تزریق کد هستند. در مقابله با این آسیب‌پذیری، روش‌هایی ارائه‌شده که هرکدام دارای نقاط ضعف و قوت هستند. ازجمله نقاط ضعف عمده در این روش‌ها، وابستگی آن‌ها به فناوری توسعه وب است. اکثر این روش‌ها نیازمند تغییر رمز منبع برنامه هستند. از سال ۲۰۰۳، تزریق کد در ۱۰ فهرست بالا از آسیب‌پذیری رایج[2] باقی‌مانده است. ۳۲۶۰ آسیب‌پذیری بین سال‌های ۲۰۰۳ و ۲۰۱۱. در سال ۲۰۱۲ یک نماینده از شرکت بارکی کارد[3] ادعا کرد که ۹۷ درصد از نقض داده درنتیجه تزریق کد است. در اواخر سال ۲۰۱۱ و اوایل سال ۲۰۱۲ از این طریق، یعنی تنها در یک ماه، بیش از یک میلیون صفحات وب توسط حمله تزریق کد قرار گرفتند. سال ۲۰۰۸ شاهد نابسامانی اقتصادی باورنکردنی به‌عنوان یک نتیجه از تزریق کد بوده است. حتی وبگاه رسمی سازمان ملل متحد در سال ۲۰۱۰ قربانی یک حمله تزریق کد شد و از کار افتاد. در سال ۲۰۱۱ ایالات متحده آمریکا وزارت امنیت داخلی، مؤسسه میتر[4] و سنز[5] تزریق کد را به‌عنوان خطرناک‌ترین آسیب‌پذیری امنیتی در صدر قرار داده‌اند. پس چرا، بیش از ۲۳ سال، هنوز هم آسیب‌پذیری شماره یک و به‌ظاهر غیر قابل جلوگیری است [۷].

سیستم تشخیص نفوذ مبتنی بر ناهنجاری[6] یک سیستم تشخیص نفوذ برای تشخیص نفوذ شبکه و رایانه و سوءاستفاده از طریق نظارت بر فعالیت سیستم و طبقه‌بندی آن به‌عنوان عادی یا غیرعادی است. این طبقه‌بندی به‌جای الگوها یا امضاها بر اساس موارد اکتشافی یا موارد بر اساس الگوها است و تلاش می‌کند تا هر نوع سوءاستفاده‌ای را که خارج از عملکرد عادی سیستم است، شناسایی کند. این برخلاف سیستم‌های مبتنی بر امضا است که فقط می‌توانند حملاتی را شناسایی کنند که قبلاً برای آن‌ها امضا ایجادشده است [۸].

در سال ۲۰۲۴ الگوریتم‌های مختلف یادگیری ماشین برای شناسایی حملات تزریق کد مورد آزمایش قرار گرفتند. در بخش پیش‌پردازش داده‌ها، استخراج ویژگی با استفاده از فن‌های پردازش زبان طبیعی انجام شد. درحالی‌که ارتباط عبارات با یکدیگر با روش آماری[7] در سطح کلمات محاسبه شد، همچنین جستجوی عبارت نیز انجام شد. در این مطالعه، فرایند طبقه‌بندی با استفاده از شبکه عصبی پیچیده[8]، ماشین بردار پشتیبان[9]، بیزهای ساده[10] و مدل آماری رگرسیون متغیرهای وابسته[11] از روش‌های یادگیری ماشین انجام شد. در پایان مطالعه تجربی، مشخص شد که شبکه عصبی پیچیده، مدلی است که حملات تزریق کد را با بیشترین موفقیت شناسایی می‌کند [۹].

در سال ۲۰۲۴ از الگوریتم بهینه‌سازی حداقل متوالی[12] برای ارائه یک روش منحصربه‌فرد برای شناسایی حملات تزریق کد درداده نای ترافیک شبکه استفاده‌شده. این مطالعه یک استراتژی منحصربه‌فرد را پیشنهاد می‌کند که از یادگیری ماشین برای حل نیازهای حیاتی برای روش‌های تشخیص کارآمد و مؤثر استفاده می‌کند. این مطالعه به طور خاص بر استفاده از فن بهینه‌سازی حداقل متوالی برای شناسایی و کاهش تهدیدات تزریق کد با استفاده از داده‌های ترافیک شبکه متمرکز است. توالی تماس‌ها بین میزبان‌ها یا داده‌های جریان شبکه، اطلاعات زیادی را برای شناسایی الگوهای غیرمعمول حاکی از فعالیت حمله فراهم می‌کند [۱۰].

۳- روش‌های تشخیص تزریق کد

به‌منظور بهبود امنیت سیستم‌های اطلاعاتی روزمره با آسیب‌پذیری‌ها و خطرات تهدید می‌شوند و همه نیاز به تجزیه‌وتحلیل دارند تا برای مقابله با آن‌ها اقدامات متناسب با آن صورت پذیرد [۱۱]. یکی از چالش‌های اصلی شناسایی حملات به‌منظور جلوگیری ازاین‌گونه حملات است. به‌منظور مقابله با تهدیدات روبه‌رشد و چشمگیر، سیستم‌های تشخیص نفوذ و پیشگیری از نفوذ و همچنین دیواره‌های آتش کاربردی وب[13] همراه با سایر لایه‌های دفاع استفاده می‌شود.

۳-۱- تطبیق الگو:

الگوریتم جستجوی رشته بویر مور[14] معمولاً برای جستجوی مقادیر زیاد داده‌ها در یک‌زمان کوتاه ازجمله، جستجوی الگوهای ویروس و پایگاه‌های داده استفاده می‌شود. این الگوریتم یکی از سریع‌ترین الگوریتم‌های جستجوی رشته است زیرا رشته‌ای را برای الگوریتم جستجو می‌کند، اما نه هر یک از رشته‌ها. با افزایش طول الگو، الگوریتم سریع‌تر اجرا می‌شود. این الگوریتم متن را برای مقایسه و کلمه کلیدی مرتب می‌کند تا کلمه کلیدی

[1] National Institute of Standards and Technology
[2] Common Vulnerabilities and Exposures
[3] Barclaycard
[4] MIT Research Establishment
[5] SysAdmin, Audit, Network and Security
[6] Anomaly-based intrusion detection system

[7] Term Frequency – Inverse Document Frequency(TF-IDF)
[8] Convolutional Neural Network(CNN)
[9] Support Vector Machine(SVM)
[10] Naive Bayes
[11] Logistic Regression
[12] Sequential Minimal Optimization(SMO)
[13] Web Application Firewall(WAF)
[14] Boyer-Moore

را بتوان از چپ به راست در طول متن بررسی کرد. چک با حرف آخر کلمه کلیدی شروع و پایان آن با اولین آغاز می‌شود. برای تعیین تعداد نوبت موردنیاز برای کشیدن پنجره روی متن، پسوند خوب و پسوند بد مورداستفاده قرار می‌گیرد [۱۲].

باوجود اثبات مؤثربودن عدم تولید نادرست‌های نادرست، فن‌های تطبیق الگو را می‌توان با استفاده از روش‌هایی مانند کدگذاری جایگزین برای تغییر نحوه نمایش داده‌های زبان جستجوی ساختاریافته دور زد [۱۳]. همانند روش‌های تحلیل استاتیک که قبلاً مورد بحث قرار گرفته است، اگر سیستم در مورد الگوی یا رشته چیزی نداند، بنابراین نمی‌تواند هر حمله جدیدی را شناسایی کند، به این معنی که باید به طور مداوم بروز شده تا به‌عنوان حملات جدید کشف شود.

۳-۲- یادگیری ماشین:

تام میچل[1] یک تعریف مختصر برای توضیح روش شناختن یادگیری ماشین را ارائه می‌دهد که عبارت است از: «زمینه یادگیری ماشین مربوط به مسئله چگونگی ساخت برنامه‌های رایانه‌ای است که به طور خودکار باتجربه بهبود می‌یابد [۱۴].» بر اساس تعریف تام میچل، جیسون براونلی[2] اخیراً آن را برای تعریف مدرن‌تر و مناسب‌تر مطرح کرده است: «یادگیری ماشین، آموزش یک مدل از اطلاعات است که تصمیم‌گیری را در برابر یک معیار عملکرد بهینه می‌کند [۱۵].» این بدان معنی است که الگوریتم را با نمونه‌های آموزشی برای ایجاد یک مدل ارائه می‌دهد، این مدل از تجربیات به‌دست‌آمده در هنگام پردازش نمونه‌های آموزشی ساخته‌شده است. سپس الگوریتم می‌تواند بر اساس ورودی‌های فعلی و پیش‌بینی‌های مربوط به ورودی‌هایی که در آینده دنبال می‌کنند، تصمیم‌گیری کند، برای یک الگوریتم که تصمیم‌گیری خوبی را انجام دهد، به مرحله آموزش خوب متکی است و به یک مجموعه داده‌ای باکیفیت بسیار بالا موردنیاز است.

۳-۳- تشخیص ناهنجاری:

روش‌های مبتنی بر تشخیص ناهنجاری[3] معمولاً مدل‌هایی را ایجاد می‌کنند که الگوی طبیعی و غیرطبیعی در ترافیک را تعیین می‌کنند. از الگوریتم‌های آموزش بدون ناظر در این مدل‌ها استفاده می‌شود. تشخیص ناهنجاری تلاش می‌کند رفتار قابل‌قبول یا معمولی یک کاربر را اندازه‌گیری کند و سپس انتظارات بالقوه را نشان می‌دهد [۱۶].

هنگامی‌که از روش تشخیص ناهنجاری استفاده می‌شود، ممکن است هشدارهای منفی کاذب وجود داشته باشد، چراکه همه ترافیک‌های غیرعادی ناسازگار نیستند، در هر دو روش، ترافیک مخرب می‌تواند به‌عنوان ترافیک نرمال عبور کند که منجر به مثبت کاذب شود [۱۷].

۳-۴- مبتنی بر سوءاستفاده:

روش‌های مبتنی بر سوءاستفاده[4] باتوجه‌به تولید هشدارهای مثبت کاذب و منفی کاذب درروش تشخیص ناهنجاری، سیستم تشخیص مبتنی بر سوءاستفاده بیشتر برای تشخیص نفوذ قابل‌توجه است.

این روش‌های تشخیص بیشتر در سیستم‌های تجاری تشخیص نفوذ استفاده می‌شود. این روش بسته‌های گرفته‌شده و پایگاه داده‌ای از حملات شناخته‌شده را می‌گیرد و بسته را در برابر تمام قوانین حمله در پایگاه‌داده مقایسه می‌کند [۱۷].

اغلب سیستم‌های تشخیص و جلوگیری از نفوذ تجاری و سیستم‌های ضدویروس از سیستم‌های تشخیص مبتنی بر سوءاستفاده می‌کنند [۱۸]. به دلیل این است که سیستم‌های مبتنی بر سوءاستفاده دقت بالایی را ارائه می‌دهند، زیرا که آن‌ها در حملات شناخته‌شده، مثبت کاذب و منفی کاذب را تولید نمی‌کنند.

از الگوریتم‌های آموزش با ناظر در این مدل‌ها استفاده می‌شود. این روش‌های تشخیص بیشتر در سیستم‌های تجاری تشخیص نفوذ استفاده می‌شود. این روش بسته‌های گرفته‌شده و پایگاه داده‌ای از حملات شناخته‌شده را می‌گیرد و بسته را در برابر تمام قوانین حمله در پایگاه‌داده مقایسه می‌کند [۱۷]. اغلب سیستم‌های تشخیص و جلوگیری از نفوذ تجاری و سیستم‌های ضدویروس از سیستم‌های تشخیص مبتنی بر سوءاستفاده، بهره‌برداری می‌کنند [۱۸]. به دلیل این است که سیستم‌های مبتنی بر سوءاستفاده دقت بالایی را ارائه می‌دهند، زیرا که آن‌ها در حملات شناخته‌شده، مثبت کاذب و منفی کاذب را تولید نمی‌کنند.

۴- معیارهای مقایسه

طیف وسیعی از معیارها وجود دارد که می‌تواند برای ارزیابی عملکرد هر الگوریتم مورداستفاده قرار گیرد، اما بعضی از آن‌ها می‌توانند بسته به پارامترهای آزمایش متفاوت باشند، به‌عنوان‌مثال ممکن است به مجموعه‌داده بستگی داشته باشد و بعضی ممکن است اطلاعات مربوط به نوع حمله را ارائه دهند [۱۹]. بااین‌حال، بر اساس تحقیق مورد بررسی، محبوب‌ترین معیارهای مورداستفاده برای ارزیابی الگوریتم‌های یادگیری ماشین شامل دقت[5]، صحت[6]، مثبت درست[7]، مثبت کاذب[8]، منفی

[1] Tom Mitchell
[2] Jason Brownlee
[3] Anomaly Based
[4] Misuse Based
[5] Accuracy
[6] Precision
[7] True Positive (TP)

درست[۱]، منفی کاذب[۲] هستند. معیارهای دیگر مانند یادآوری یا بازخوانی[۳]، کمیت وزن‌دار اف[۴] و کلاس[۵] نیز می‌توانند برای مقایسه مورداستفاده قرار بگیرند. با استفاده از موارد فوق جدول (۱) بنام ماتریس درهم‌ریختگی یا اغتشاش[۶] تولید می‌گردد[۲۰].

جدول (۱). ماتریس درهم‌ریختگی یا اغتشاش

Confusion Matrix		Detection Classification مقادیر پیش‌بینی‌شده	
		Positive پیش‌بینی مثبت	Negeative پیش‌بینی منفی
Real Lable مقادیر واقعی	Positive (True) مثبت واقعی	TP درست مثبت	FN نادرست منفی
	Negeative (False) منفی واقعی	FP نادرست مثبت	TN درست منفی

نتایج ماتریس درهم‌ریختگی درواقع بیانگر آن است که یک مثبت درست نشان‌دهنده‌ی این است که درخواست‌های نرمال و درست به‌صورت صحیح دسته‌بندی‌شده است. یک منفی درست نشان‌دهنده‌ی این است که درخواست‌های غیرنرمال و نادرست به‌صورت صحیح دسته‌بندی‌شده است. منفی کاذب تشخیص نادرست ترافیک حمله به‌عنوان خوب، یعنی هنگامی‌که حمله باشد و درخواست‌ها غیرنرمال بوده؛ ولی به‌صورت درست طبقه‌بندی شوند، مثبت کاذب تشخیص نادرست ترافیک خوب به‌عنوان حمله، یعنی اگر حمله‌ای نباشد و درخواست نرمال بوده؛ ولی به‌عنوان حمله شناخته شود. میزان هشدارهای اشتباه از عوامل کلیدی برای مقایسه کارایی سیستم‌های تشخیص است [۲۱]. اما در مطالعات اخیر میشل و چن[۷] ادعا دارند که در حال حاضر محققان عملکرد سیستم‌ها را با استفاده از دو شاخص اصلی ازجمله نرخ مثبت کاذب[۸] و نرخ منفی واقعی[۹] اندازه‌گیری می‌کنند [۲۲]. این نرخ‌ها را می‌توان با استفاده از داده‌ها جمع‌آوری‌شده از ماتریس درهم‌ریختگی که:

$$FPR = FP / (TN + FP)$$

$$TPR = TP / (TP + FP)$$

محاسبه کرد. برای آزمایش‌های انجام‌شده در این مطالعه، این دو معیار به‌عنوان روش اصلی برای ارزیابی و به دلیل استفاده گسترده آن‌ها در پژوهش‌های فعلی مورداستفاده قرار می‌گیرد.

[8] False Positive (FP)
[1] True Negative (TN)
[2] False Negative (FN)
[3] Recall
[4] F-Measure
[5] Class
[6] Confusion Matrix
[7] Michelle and Chen
[8] False Positive Rates
[9] True Negative Rates

بااین‌وجود نتایج برای معیارهای دیگر نیز در صورت لزوم برای ارزیابی دقیق استفاده می‌شود.

۵- مجموعه‌داده:

تمام تحقیقات اخیر که در جستجوی تشخیص تزریق کد با استفاده از یادگیری ماشین انجام‌شده است، هرکدام مجموعه‌داده‌های خود را برای آزمایش الگوریتم‌های یادگیری ماشین مورداستفاده قرار داده‌اند، زیرا هرکدام از مجموعه‌داده‌ها به طور خاص که برای آزمایش‌های آن‌ها ساخته‌شده‌اند در دسترس عموم نیستند. به‌عنوان یک نتیجه از این تحقیق، مجموعه‌داده‌های استانداردشده برای ارزیابی الگوریتم‌های یادگیری ماشین برای تشخیص حملات تزریق کد تعریف‌نشده است. فهرستی از مشکلات موجود در هنگام دستیابی به مجموعه‌داده‌ها برای آزمودن سیستم تشخیص نفوذ وجود دارد:

- مجموعه‌داده‌ها به‌اندازه کافی برچسب‌گذاری نشده‌اند.
- مجموعه‌داده‌هایی که به طور گسترده استفاده می‌شود شامل ترافیک وب نیست.
- مقدار زیادی از مجموعه‌داده‌های مورداستفاده در پژوهش‌های مدرن به‌صورت عمومی در دسترس نیست یا دستیابی به آن‌ها دشوار است.
- مجموعه‌داده‌های معمول تا به امروز به‌روزرسانی نشده‌اند. (کی دی دی کاپ ۱۹۹۹ [۱۰])
- مقدار متعادلی از درخواست‌های نرمال و مخرب در آن‌ها نیست.

جدول (۲) مجموعه‌داده‌هایی است که به طور گسترده در تحقیقات اخیر مورداستفاده قرار گرفته است.

جدول (۲) مشخصات مجموعه‌داده‌های تحقیقات اخیر [۲۳]

Dataset	Public access	HTTP traffic	Labeled	Two classes	Modern attacks	Not anonymized
UNB ISC	✗	✓	✓	✓	✓	✓
ECML/PKDD	✗	✓	✓	✓	✓	✗
LBNL	✓	✗	✓	✓	✗	✗
DEFCON	✓	✓	✓	✗	✓	✓
DARPA 98/99	✓	✓	✓	✓	✗	✓
Captured traffic	✓	✓	✗	✗	✓	✓
CSIC	✓	✓	✓	✓	✓	✓

در جدول (۲) از مجموعه‌داده ردیف دوم چندین بار برای ارزیابی الگوریتم‌های یادگیری ماشین در تشخیص حملات وب در تحقیقات اخیر استفاده‌شده است [۲۰] [۲۴] [۲۵] [۲۳]. با نگاهی به این تحقیقات و نتایج مختلف، در مورد استخراج و انتخاب ویژگی، انتخاب الگوریتم و نتایج، روند قابل‌توجهی وجود دارد.

این مجموعه‌داده برای رقابت‌های داده‌کاوی در ۱۸ اجلاس اروپایی از یادگیری ماشین برگزارشده، طراحی‌شده بود.

[10] KDD CUP1999 Dataset

مجموعه‌داده‌های درمجموع حدود ۵۰۰۰۰ درخواست دارد که ۷۰ درصد از درخواست‌ها درست و ۳۰ درصد آن ترافیک مخرب بود و شامل انواع حملات[1] تزریق کد است [۲۵].

شکل (۲) اولین نمونه مجموعه‌داده ردیف دوم جدول (۲) که ساختار زبان نشانه‌گذاری گسترش‌پذیر[2] دارد و در این تحقیق از آن استفاده‌شده است، را نشان می‌دهد.



شکل (۲). نمونه‌ای از مجموعه‌داده ECML/PKDD

دومین مجموعه‌داده انتخابی بعدی، ردیف آخر جدول (۲) در سال ۲۰۱۰ بوده که قالب آن به‌صورت متنی بوده، اگرچه مجموعه‌داده به‌درستی برچسب‌گذاری نشده است و شامل حملات تزریق کد برای اهداف آموزشی به‌خوبی مجموعه‌داده اول نبوده، ولی باتوجه‌به کار خوبی که در مرحله پیش‌پردازش بر روی آن انجام‌شده و باتوجه‌به اینکه ویژگی‌های استخراج‌شده دارای کیفیت بالایی بوده، نتایج حاصله بسیار خوبی را نسبت به تحقیقات قبلی به دست آوردیم. این مجموعه‌داده شامل ۳۶۰۰۰ درخواست نرمال و ۲۴۶۶۸ درخواست مخرب است.

مجموعه‌داده دوم مطابق با تمام الزامات تشخیص حمله برنامه‌های تحت وب است که در دیواره‌های آتش تحت وب مورداستفاده قرار گرفته است. مجموعه‌داده نیز با الزامات این تحقیق مطابقت دارد، زیرا شامل حملات تزریق کد، ترافیک طبیعی و حملات سرقت اطلاعات کاربران[3] است.

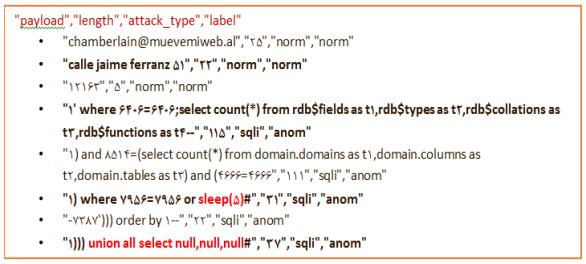
شکل (۳) نمونه‌ای از یکی از درخواست‌ها از مجموعه‌داده‌های دوم را نشان می‌دهد. در مقاله سال ۲۰۱۹ این مجموعه‌داده به‌عنوان یک مجموعه‌داده مدرن نام‌برده شده است [۲۶].



شکل (۳). نمونه‌ای از مجموعه‌داده CSIC2010

[1] Types of attacks (SQLi, LDAP, XPATH Injection attacks, XSS, SSI attacks, Path Traversal and Command Execution)
[2] Extensible Markup Language(XML)
[3] Cross Site Scripting(XSS)

مجموعه‌داده آخر[4] مربوط به سال ۲۰۱۶ است که این مجموعه‌داده نیز یکی از مجموعه‌داده‌های مفید است که با ابزار انمپ[5] ساخته‌شده است و البته سایر پارامترهای مربوط به درخواست‌های وب را ندارد و فقط پیلود[6] حملات را دارد. قالب آن‌هم به‌صورت متنی هست. این مجموعه‌داده در کل ۲۰۱۰۵ سطر، شامل ۱۲۸۷۰ درخواست نرمال و ۷۲۳۵ درخواست مخرب است. شکل (۴)



"payload","length","attack_type","label"
- "chamberlain@muevemiweb.al","۲۵","norm","norm"
- "calle jaime ferranz ۵۱","۲۲","norm","norm"
- "۱۲۱۶۳","۵","norm","norm"
- "\' where ۶۴۰۶=۶۴۰۶;select count(*) from rdb$fields as t۱,rdb$types as t۲,rdb$collations as t۳,rdb$functions as t۴--","۱۱۵","sqli","anom"
- "۱) and ۸۵۱۴=(select count(*) from domain.domains as t۱,domain.columns as t۲,domain.tables as t۳) and (۴۶۶۶=۴۶۶۶","۱۱۱","sqli","anom"
- "۱) where ۷۹۵۶=۷۹۵۶ or sleep(۵)#","۲۱","sqli","anom"
- "-۷۲۸۷'))) order by ۱۰--","۲۲","sqli","anom"
- "۱)))) union all select null,null,null#","۲۷","sqli","anom"

شکل (۴). نمونه‌ای از مجموعه‌داده HttpParamsDataset

۶- استخراج ویژگی

الگوریتم یادگیری ماشین برای تولید قوانین، جهت فرایند طبقه‌بندی تشخیص است و ویژگی‌ها باید از مجموعه‌داده ورودی در مرحله پیش‌پردازش استخراج شوند. این مرحله جایی است که پرس‌وجوها به آیتم‌های قابل‌توجهی از رشته‌های منظم که رخ می‌دهند در داده‌های اصلی شکسته می‌شوند [۲۷].

جدول (۳) روش‌های استخراج ویژگی که بیشترین استفاده رادارند، به روش جداسازی خالی[7] و توکنایز[8] شناخته می‌شود [۲۸]. جداسازی خالی بدین معنی است که کلمات یا اصطلاح از مجموعه‌داده‌ها بر اساس فضاهای خالی استخراج می‌شود و تعداد وقوع هر اصطلاح نیز محاسبه می‌شود [۲۸].

OR' 'a' = 'a' - '

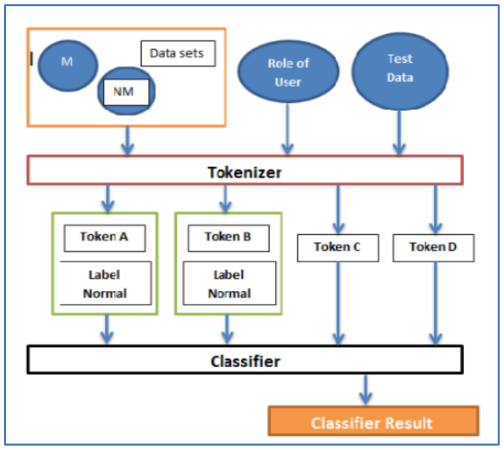
جدول (۳). روش جداسازی خالی [۲۸]

Name	Number of Terms
'	5
OR	1
a	2
=	1
;	1
--	1

یکی دیگر از روش استخراج ویژگی‌ها توکنایز است، این روش علامت‌هایی را برای ویژگی‌های مختلف از کد مخرب که برای توصیف اینکه کدام ویژگی انتخاب‌شده است استفاده می‌کنند.

[4] HttpParamsDataset 2016
[5] Nmap
[6] Pyload
[7] Blank separation
[8] Tokenising

توکنایز فرایند شکستن پرس‌وجو در عناصر معنی‌دار به نام نشانه است [۲۹]. سپس نشانه‌های انتخاب‌شده از پرس‌وجوها برای فرایند طبقه‌بندی استفاده می‌شود. شکل (۵)



شکل (۵). روش استخراج ویژگی (توکنایز) [۲۹]

استخراج مناسب این ویژگی‌ها نقش بسیار مهمی در چندین جنبه طراحی سیستم‌های تشخیص تزریق کد قوی و قابل‌اجرا دارد. منطق انتخاب این نوع ویژگی‌ها، توانایی آن برای شناسایی بیشتر انواع حملات[۱] تزریق کد است؛ مانند جایی که با پرس‌وجوهای مخرب در زبان جستجوی ساختاریافته دیده می‌شوند. دیگر ویژگی‌ها می‌توانند برای افزایش مقیاس‌پذیری سیستم برای شناسایی رمزهای مخرب جدیدتر استفاده شوند. علاوه بر این، با کاهش تعداد ویژگی‌ها (با حذف ویژگی‌های غیرمستقیم یا غیرضروری) و همچنین عملکرد سیستم قوانین انجمنی می‌تواند به‌خوبی عملکرد طبقه‌بندی قوانین تولیدشده بهبود یابد [۳۰].

استخراج ویژگی‌ها، مجموعه‌ای از ویژگی‌هایی است که در درخواست‌های رشته صفحات وب[۲] استخراج می‌شود. این ویژگی‌ها مربوط به مشاهدات مجموعه‌ای از زیررشته‌ها یا توکن‌ها در درخواست رشته صفحات وب است. معیارهای مورداستفاده در انتخاب این نشانه‌ها، استفاده از یک مجموعه ثابت از کلمات تعیین‌شده توسط یک متخصص امنیتی که در حملات وب سررشته دارد و یا به طور خودکار از یک مجموعه‌داده آموزشی به روش مدل‌سازی زبانی آماری انگرام[۳] استخراج می‌گردند [۳۱].

استخراج ویژگی بر اساس محاسبه ویژگی‌های استاتیک صفت رشته؛ مانند طول و تعداد کل رقم‌ها، حروف و شخصیت‌های خاص است. استخراج ویژگی در بخش‌های (متد، رشته صفحات وب، متن جستجو، هدر و بدنه) [۴] از درخواست رشته صفحات وب انجام‌شده است. کلیدواژه‌های حمله تزریق کد کلمات کلیدی محبوب در زبان جستجوی ساختاریافته است که به‌طورکلی برای انجام عملیات بر روی جداول در یک پایگاه‌داده استفاده می‌شود. این رویکرد، توکن‌ها (کلمات کلیدی) را از شخصیت‌های خاصی که مربوط به رمزهای وب مخرب و به‌عنوان ویژگی‌ها تشکیل‌شده است، استخراج می‌کند [۳۱].

استخراج ویژگی بر اساس محاسبه ویژگی‌های استاتیک صفت رشته؛ مانند طول و تعداد کل رقم‌ها، حروف و شخصیت‌های خاص است. ازآنجاکه ویژگی رشته در اکثر مجموعه‌داده‌های حمله وب مشکل‌ساز است، تمام ویژگی‌های رشته به‌عنوان مجموعه‌ای از ویژگی‌های عددی که نمایانگر وقایع شخصیت هستند، به شخصیت‌های یونی گرام[۵] تقسیم شدند [۳۲].

استخراج ویژگی در بخش‌های مختلف (متد، رشته صفحات وب، متن جستجو، هدر و بدنه) از رشته درخواست صفحات وب انجام‌شده است [۳۳].

مهندسی استخراج ویژگی شامل موارد زیر است [۳۴]:

تبدیل ویژگی[۶]: در مورد ایجاد ویژگی‌های جدید از ویژگی‌های موجود است.

تولید ویژگی[۷]: در مورد تولید ویژگی‌های جدید است که غالباً نتیجه تبدیل ویژگی‌ها نیستند. از الگوهای استخراج‌شده از داده‌ها می‌توان برای تولید ویژگی‌های جدید استفاده کرد. از اصطلاحات ساخت ویژگی[۸] و استخراج ویژگی برای تولید ویژگی استفاده می‌شود.

انتخاب ویژگی[۹]: انتخاب ویژگی‌ها مربوط به انتخاب یک مجموعه کوچک از ویژگی‌ها از یک مجموعه بزرگی از ویژگی‌ها است. انتخاب ویژگی همچنین ممکن است منجر به بهبود کیفیت درنتیجه الگوریتم‌های یادگیری ماشین شود.

مهندسی ویژگی خودکار[۱۰]: در مورد رویکردهای عمومی برای تولید خودکار تعداد زیادی از ویژگی‌ها و انتخاب زیرمجموعه مؤثر از ویژگی‌های تولیدشده است. مثل انگرام.

تجزیه تحلیل و ارزیابی ویژگی[۱۱]: درباره ارزیابی سودمندی ویژگی‌ها و یا مجموعه ویژگی‌ها است. گاهی اوقات به‌عنوان بخشی از انتخاب ویژگی‌ها هم نامیده می‌شود.

۷- انتخاب ویژگی

انتخاب ویژگی‌های مناسب جهت رسیدن به بهترین نتیجه در طبقه‌بندی داده از مباحث چالش‌برانگیز دهه‌های اخیر است.

[1] Types of attacks (tautologies ،union ،piggybacked ،illegal / logically incorrect ،alternate encodings and stored procedures)
[2] Uniform Resource Locator*(URL)*
[3] N-gram
[4] Sections (HTTP method, URL, query string, headers, body)

[5] Unigram
[6] Feature transformation
[7] Feature generation
[8] Feature construction
[9] Feature selection
[10] Automatic feature engineering
[11] Feature analysis and evaluation

اگرچه از دیدگاه نظریه، یادگیری از تعداد ویژگی‌های بیشتر باعث افزایش دقت پیش‌بینی می‌شود، اما شواهد عملی بیانگر آن است که این امر همواره صادق نیست؛ زیرا تمام ویژگی‌ها برای تشخیص برچسب کلاس داده‌ها، مهم نیستند و یا برخی از آن‌ها اساساً با برچسب داده‌ها بی‌ارتباط هستند. ازآنجاکه عامل‌های بسیاری ازجمله کیفیت در موفقیت یک الگوریتم یادگیری مؤثر است، اگر داده‌ها حاوی ویژگی‌ها و یا اطلاعات تکراری، نامربوط، نویزی و یا نامطمئن باشند، اخذ دانش از آن داده‌ها مشکل می‌گردد. علاوه بر این، وجود تعداد زیاد ویژگی‌ها موجب افزایش نامطلوب زمان اجرای الگوریتم یادگیری نیز می‌شوند [۳۵].

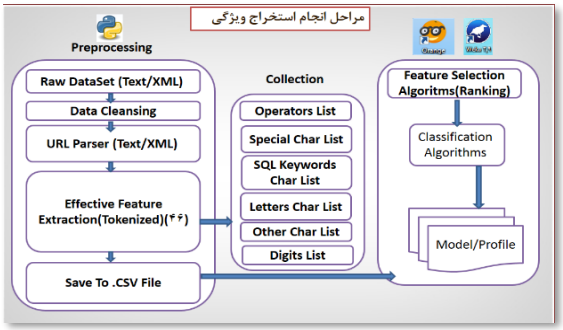
ازاین‌رو شیوه‌های کاهش ویژگی بسیار متداول هستند و شامل دودسته استخراج ویژگی (تبدیل ویژگی) و انتخاب ویژگی تقسیم‌بندی می‌شوند. دسته نخست شامل تحلیل مؤلفه‌های اصلی[1] و آنالیز تشخیصی خطی[2] و با تصویرکردن ویژگی‌های اصلی در فضایی کوچک‌تر، ویژگی‌های جدیدی تولید می‌کنند. بزرگ‌ترین مشکل شیوه‌های استخراج ویژگی این است که برای کاربران قابل‌فهم نیست. در مقابل، روش انتخاب ویژگی زیرمجموعه‌های از ویژگی‌های داده را به گونه‌های انتخاب می‌کند که جداپذیری داده‌ها حفظ شود (ویژگی‌های اولیه را تغییر نمی‌دهد) و بدین ترتیب موجب بهبود کیفیت داده‌ها نیز می‌گردد. به دلیل مزایای زیاد الگوریتم‌های انتخاب ویژگی ازجمله پایداری در برابر نویز[3] و افزایش دقت[4]، موردتوجه بسیاری از محققان قرار گرفته است و در حوزه‌هایی مثل دسته‌بندی متون، بازیابی تصویر، زیست‌شناسی و تشخیص نفوذ مورداستفاده قرار می‌گیرد [۳۵].

باتوجه‌به نحوه استفاده از الگوریتم یادگیری در پروسه‌ی انتخاب ویژگی، الگوریتم‌های انتخاب ویژگی به دودسته کلی روش‌های مبتنی بر فیلتر[5] و روش‌های مبتنی بربسته بندی رپر[6] تقسیم می‌شوند.

۸- مدل روش پیشنهادی

مجموعه‌داده‌ها که در دو قالب زبان نشانه‌گذاری گسترش‌پذیر و متنی قرار دارند، تعدادی از آیتم‌ها در نمونه‌های مختلف وجود دارد که هیچ مقدار داده‌ای ندارند، به این معنی که تمام ویژگی‌ها باید به‌درستی استخراج شود. برای مقابله با این موضوع یک اسکریپت نوشته شد که فایل زبان نشانه‌گذاری گسترش‌پذیر و متن اصلی را تجزیه‌وتحلیل می‌کند و یک فایل اکسل ایجاد می‌کند و هر درخواست را در یک ردیف جداگانه قرار می‌دهد. پایتون به‌عنوان بهترین زبان برنامه‌نویسی برای حل این مشکل انتخاب شد، زیرا دارای بسته‌هایی است که در خواندن و نوشتن فایل اکسل و تجزیه زبان نشانه‌گذاری گسترش‌پذیر و متن کمک زیادی می‌کند.

طبق شکل (۶) این مرحله شامل چندین گام است که در ابتدا مجموعه‌داده خام وارد می‌شود و سپس پاک‌سازی داده روی آن انجام می‌گیرد، در ادامه توسط یک پارسر[7] به زبان پایتون مجموعه‌داده بازشده و پس از آن موتور استخراج ویژگی‌ها شروع به کارکرده و ویژگی‌های مؤثر که در بیشتر در قسمت پیلود درخواست است و تعدادی هم در هدر[8]، مسیر[9] و دامین[10] است، استخراج می‌گردد که همان ویژگی‌های مؤثری هستند که در این تحقیق به آن‌ها رسیدیم و شامل ۴۶ ویژگی هستند که تعداد ۳۰ تا از آن‌ها در تحقیقات قبلی ذکرشده و تعداد ۱۶ عدد را هم به آن اضافه کرده‌ایم که نو و جدید هستند. در آخر همه ویژگی‌ها در یک فایل با قالب اکسل ذخیره می‌گردد و سپس به‌عنوان ورودی پیمانه یادگیری ماشین برای ارزیابی و تحلیل نتایج و ساخت مدل ارسال می‌گردد.



شکل (۶). مدل روش پیشنهادی

برای ارزیابی ویژگی‌های استخراج‌شده، با استفاده از هشت الگوریتم مربوط به انتخاب ویژگی ازجمله (ضریب جینی، کسب اطلاعات، نسبت بهره، رتبه‌بندی ویژگی، متغیر تصادفی و منطق فازی) [11] انجام‌شده. از روش بسته‌بندی رپر و کمینه افزونگی، بیشینه وابستگی[12] از سه مجموعه‌داده استفاده‌شده، بهترین زیرمجموعه (از ۱ تا ۲۰ ویژگی) از ۴۶ ویژگی که دارای بالاترین رتبه بودند را با استفاده از پیمانه رتبه‌بندی[13] در محیط داده‌کاوی اورنج[14] انتخاب کرده که برای ارزیابی از آن‌ها استفاده می‌شود.

[1] Principal Component Analysis(PCA)
[2] Linear Discriminant Analysis(LDA)
[3] Over-Fitting
[4] Increased Accuracy
[5] Filter
[6] Wrapper

[7] Parser
[8] Header
[9] Path
[10] Domain
[11] (GiniIndex-InformationGain-GainRatio-ReliefF-Chi2-CFS)
[12] Minimum Redundancy Maximum Relevance(mRMR)
[13] Ranking
[14] Orange

در ابتدا بر روی ۴۶ ویژگی مؤثر مربوط به مجموعه‌داده اول، آزمایش‌های خود را توسط الگوریتم‌های مختلف یادگیری ماشین در محیط وکا[1] بکار گرفتیم و همچنین از الگوریتم انتخاب ویژگی کمینه افزونگی، بیشینه وابستگی و منطق فازی جهت انتخاب زیرمجموعه‌ای از ویژگی‌ها استفاده کرده و نتایج زیر را در جدول (۴) و جدول (۵) به دست آوردیم.

جدول (۴). نتایج پیش‌بینی بر روی ۴۶ ویژگی مجموعه‌داده ECML/PKDD و روش CFS

		Full-Set(46)				CFS(7)			
	Method	LR	RF	CART	J48	LR	RF	CART	J48
Norm	TP Rate	**0.999**	0.999	0.996	0.994	0.995	**0.995**	0.995	0.994
	FP Rate	**0.030**	0.088	0.088	0.077	0.167	**0.148**	0.156	0.151
	Precision	**0.997**	0.991	0.990	0.992	0.982	**0.984**	0.983	0.984
	Recall	**0.999**	0.999	0.996	0.994	0.995	**0.995**	0.995	0.994
	F-Measure	**0.998**	0.995	0.993	0.993	0.989	**0.989**	0.989	0.989
Anorm	TP Rate	**0.970**	0.913	0.912	0.923	0.833	**0.852**	0.844	0.849
	FP Rate	**0.001**	0.001	0.004	0.006	0.005	**0.005**	0.005	0.006
	Precision	**0.986**	0.989	0.960	0.941	0.949	**0.947**	0.947	0.937
	Recall	**0.970**	0.913	0.912	0.923	0.833	**0.852**	0.844	0.849
	F-Measure	**0.978**	0.949	0.935	0.932	0.887	**0.897**	0.893	0.891
Accuracy		**99.5759**	99.0401	98.7587	98.6739	97.9144	**98.0686**	97.9954	97.9453
Time		**2.52**	10.05	10.92	1.7	0.41	**3.87**	2.61	0.23
LR: Logistic Regression – FR: Random Forest									

جدول (۵). نتایج پیش‌بینی بر روی ۴۶ ویژگی معمولی مجموعه‌داده ECML/PKDD و روش mRMR

		Full-Set(46)				mRMR (10)			
	Method	LR	RF	CART	J48	LR	RF	CART	J48
Norm	TP Rate	**0.999**	0.999	0.996	0.994	**0.998**	0.998	0.996	0.995
	FP Rate	**0.030**	0.088	0.088	0.077	**0.066**	0.068	0.089	0.077
	Precision	**0.997**	0.991	0.990	0.992	**0.993**	0.993	0.990	0.992
	Recall	**0.999**	0.999	0.996	0.994	**0.998**	0.998	0.996	0.995
	F-Measure	**0.998**	0.995	0.993	0.993	**0.995**	0.995	0.993	0.993
Anorm	TP Rate	**0.970**	0.913	0.912	0.923	**0.934**	0.932	0.911	0.923
	FP Rate	**0.001**	0.001	0.004	0.006	**0.002**	0.002	0.004	0.005
	Precision	**0.986**	0.989	0.960	0.941	**0.979**	0.979	0.958	0.952
	Recall	**0.970**	0.913	0.912	0.923	**0.934**	0.932	0.911	0.923
	F-Measure	**0.978**	0.949	0.935	0.932	**0.956**	0.955	0.934	0.938
Accuracy		**99.5759**	99.0401	98.7587	98.6739	**99.1442**	99.1365	98.7317	98.7895
Time		**2.52**	10.05	10.92	1.7	**0.61**	2.67	2.88	0.34
LR: Logistic Regression – FR: Random Forest									

همان‌طور که در دو جدول فوق دیده می‌شود خروجی حاصل از انتخاب همه ۴۶ ویژگی مؤثر، در دو بخش عادی و غیرعادی با استفاده از ۴ الگوریتم یادگیری ماشین بنام‌های مدل آماری رگرسیون متغیرهای وابسته، درخت تصمیم‌گیری، درخت طبقه‌بندی کارت[2] و الگوریتم جی ۴۸[3] تقسیم‌بندی شده و نتایج پارامترهای معیار اف[4]، نرخ مثبت کاذب، نرخ منفی واقعی، صحت و بازخوانی در هر دو بخش آورده شده و در کل دقت مدل برای ۴۶ ویژگی مؤثر با الگوریتم مدل آماری رگرسیون متغیرهای وابسته به ۹۹.۵۷۵۹ درصد و زمان ساخت مدل ۲.۵۲ ثانیه بوده است. در ادامه با استفاده از الگوریتم انتخاب ویژگی منطق فازی[5]، بهترین نتیجه مربوط به الگوریتم یادگیری ماشین درخت تصادفی بوده که به‌دقت ۹۸.۰۶۸۶ رسیده شد. سپس با استفاده از الگوریتم انتخاب ویژگی کمینه افزونگی، بیشینه وابستگی، بهترین نتیجه مربوط به الگوریتم یادگیری ماشین مدل آماری رگرسیون متغیرهای وابسته بوده که به‌دقت ۹۹.۱۴۴۲ رسیده شد. از آزمایش فوق در محیط اورنج نیز نتایج مشابهی بر اساس تصاویر زیر به‌دست‌آمده که نشان دهند این نکته است که محیط‌های پیش‌بینی تأثیر چندانی در خروجی حاصله ندارند.

جدول (۶). مقادیر پیش‌بینی الگوریتم‌های یادگیری ماشین با اورنج بر روی ۴۶ ویژگی مؤثر مجموعه‌داده ECML/PKDD

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.999	0.996	0.996	0.996	0.996
Neural Network	0.998	0.992	0.992	0.992	0.992
Random Forest	0.988	0.989	0.989	0.989	0.989
Tree	0.902	0.979	0.978	0.979	0.979
SVM	0.986	0.966	0.967	0.969	0.966
kNN	0.896	0.960	0.956	0.961	0.960
Naive Bayes	0.904	0.801	0.836	0.913	0.801

در جدول (۶) الگوریتم مدل آماری رگرسیون متغیرهای وابسته دقت ۹۹.۶ درصد در پارامتر ستون دوم که همان دقت است را نشان می‌دهد که با مقدار ۹۹.۵۷۵۹ در وکا[6] مشابه است. در جدول (۷) نتایج ماتریس درهم‌ریختگی مقادیر مثبت درست، مثبت کاذب، منفی درست، منفی کاذب را برای الگوریتم مدل آماری رگرسیون متغیرهای وابسته نمایش می‌دهد.

جدول (۷). ماتریس درهم‌ریختگی ۴۶ ویژگی مؤثر مجموعه‌داده ECML/PKDD

Random Forest
Logistic Regression
Tree
SVM
kNN
Naive Bayes
Neural Network

Show: Number of instances

Actual \ Predicted	anorm	norm	Σ
anorm	2472	88	2560
norm	30	23350	23380
Σ	2502	23438	25940

همان‌طور که در جدول (۷) پیداست از ۲۵۶۰ نمونه غیرنرمال تعداد ۲۴۷۲ نمونه منفی درست و تعداد ۸۸ نمونه منفی کاذب تشخیص‌داده‌شده است و از ۲۳۳۸۰ نمونه نرمال تعداد ۲۳۳۵۰ نمونه مثبت درست و تعداد ۳۰ نمونه مثبت کاذب تشخیص‌داده‌شده است. در جدول (۸) با استفاده از الگوریتم‌های انتخاب ویژگی ازجمله (ضریب جینی، کسب اطلاعات، نسبت بهره، رتبه‌بندی ویژگی، متغیر تصادفی و منطق فازی) در محیط اورنج یک زیرمجموعه‌ای از ویژگی‌ها را انتخاب کرده و سپس با استفاده از الگوریتم یادگیری ماشین مدل آماری رگرسیون متغیرهای وابسته مورد آزمایش قرار داده و مقایسه می‌کنیم.

[1] Weka
[2] CART Algoritm
[3] J48 Algoritm
[4] F-Measure
[5] Container Freight Station(CFS)
[6] Weka

جدول (۸). انتخاب ویژگی به پنج روش در اورنج

	#	Info. gain	Gain ratio	Gini	χ^2	ReliefF
max_len_value_in_arq		0.198	0.099	0.053	4060.007	0.063
nr_of_keywords		0.191	0.124	0.068	11565.890	0.065
operators		0.175	0.088	0.049	3577.102	0.086
Len_of_allArqValue		0.092	0.046	0.023	2284.842	0.030
num_of_spesialChInArgument		0.091	0.046	0.022	2340.133	0.046
token_count_url		0.062	0.031	0.015	1679.452	0.027
payloadUpperChars		0.059	0.030	0.014	1521.812	0.019
payloadDigits		0.058	0.029	0.013	1436.132	0.022
LenOfPayload		0.055	0.028	0.012	1353.795	0.025
Length_of_url		0.045	0.023	0.011	1243.862	0.023
num_of_DistinctByteInPayload		0.041	0.021	0.009	1009.417	0.064
avg_token_length_url		0.039	0.019	0.009	1072.210	0.028
num_of_letterInArq		0.033	0.017	0.007	789.055	0.022
payloadLowerChars		0.024	0.012	0.005	552.416	0.020
min_len_value_in_arq		0.010	0.005	0.002	412.175	0.019
Len_of_allArqName		0.010	0.005	0.002	139.185	0.034
number_of_arguments		0.009	0.005	0.002	190.486	0.041
entropy		0.005	0.003	0.001	153.971	0.024
LenOfRequest		0.004	0.002	0.001	104.251	0.015
No_of_dots_url		0.004	0.002	0.001	94.752	0.014

بـا بـه‌کارگیری الگوریتم‌هـای انتخـاب ویژگـی ذکرشـده، ویژگی‌هایی که بالاترین رتبه‌بندی را داشته باشـند بـه ترتیـب از بالابه‌پایین مرتب می‌شوند که ۲۰ ویژگی اول را شـامل می‌گـردد. بر اساس جدول فوق از بالابه‌پایین، ابتدا اولین ویژگی با رتبه‌بندی بالا را انتخاب کرده، سپس دو ویژگـی بـا رتبه‌بنـدی بـالا را و بـه همین ترتیب تا ترکیب ۲۰ ویژگی ادامه می‌دهیم و نتایج حاصـله که شامل دقت است را در قالب جدول (۹) به دست آوردیم.

جدول (۹). نتایج پیش‌بینی انتخاب ویژگی به پنج روش

CHI2		Info Gain		Gain Ratio		Gini Index		ReliefF	
FS	ACC	FS	ACC	FS	ACC	FS	ACC	FS	ACC
۱	۹۳.۱	۱	۹۷.۵	۱	۹۳.۱	۱	۹۳.۱	۱	۹۲.۹
۲	۹۷.۶	۲	۹۷.۶	۲	۹۷.۶	۲	۹۷.۶	۲	۹۲.۹
۳	۹۷.۹	۳	۹۷.۹	۳	۹۷.۹	۳	۹۷.۹	۳	۹۷.۷
۴	۹۸.۹	۴	۹۸.۹	۴	۹۸.۹	۴	۹۸.۹	۴	۹۸.۴
۵	۹۹.۰	۵	۹۹.۰	۵	۹۹.۰	۵	۹۹.۰	۵	۹۹.۱
۶	۹۹.۱	۶	۹۹.۱	۶	۹۹.۱	۶	۹۹.۱	۶	۹۸.۶
۷	۹۹.۱	۷	۹۹.۱	۷	۹۹.۱	۷	۹۹.۱	۷	۹۸.۶
۸	۹۹.۱	۸	۹۹.۱	۸	۹۹.۱	۸	۹۹.۱	۸	۹۸.۶
۹	۹۹.۱	۹	۹۹.۱	۹	۹۹.۱	۹	۹۹.۱	۹	۹۹.۰
۱۰	۹۹.۲	۱۰	۹۹.۲	۱۰	۹۹.۲	۱۰	۹۹.۲	۱۰	۹۹.۱
۱۱	۹۹.۲	۱۱	۹۹.۲	۱۱	۹۹.۲	۱۱	۹۹.۲	۱۱	۹۹.۱
۱۲	۹۹.۲	۱۲	۹۹.۲	۱۲	۹۹.۲	۱۲	۹۹.۲	۱۲	۹۹.۱
۱۳	۹۹.۳	۱۳	۹۹.۳	۱۳	۹۹.۳	۱۳	۹۹.۳	۱۳	۹۹.۱
۱۴	۹۹.۳	۱۴	۹۹.۳	۱۴	۹۹.۳	۱۴	۹۹.۳	۱۴	۹۹.۱
۱۵	۹۹.۳	۱۵	۹۹.۳	۱۵	۹۹.۳	۱۵	۹۹.۳	۱۵	۹۹.۴
۱۶	۹۹.۳	۱۶	۹۹.۵	۱۶	۹۹.۵	۱۶	۹۹.۳	۱۶	۹۹.۵
۱۷	۹۹.۳	۱۷	۹۹.۶	۱۷	۹۹.۶	۱۷	۹۹.۵	۱۷	۹۹.۴
۱۸	۹۹.۶	۱۸	۹۹.۶	۱۸	۹۹.۶	۱۸	۹۹.۶	۱۸	۹۹.۵
۱۹	۹۹.۵	۱۹	۹۹.۵	۱۹	۹۹.۶	۱۹	۹۹.۶	۱۹	۹۹.۵
۲۰	۹۹.۵	۲۰	۹۹.۵	۲۰	۹۹.۶	۲۰	۹۹.۵	۲۰	۹۹.۵

برای الگوریتم‌های (ضریب جینـی، کسـب اطلاعـات، نسـبت بهره، رتبه‌بندی ویژگی، متغیر تصادفی و منطق فازی) بـه ترتیـب در سـطرهای ۱۸،۱۷،۱۷،۱۸،۱۶ بـه بهتـرین زیرمجموعـه‌ای از ویژگی‌ها که بهترین نتیجه را به دست آورده‌اند رسیدیم.

در ادامه بر روی ۴۵ ویژگی مؤثر مربوط به مجموعـه‌داده دوم آزمایش‌های خود را توسط الگوریتم‌های مختلف یادگیری ماشین در محیط وکا بکار گرفتیم و همچنین از الگوریتم انتخاب ویژگـی کمینه افزونگی، بیشینه وابستگی و منطـق فـازی جهـت انتخـاب زیرمجموعه‌ای از ویژگی‌ها استفاده کرده و نتـایج جـدول (۱۰) و جدول (۱۱) را به دست آوردیم.

جدول (۱۰). نتایج پیش‌بینی بر روی ۴۵ ویژگی معمولی مجموعه‌داده CSIC 2010 و CFS

		Full-Set(45)				CFS(6)			
	Method	LR	**RF**	CART	J48	LR	**RF**	CART	J48
Norm	TP Rate	0.982	**0.997**	0.997	0.996	0.601	**0.931**	0.927	0.923
	FP Rate	0.051	**0.009**	0.012	0.010	0.125	**0.007**	0.007	0.007
	Precision	0.969	**0.994**	0.992	0.994	0.750	**0.989**	0.988	0.944
	Recall	0.982	**0.997**	0.997	0.996	0.601	**0.931**	0.927	0.988
	F-Measure	0.975	**0.995**	0.995	0.995	0.667	**0.959**	0.957	0.955
Anorm	TP Rate	0.949	**0.991**	0.988	0.990	0.875	**0.993**	0.993	0.993
	FP Rate	0.018	**0.003**	0.003	0.004	0.399	**0.069**	0.073	0.077
	Precision	0.970	**0.994**	0.995	0.994	0.779	**0.958**	0.956	0.954
	Recall	0.949	**0.991**	0.988	0.990	0.875	**0.993**	0.993	0.993
	F-Measure	0.960	**0.993**	0.991	0.992	0.824	**0.976**	0.974	0.973
Accuracy		96.9323	**99.438**	99.3226	99.3687	76.9823	**96.9323**	96.7937	96.6282
Time		3.02	**7.74**	5.78	1.41	0.44	**3.67**	1.98	0.33
LR: Logistic Regression – FR: Random Forest									

جدول (۱۱). نتایج پیش‌بینی بر روی ۴۵ ویژگی معمولی مجموعه‌داده CSIC 2010 و mRMR

		Full-Set(45)				mRMR (10)			
	Method	LR	**RF**	CART	J48	LR	**RF**	CART	J48
Norm	TP Rate	0.982	**0.997**	0.997	0.996	0.966	**0.992**	0.990	0.993
	FP Rate	0.051	**0.009**	0.012	0.010	0.330	**0.046**	0.050	0.059
	Precision	0.969	**0.994**	0.992	0.994	0.824	**0.972**	0.969	0.964
	Recall	0.982	**0.997**	0.997	0.996	0.966	**0.992**	0.990	0.993
	F-Measure	0.975	**0.995**	0.995	0.995	0.889	**0.982**	0.980	0.979
Anorm	TP Rate	0.949	**0.991**	0.988	0.990	0.670	**0.954**	0.950	0.941
	FP Rate	0.018	**0.003**	0.003	0.004	0.037	**0.008**	0.010	0.007
	Precision	0.970	**0.994**	0.995	0.994	0.924	**0.986**	0.984	0.989
	Recall	0.949	**0.991**	0.988	0.990	0.670	**0.954**	0.950	0.941
	F-Measure	0.960	**0.993**	0.991	0.992	0.777	**0.970**	0.966	0.964
Accuracy		96.9323	**99.438**	99.3226	99.3687	85.2117	**97.7021**	97.4596	97.3172
Time		3.02	**7.74**	5.78	1.41	0.41	**4.56**	2.78	0.45
LR: Logistic Regression – FR: Random Forest									

همان‌طور که در دو جدول فوق دیده می‌شود خروجی حاصل از انتخاب همه ۴۵ ویژگی مؤثر، در دو بخش غیرعادی و عادی بـا استفاده از ۴ الگوریتم یـادگیری ماشـین بنام‌هـای مـدل آمـاری رگرسـیون متغیرهـای وابسـته، درخـت تصـمیم‌گیری، درخـت طبقه‌بندی کارت و الگوریتم جی ۴۸ تقسـیم‌بندی شـده و نتـایج پارامترهای معیار اف، نرخ مثبت کاذب، نرخ منفی واقعی، صـحت و بازخوانی در هر دو بخش آورده شده و در کل دقت مـدل بـرای ۴۵ ویژگی مؤثر با الگوریتم درخت تصادفی بـه ۹۹.۴۳۸ درصـد و زمان ساخت مدل ۷.۷۴ ثانیه بوده است. در ادامـه بـا اسـتفاده از الگوریتم انتخاب ویژگی منطق فازی، بهتـرین نتیجـه مربـوط بـه الگـوریتم یـادگیری ماشـین درخـت تصـادفی بـوده کـه بـه‌دقت ۹۶.۹۳۲۳ رسیده شـد. سـپس بـا اسـتفاده از الگـوریتم انتخـاب ویژگی کمینه افزونگی، بیشینه وابستگی، بهترین نتیجه مربوط به الگـوریتم یـادگیری ماشـین درخـت تصـادفی بـوده کـه بـه‌دقت ۹۷.۷۰۲۱ رسیده شد.

در آخر بر روی ۱۷ ویژگی مؤثر مربوط به مجموعه‌داده سـوم آزمایش‌های خود را توسط الگوریتم‌های مختلف یادگیری ماشین در محیط وکا بکار گرفتیم و همچنین از الگوریتم انتخاب ویژگـی کمینه افزونگی، بیشینه وابستگی و منطـق فـازی جهـت انتخـاب زیرمجموعه‌ای از ویژگی‌ها استفاده کرده و نتـایج جـدول (۱۲) و جدول (۱۳) را به دست آوردیم.

جدول (۱۲). نتایج پیش‌بینی بر روی ۱۷ ویژگی معمولی مجموعه‌داده HttpParam2016 و CFS

		Full-Set(17)				CFS(5)			
	Method	LR	RF	CART	J48	LR	RF	CART	J48
Norm	TP Rate	1.000	**1.000**	0.999	0.999	0.998	**0.999**	0.999	0.999
	FP Rate	0.001	**0.001**	0.001	0.001	0.003	**0.002**	0.003	0.003
	Precision	0.999	**1.000**	0.999	0.999	0.998	**0.999**	0.999	0.999
	Recall	1.000	**1.000**	0.999	0.999	0.998	**0.999**	0.999	0.999
	F-Measure	1.000	**1.000**	0.999	0.999	0.998	**0.999**	0.999	0.999
Anorm	TP Rate	0.999	**0.999**	0.999	0.999	0.997	**0.998**	0.997	0.997
	FP Rate	0.000	**0.000**	0.001	0.001	0.002	**0.001**	0.001	0.001
	Precision	0.999	**0.999**	0. 998	0.998	0.997	**0.998**	0.998	0.998
	Recall	0.999	**0.999**	0.999	0.999	0.997	**0.998**	0.997	0.997
	F-Measure	0.999	**0.999**	0.999	0.999	0.997	**0.998**	0.998	0.998
Accuracy		99.937	**99.9469**	99.9038	99.9038	99.7777	**99.8441**	99.8275	99.8242
Time		1.67	**4.28**	3.12	0.52	0.42	**2.46**	1.56	0.14
LR: Logistic Regression – FR: Random Forest									

جدول (۱۳). نتایج پیش‌بینی بر روی ۱۷ ویژگی معمولی مجموعه‌داده HttpParam2016 و mRMR

		Full-Set(17)				mRMR (10)			
	Method	LR	RF	CART	J48	LR	RF	CART	J48
Norm	TP Rate	1.000	**1.000**	0.999	0.999	0.999	**0.999**	0.999	0.999
	FP Rate	0.001	**0.001**	0.001	0.001	0.002	**0.001**	0.001	0.001
	Precision	0.999	**1.000**	0.999	0.999	0.999	**1.000**	0.999	0.999
	Recall	1.000	**1.000**	0.999	0.999	0.999	**0.999**	0.999	0.999
	F-Measure	1.000	**1.000**	0.999	0.999	0.999	**1.000**	0.999	0.999
Anorm	TP Rate	0.999	**0.999**	0.999	0.999	0.998	**0.999**	0.999	0.999
	FP Rate	0.000	**0.000**	0.001	0.001	0.001	**0.001**	0.001	0.001
	Precision	0.999	**0.999**	0. 998	0.998	0.998	**0.999**	0.998	0.999
	Recall	0.999	**0.999**	0.999	0.999	0.998	**0.999**	0.999	0.999
	F-Measure	0.999	**0.999**	0.999	0.999	0.998	**0.999**	0.999	0.999
Accuracy		99.937	**99.9469**	99.9038	99.9038	99.8607	**99.927**	99.9005	99.9137
Time		1.67	**4.28**	3.12	0.52	0.67	**3.58**	2.2	0.27
LR: Logistic Regression – FR: Random Forest									

همان‌طور که در دو جدول فوق دیده می‌شود خروجی حاصل از انتخاب همه ۱۷ ویژگی مؤثر، در دو بخش غیرعادی و عادی با استفاده از ۴ الگوریتم یادگیری ماشین بنام‌های مدل آماری رگرسیون متغیرهای وابسته، درخت تصمیم‌گیری، درخت طبقه‌بندی کارت و الگوریتم جی ۴۸ تقسیم‌بندی شده و نتایج پارامترهای معیار اف، نرخ مثبت کاذب، نرخ منفی واقعی، صحت و بازخوانی در هر دو بخش آورده شده و در کل دقت مدل برای ۱۷ ویژگی مؤثر با الگوریتم درخت تصادفی به ۹۹.۹۴۶۹ درصد و زمان ساخت مدل ۴.۲۸ ثانیه بوده است. در ادامه با استفاده از الگوریتم انتخاب ویژگی منطق فازی، بهترین نتیجه مربوط به الگوریتم یادگیری ماشین درخت تصادفی بوده که به‌دقت ۹۹.۸۴۴۱ رسیده شد. سپس با استفاده از الگوریتم انتخاب ویژگی کمینه افزونگی، بیشینه وابستگی، بهترین نتیجه مربوط به الگوریتم یادگیری ماشین درخت تصادفی بوده که به‌دقت ۹۹.۹۲۷ رسیده شد.

جدول (۱۴). نتایج سه مجموعه‌داده، با همه ویژگی‌ها و الگوریتم‌های انتخاب ویژگی

Compare Accuracy In Dataset								
Dataset	No FS	Info.Gain	Gain.Ratio	Chi2	ReliefF	CFS	mRMR	Wrapper
ECML/PKDD 2007(LR)	**99.5759**	99.2251	98.882	99.2251	99.0054	98.0686	99.1442	90.4588
CSIC 2010(RF)	**99.438**	97.5789	97.2363	97.6097	97.5943	96.9323	97.7021	81.4396
HttpParam2016(RF)	**99.9469**	99.927	99.9104	99.927	99.937	99.8441	99.927	99.9436

Compare Time In Dataset								
Dataset	No FS	Info.Gain	Gain.Ratio	Chi2	ReliefF	CFS	mRMR	Wrapper
ECML/PKDD 2007	2.52	0.55	0.48	0.55	0.47	3.87	0.61	0.38
CSIC 2010	7.74	5.53	4.68	5.67	4.88	3.67	4.56	10.18
HttpParam2016	4.28	3.39	3.27	3.31	3.34	2.46	3.58	3.26

جدول (۱۴) مقایسه سه مجموعه‌داده و نتایج حاصله از انتخاب همه ویژگی‌ها (ستون اول) و استفاده از الگوریتم‌های انتخاب ویژگی (سایر ستون‌ها) را نشان می‌دهد.

مقایسه با سایر کارهای مشابه:

کار ما (ستون‌های سبزرنگ دوم) با کارهای مشابه انجام‌شده دررابطه‌با مجموعه‌داده اول و دوم جدول (۱۵) و (۱۶) در سال ۲۰۱۱ [۳۶].

جدول (۱۵). نتایج مربوط به مجموعه‌داده ECML/PKDD

Table 4. Experimental results on the ECML/PKDD 2007 dataset									
Classifiers	Detection rate						False Positive Rate		
	Full-set		CFS		mRMR		Full-set	CFS	mRMR
C4.5	96.37	98.6739	86.45	97.9453	91.62	98.7895	3.7	17.6	9.9
CART	96.11	98.7587	86.45	97.9954	91.54	98.7317	4.3	17.6	10
Random Tree	96.89	97.818	86.39	97.0779	93.41	98.3847	2.6	17.7	6.4
Random Forest	98.80	99.0401	86.39	98.0686	95.18	99.1365	1.2	17.7	5.0
Average	**97.04**	98.5726	**86.42**	97.7718	**92.93**	98.7606	**2.95**	**17.6**	**7.8**

جدول (۱۶). نتایج مربوط به مجموعه‌داده CSIC

Table 3. Experimental results on the CSIC 2010 dataset									
Classifiers	Detection rate						False Positive Rate		
	Full-set		CFS		mRMR		Full-set	CFS	mRMR
C4.5	94.49	99.3687	94.06	96.6282	79.80	97.3172	5.9	6.8	25.7
CART	94.12	99.3226	93.71	96.7937	79.85	97.4596	6.2	6.8	25.3
Random Tree	92.30	98.9453	92.70	96.9669	71.36	97.6405	8.3	7.8	30.6
Random Forest	93.71	99.438	93.68	96.9323	71.70	97.7021	7.2	7.2	30.5
Average	**93.65**	99.2686	**93.53**	96.8302	**75.67**	97.5298	**6.9**	**7.1**	**28**

کار ما (ستون‌های سبزرنگ دوم) با کارهای مشابه انجام‌شده دررابطه‌با مجموعه‌داده دوم جدول (۱۷) در سال ۲۰۱۴ [۳۷].

جدول (۱۷). نتایج مربوط به مجموعه‌داده CSIC

Classifiers	Detection rate			
	Full-set		CFS	
C4.5	97.8	99.3687	91.74	96.6282
CART	97.46	99.3226	92.01	96.7937
Random Tree	97.62	98.9453	92.64	96.9669
Random Forest	99.18	99.438	92.64	96.9323
Average	**98.02**	99.2686	**93.53**	96.8302

کار ما (ستون‌های سبزرنگ دوم) با کارهای مشابه انجام‌شده دررابطه‌با مجموعه‌داده اول و دوم جدول (۱۸) و (۱۹) در سال ۲۰۱۵ [۲۳].

جدول (۱۸). نتایج مربوط به مجموعه‌داده ECML/PKDD

Detection rate and false positive rate of four decision trees performed on the *expert knowledge* subset of the *ECML/PKDD* dataset									
Classifiers	Detection rate						False Positive Rate		
	Full-set		CFS		mRMR		Full-set	CFS	mRMR
C4.5	95.42	98.6739	89.07	97.9453	92.30	98.7895	7.8	23.2	15.1
CART	95.51	98.7587	89.12	97.9954	92.23	98.7317	7.9	23.1	14.7
Random Tree	92.43	97.818	89.11	97.0779	88.81	98.3847	10.6	23.2	16.3
Random Forest	95.80	99.0401	89.10	98.0686	91.53	99.1365	8.1	23.2	15.1
Average	94.79	98.5726	89.10	97.7718	91.22	98.7606	8.6	23.18	15.3

جدول (۱۹). نتایج مربوط به مجموعه‌داده CSIC

Detection rate and false positive rate of four decision trees performed on the *expert knowledge* subset of the *CSIC* dataset									
Classifiers	Detection rate						False Positive Rate		
	Full-set		CFS		mRMR		Full-set	CFS	mRMR
C4.5	94.98	99.3687	94.07	96.6282	79.8	97.3172	5.3	6.8	25.7
CART	94.31	99.3226	93.72	96.7937	79.86	97.4596	7.0	6.8	25.3
Random Tree	92.76	98.9453	92.71	96.9669	71.36	97.6405	7.7	7.8	30.6
Random Forest	93.93	99.438	93.69	96.9323	71.70	97.7021	6.0	7.2	30.5
Average	93.99	99.2686	93.55	96.8302	**75.5**	97.5298	6.5	7.15	28.02

کار ما (ستون‌های سبزرنگ دوم) با کارهای مشابه انجام‌شده دررابطه‌با مجموعه‌داده دوم جدول (۲۰) در سال ۲۰۱۶ [۳۸]

جدول (۲۰). نتایج مربوط به مجموعه‌داده CSIC

Best performance of CISIC2010										
	Method	RdF		LRg		DTr		ABc		SGDc	
anom	Precision	79.70	99.4	99.39	97.0	88.10	99.4	67.24	81.0	72.45	96.1
	Recall	87.11	99.1	93.05	94.9	88.28	98.8	89.19	85.8	92.04	85.2
	F-Measure	83.24	99.3	96.11	96.0	88.19	99.1	76.68	83.4	81.08	90.3
norm	Precision	83.37	99.4	92.54	96.9	86.48	99.2	80.06	90.8	86.69	91.4
	Recall	74.46	99.7	99.34	98.2	86.26	99.6	49.98	87.5	59.70	97.9
	F-Measure	78.67	99.5	95.82	97.5	86.37	99.4	61.54	89.1	70.71	94.5
	Average	81.53	99.4	95.965	96.95	87.29	99.3	73.65	85.9	79.57	93.75

SGD:stochastic gradient descent-DT:Decision Tree-ABC:ADBoost

کارهای مشابه انجام‌شده دررابطه‌با مجموعه‌داده اول، دوم و سوم جدول (۲۱) در سال ۲۰۱۹ [۳۹]

جدول (۲۱). نتایج مربوط به سه مجموعه‌داده

Paper	SBD	ABD	Detection Method	Dataset	[%]
(Stephan et al. 2015)[4]	×	√	YSA	HTTP Dataset	95.00
(Nguyen et al. 2013)[5]	×	√	C4.5	CSIC 2010	94.49
				ECML-PKDD 2007	96.30
			CART	CSIC 2010	94.12
				ECML-PKDD 2007	96.11
			Random Tree	CSIC 2010	92.30
				ECML-PKDD 2007	96.89
			Random Forest	CSIC 2010	93.71
				ECML-PKDD 2007	98.80
(Kirchner 2010) [26]	×	√	K-Means	HTTP Dataset	90.90
(Zolotukhin et al. 2012) [27]	×	√	N Gram	HTTP Dataset	97.70
(Tekerek et al. 2016)[11]				WAF 2015	97.60
	√	√	Bayes	CSIC 2010	94.50
				ECML-PKDD 2007	93.30
				WAF 2015	98.52
Proposed Model	√	√	ANN	CSIC 2010	96.74
				ECML-PKDD 2007	94.53

Tab. II *Comparison for datasets, feature, and detection method.*

در جدول (۲۲) نتایج کار ما است که در مقایسه با کارهای قبلی نتایج بسیار بهتری به‌دست‌آمده و نشانگر این است که ویژگی‌هایی که ما در این تحقیق استخراج کرده‌ایم باتوجه‌به انتخاب مناسب جداول پایه‌ای مربوط به ویژگی‌ها، تأثیر بسیار زیادی در خروجی و تشخیص موارد نرمال و غیرنرمال را داشته است.

جدول (۲۲). نتایج پیش‌بینی ما در سه مجموعه‌داده

Dataset	پیش بینی ما
ECML/PKDD ۲۰۰۷	۹۹.۵۷۵۹
CSIC ۲۰۱۰	۹۹.۴۳۸
HttpParam۲۰۱۶	۹۹.۹۴۶۹

۹- نتیجه‌گیری

استخراج ویژگی موضوع بسیار مهمی در تشخیص حملات وب است. چراکه اکثر حملات وب مجموعه‌ای از به‌کارگیری شخصیت‌های خاص در حملات توسط حمله‌کننده‌ها است. با انتخاب دسته‌بندی مناسب شخصیت‌های حمله که شامل پنج دسته‌بندی اصلی بوده، باتوجه‌به نتایجی که به‌دست‌آمده نشان‌دهنده این موضوع است که ویژگی‌های استخراج‌شده کیفیت بالایی را دارا بوده، چراکه این پنج دسته‌بندی انواع حملات مربوط به تزریق کد را شامل می‌شود. پس از ارزیابی برخی از الگوریتم‌هایی که به طور گسترده مورداستفاده برای تشخیص حملات تزریق کد قرار گرفته است، مشاهده شد که درخت تصمیم‌گیری به‌طورکلی بهتر از دیگر الگوریتم‌های مورداستفاده برای تشخیص تزریق کد است. از این آزمایش مشخص‌شده است که داشتن ویژگی‌های بیشتر همیشه به معنای میزان دقت بالاتر نیست و برعکس.

استفاده از انتخاب ویژگی‌های خودکار و همچنین انتخاب ویژگی‌های دستی هر دو به طور گسترده‌ای در تحقیقات مورداستفاده قرار می‌گیرد، ارزیابی این روش‌های انتخاب ویژگی ثابت می‌کند که داشتن دانش تخصصی در مورد حوزه موردنظر اجازه می‌دهد تا از ویژگی‌های مرتبط به هم که از مجموعه‌داده‌ها استخراج می‌شوند، منجر به ارائه بهترین نتیجه تشخیص شود.

درنهایت، از نتایج به‌دست‌آمده در هنگام انتخاب همه ویژگی‌های مجموعه‌داده‌ها و سپس انتخاب تنها تعدادی از ویژگی‌های مربوط به تزریق کد تفاوت کمی در دقت وجود دارد. همچنین مشاهده شد که مقدار مشخصی از ویژگی‌های بسیار مهم‌تر از بقیه است، زیرا ویژگی‌ها با استفاده از الگوریتم‌های انتخاب ویژگی از ۴۶ ویژگی به طور متوسط بین ۵ الی ۲۰ ویژگی محدودشده است، ولی بازهم نتایج به هم نزدیک بوده است. به‌طورکلی می‌توان گفت که:

۱. این تحقیق منجر به ایجاد یک مجموعه استاندارد از ویژگی‌های مؤثر در عمق برای استفاده در آزمایش‌های گردید که در آن ویژگی‌های به‌دست‌آمده به طور مستقیم به کشف تزریق کد کمک می‌کنند.

۲. باتوجه‌به خروجی‌های بسیار خوب پیش‌بینی در سه مجموعه‌داده ذکرشده نشان‌دهنده این است که ویژگی‌های استخراج‌شده دارای کیفیت بسیار بالایی هستند و بر روی هر مجموعه‌داده حاوی حملات وب، نتایج عملکرد بسیار خوبی دارند.

۳. نتایج حاصله نشان‌دهنده این است که اگر ویژگی‌های مؤثر باکیفیت بسیار خوب استخراج‌شده باشند، محیط‌های داده‌کاوی فرق آن‌چنانی در درصد پیش‌بینی ندارند و نتایج خروجی بسیار شبیه به هم هستند.

۴. استفاده از الگوریتم‌های انتخاب ویژگی درجایی که ویژگی‌ها زیاد باشند و به‌صورت دستی نتوان ویژگی‌هایی که تأثیر بیشتری در دقت تشخیص دارند را انتخاب کرد، بسیار چاره‌ساز خواهد بود.

۵. از کاربردهای خوب این ویژگی‌ها، استفاده مؤثر در دیواره آتش تحت وب، جهت جلوگیری از انواع حملات سمت پایگاه‌داده است.

۶. هر چه‌کار بیشتری در مرحله پیش‌پردازش انجام گردد، یعنی دستیابی به ویژگی‌های مؤثر، نتیجه نهایی و خروجی بسیار قابل‌اطمینان‌تر است.

۱۰- مراجع

[1] M. C. R. R. a. W. H. Rafał Kozik ،"Semi-unsupervised Machine Learning for Anomaly Detection in HTTP Traffic "،*Institute of Telecommunications and Computer ScienceUTP University of Science and Technology Bydgoszcz Poland* ،2016 .

[2] H. K. K. a. P. B. Rupam Kr. Sharma ،"Analysis of Machine Learning Techniques Based Intrusion Detection Systems "، *North Eastern Hills UniversityShillongIndia, Assam Don Bosco UniversityGuwahatiIndia* ،2016 .

[3] J.-G. Kim ،"Injection Attack Detection using the Removal of SQL Query Attribute Values "،*Department of Computer Science, Namseoul University 21 Maeju-ri Sunghwan-eub Cheonan-si Chungnam, Korea* ،2011 .

[4] P. Aaby ،"Evaluating Web App Datasets towards Detection of SQL Injection Attacks with Machine Learning Techniques "،*Edinburgh Napier University* ،2016 .

[5] OWASP, 2018. [Online]. Available: https://owasp.org/Top10/#welcome-to-the-owasp-top-10-2021.

[6] J. W. f. Wikipedia, "https://fa.wikipedia.org/wiki," 2003. [Online].

[7] D. Litchfield ،"Data-mining with SQL Injection and Inference "،*An NGSSoftware Insight Security Research (NISR) Publication* .۲۰۰۵ ،

[8] K. Wang ،"Anomalous Payload-Based Network Intrusion Detection "،*Lecture Notes in Computer Science. Springer Berlin. 3224: 203–222. doi:10.1007/978-3-540-30143-1_11. ISBN 978-3-540-23123-3* ،2004 .

[9] O. M. Ü. S. T. Carsamba Ticaret Borsasi Meslek Yüksekokulu, "Detection of SQL Injection Attacks Using Machine Learning Algorithms Based on NLP-Based Feature Extraction," in *2024 9th International Conference on Computer Science and Engineering (UBMK), Antalya, Turkiye, 2024, pp. 468-472, doi: 10.1109/UBMK63289.2024.10773483.*, 2024.

[10] S. A. M. M. C. a. A. D. M. Thilakraj, "Detection of SQL Injection Attacks," in *2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 2024, pp. 1515-1520, doi: 10.1109/ICICT60155.2024.10544579.*, 2024.

[11] S. S. L. C. B. Souissi ،"A Novel Security Architecture Based on Multi-level Rule Expression Language "،*Hybrid Intelligent Systems. HIS 2016. Advances in Intelligent Systems and Computing, vol 420. Springer, Cham* ،2016 .

[12] T. F. A. B. A. G. A. K. A. F. M. Rahman ،"SQL Injection Attack Scanner Using Boyer-Moore String Matching Algorithm "،*Journal of Computers* ،2015 .

[13] R. K. S. Mahapatra ،"A Survey of SQL Injection Countermeasures "،*International Journal of Computer Science & Engineering Survey* ،pp. 55-75 ،2012 .

[14] T. Mitchel ،"Machine Learning،" McGraw Hill Science/Engineering/Math ،1997.

[15] J. Brownlee, "Master the Defintions and Concepts," 2014. [Online]. Available: http://www.datascienceassn.org/sites/default/files/Machine%20Learning%20Resource%20Guide%202014.pdf.

[16] S. Kumar ،"Classification and Detection of Computer Intrusions،" 1995.

[17] M. C. R. Moghaddam, "Network Intrusion Detection," in *Proceesings of the 28th International Florida Artificial Intelligence reserach Society Conference*, 2015.

[18] C. Maimbo ،"Exploring the Applicability of SIEM Technology in IT Security،" 2014.

[19] R. Sommer و V. Paxon ،"Outside the Closed World: On Using Machine Learning For Network Intrusion Detection،" International Computer Science Institute ،2010.

[20] H. T. F. K. P. S. Nguyen, "Feature Extraction Methods for Intrusion Detection Systems," in *Threats, Countermeasures, and Advances in Applied Information Security*, 2012, pp. 23-52.

[21] M. Elhamahmy ،H. elmahdy و I. Saroit ،"A New Approach for Evaluating Intrusion Detection System "،*CiiT International Journal of Artificial Intelligent Systems and Machine Learning* ،pp. 290-299 ،2010 .

[22] R. C. I. Mitchel ،"A survey of intrusion detection in wireless network applications "،*Computer Communications* ،pp. 1-23 ،2014 .

[24] C. Giminez ،" Study of Stochastic and Machine Learning Techniques for Anomlay-based Web Attack Detection "،. Madrid: University of Madrid, Department of IT. ،2015.

[25] C. Torrano-Gimenez ،H. T. Nguyen ،G. Alvarez و k. Frank، "Combining Expert Knowledge with Automatic Feature Extraction for Reliable Web Attack Detection "،*Security and Communication Networks* ،pp. 2750-2767 .۲۰۱۲ ،

[26] B. Gallagher و T. Eliassi-Rad ،"LLNL-TR-414570 ،" Lawrence Livermore National Laboratory. ،2009.

[27] K. S. H. a. P. P. K. Rajagopal Smitha ،"A Machine Learning Approach for Web Intrusion Detection: MAMLS Perspective "،*Department of Computer Applications, Manipal Institute of Technology, Manipal Academy of* ، 2019 .

[28] J. Choi, H. Kim, C. Choi and P. Kim, "Efficient Malicoius Code Detection Using N-Gram Analysis and SVM.," in *International Conference on Network-Based Informaiton Systems.*, 2011.

[29] A. G. V. Joshi, "SQL Injection detection using Machine Learning," in *Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2014.

[30] R. I. P. H. M. Komiya ،"Classification of malicious web code by machine learning "،*Awareness Science and Technology(iCAST)* ،pp. 406-411 ،2011 .

[31] A. e. S. D. Christine Basta ،"Detection of SQL Injection Using a Genetic Fuzzy Classifier System "،*(IJACSA) International Journal of Advanced Computer Science and Applications* ،2016 .

[32] E. G. R. M. a. Á. P. Gustavo Betartey ،"Machine learning-assisted virtual patching of web applications "،*Computer Science < Cryptography and Security* ،2018 .

[33] E. K. Y. H. Mohammed BABIKER ،"A hybrid feature-selection approach for finding the digital evidence of web application attacks "،*Turk J Elec Eng & Comp Sci* ،2019 .

[34] R. F. C. C. Nico Epp ،"Anomaly-based Web Application Firewall using HTTP-specific features and One-Class SVM "،*Special Issue: 15 Escola Regional de Redes de Computadores Special Issue (ERRC)* ،2017 .

[35] G. Dong, "A Quick Guide to Feature Engineering," https://www.analyticsvidhya.com, 2019. [Online].

[36] M. M. N. H. M. R. K. B. Salehi ،"Two new methods of class-dependent feature selection and their application in Persian handwritten digit recognition "،*Computer Department, Shiraz University(in Persian)* ،1390 .

[37] *C. T.-G. G. A. P. K. F. Hai Thanh Nguyen, "Application of the Generic Feature Selection Measure in Detection of Web Attacks," Á. Herrero and E. Corchado (Eds.): CISIS 2011, LNCS 6694, pp. 25–32, 2011., 2011.*

[38] *H. T. N. G. A. S. P. ,. K. F. Carmen Torrano-Gimenez, "Applying Feature Selection to Payload-Based Web Application Firewalls," 2014 .*

[39] *T. S. Pham, T. H. Hoang و V. C. Vu, "Machine learning techniques for web intrusion detection – a comparison," Eighth International Conference on Knowledge and Systems Engineering (KSE), 2016 .*

[40] *A. T. Adem Tekerek, "DESIGN AND IMPLEMENTATION OF AN ARTIFICIAL INTELLIGENCE-BASED WEB APPLICATION FIREWALL MODEL," Information Technology Department, Gazi University, 2019 .*