Electronic and Cyber Defens ۲۰۲۵/ *Vol.* ۱۳/ *No.* ۲/ ۱۱۱-۱۲۲

Electronic and Cyber Defens

ISSN: ۲۳۲۲-٤۳٤۷ , *E-ISSN:* ۲۹۸۰-۸۹۷۹

Using Linear Discriminant Analysis for Risk Score Estimation of Unified Resource Locators (URLs)

M. Deypir [۱*], **K. Halili** [۲]

[۱] Associate Professor, Shahid Sattari University of Aeronautical Science and Technology, Tehran, Iran (*Correspondence: mdeypir@ssau.ac.ir)

[۲] Assistant Professor, Shahid Sattari University of Aeronautical Science and Technology, Tehran, Iran

ARTICLE INFO

Article history:
Article Type: Research paper
Received: ۳۰ March ۲۰۲۵
Revised: ۲٦ April ۲۰۲۵
Accepted: ۰۸ June ۲۰۲۵
Available online: ۱۱ July ۲۰۲۵
*Correspondence:
mdeypir@ssau.ac.ir
Keywords:
Linear Discriminant Analysis
Malicious Urls
Security Risk
Textual Feature

ABSTRACT

Sending a malicious URL to a victim is the starting point for various types of malicious activities by attackers. Therefore, identifying malicious URLs plays a significant role in the security of Internet users. Recently, calculating the security risk of URLs instead of classifying them using machine learning-based models has received attention. This is because, on the one hand, it provides necessary warnings to users and, on the other hand, it does not have the problems of classification models. In this study, a new criterion based on linear discriminant analysis is proposed to estimate the security risk of URLs. In this criterion, previously known examples of normal and malicious URLs are mapped into a new space in which the security risk can be calculated more accurately. Although deep learning is not used in the proposed criterion and it requires little training data, it provides a realistic estimate of the security risk values of malicious and safe URLs. Experiments conducted on real datasets show that the proposed criterion outperforms the previously proposed criteria in terms of detection rate. The implementations carried out in this research, along with the datasets used in the experiments, are publicly available at https://github.com/mdeypir/LRU.

Cite this article: Khorasani, Yaghoub©, Shafiee, Ebrahim©, Shamsi, Alireza (۲۰۲۵). Using Linear Discriminant Analysis for Risk Score Estimation of Unified Resource Locators (URLs). Journal of Electronic and Cyber Defens. ۲۰۲۵; ۱۳(۲): ۱۱۱-۱۲۲.

DOI: https://dor.isc.ac/dor/۲۰.۱۰۰۱.۱.۲۳۲۲٤۳٤۷.۱٤۰٤.۱۳.۲.۱۰.۷

Publisher: Imam Hossein University.

پدافند الکترونیک و سایبری/ سال ۱۴۰۴/ دوره ۱۳/ شماره ۲/ صفحه ۱۲۲-۱۱۱

پدافند الکترونیک و سایبری

شاپا الکترونیکی: ۴۳۴۷- ۲۳۲۲۲ **شاپا چاپی:** ۸۹۷۹-۲۹۸۰

تخمین خطر امنیتی نشانی‌های اینترنتی به کمک تحلیل تفکیکی خطی

محمود دی پیر[1*]، خداداد هلیلی[2]

[1] دانشیار، دانشگاه علوم و فنون هوایی شهید ستاری، تهران، ایران (نویسنده مسئول: mdeypir@gmail.com)

[2] استادیار، دانشگاه علوم و فنون هوایی شهید ستاری، تهران، ایران (kh.halili@ssau.ac.ir)

چکیده (استایل عنوان چکیده)

ارسال یک آدرس اینترنتی(URL) مخرب به قربانی نقطه شروع انواع مختلف فعالیت‌های مخرب توسط مهاجمان است؛ بنابراین شناسایی URL های مخرب نقش بسزایی در امنیت کاربران اینترنت دارد. اخیراً محاسبه ریسک امنیتی URL ها به‌جای طبقه‌بندی آن‌ها با استفاده از مدل‌های مبتنی یادگیری ماشینی موردتوجه قرارگرفته است. این امر به این دلیل است که ازیک‌طرف هشدارهای لازم را به کاربران می‌دهد و از طرف دیگر مشکلات مدل‌های طبقه‌بندی را ندارد. در این مطالعه، معیار جدیدی بر اساس تحلیل تفکیک خطی برای تخمین ریسک امنیتی URL ها ابداع‌شده است. در این معیار، نمونه‌های شناخته‌شده قبلی از URLهای عادی و مخرب در فضای جدیدی نگاشت می‌شوند که در آن می‌توان ریسک امنیتی را با دقت بیشتری محاسبه کرد. اگرچه یادگیری عمیق در معیار پیشنهادی استفاده نمی‌شود و به داده‌های آموزشی کمی نیاز دارد، تخمین واقع‌بینانه‌ای برای مقادیر ریسک امنیتی URLهای مخرب و ایمن ارائه می‌دهد. آزمایش‌های انجام‌شده بر روی مجموعه‌های داده واقعی نشان می‌دهد که معیار پیشنهادی ازنظر میزان تشخیص نسبت به معیارهای ارائه‌شده قبلی برتری دارد. پیاده‌سازی‌های انجام‌شده در این پژوهش به همراه مجموعه داده‌های مورداستفاده در آزمایش‌های انجام‌شده در آدرس https://github.com/mdeypir/LRU در دسترس عموم قرارگرفته‌اند.

مشخصات مقاله

تاریخچه مقاله:

نوع مقاله: علمی پژوهشی

دریافت: ۱۴۰۴/۰۱/۱۰

بازنگری: ۱۴۰۴/۰۲/۶

پذیرش: ۱۴۰۴/۰۳/۱۸

ارائه آنلاین: ۱۴۰۴/۰۴/۲۰

کلید واژه‌ها:

تحلیل تفکیک خطی

URL های مخرب

خطر امنیتی

ویژگی متنی

استناد: دی پیر، محمود©، هلیلی، خداداد©. تخمین خطر امنیتی نشانی‌های اینترنتی به کمک تحلیل تفکیکی خطی. پدافند الکترونیک و سایبری. (۱۴۰۴)؛ ۱۳ (۲): ۱۲۲-۱۱۱.

https://dor.isc.ac/dor/ ۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۴۰۴.۱۳.۲.۱۰.۷

ناشر: دانشگاه جامع امام حسین(ع).

۱- مقدمه

URL[1] مخرب، پیوندی است که باهدف تسهیل کلاه‌برداری، حمله سایبری یا هر فعالیت بدخواهانه دیگری ایجادشده است. وقتی کاربر روی یک URL مخرب کلیک می‌کند، می‌تواند باعث دانلود باج افزار، دسترسی به صفحات فیشینگ، ایمیل‌های فیشینگ یا سایر اشکال نفوذ سایبری شود. حجم قابل‌توجهی از مطالعات تحقیقاتی درزمینه شناسایی URLهای مخرب انجام‌شده است. بر اساس این مطالعات، ویژگی‌های مختلفی را می‌توان از آدرس وب‌سایت، محتوای آن‌ها، دامنه مرتبط، سرور و میزبان استخراج کرد. طیف وسیعی از مدل‌هایی که از تکنیک‌های داده‌کاوی یا یادگیری ماشین استفاده می‌کنند با استفاده از این ویژگی‌ها ایجادشده‌اند [۱-۱۹]. این مدل‌های طبقه‌بندی برای تشخیص آدرس‌های مخرب از آدرس‌های مفید استفاده می‌شوند. این روش‌ها به‌طورکلی نرخ مشخصی از تشخیص یا خطای طبقه‌بندی را نشان می‌دهند. علاوه بر این، با درک اینکه چگونه این روش‌ها به‌عنوان یک جعبه سیاه یا یک جعبه سفید عمل می‌کنند، نفوذ گران می‌توانند رویکردهایی برای دور زدن این مدل‌ها ابداع کنند. درنتیجه، عملکرد مدل‌های تشخیص URL کاهش می‌یابد [۲]. برای رسیدگی به این چالش‌ها، رویکرد پیشنهادی به‌جای طبقه‌بندی آن‌ها به‌عنوان مخرب یا غیر مخرب، بر تخمین خطر امنیتی URLها تمرکز دارد. برای این منظور، معیاری را معرفی می‌کنیم. در این معیار، به‌جای برچسب کلاس باینری (مخرب یا مفید)، برای هر URL ورودی ناشناخته، یک مقدار ریسک امنیتی محاسبه می‌شود. بنابراین، هر چه میزان ریسک بالاتر باشد، احتمال مخرب بودن URL بیشتر است. بنابراین، با مشاهده ارزش ریسک یک URL ورودی، کاربر می‌تواند تصمیم بگیرد که آیا با آن تعامل داشته باشد یا خیر. ارزیابی‌های تجربی ما بر روی مجموعه داده‌های URL واقعی، اثربخشی معیار پیشنهادی را در شناسایی URLهای مخرب نسبت به معیارهای ارائه‌شده قبلی، نشان می‌دهد. علاوه بر این، معیار پیشنهادی به‌طور قابل‌توجهی از معیارهای خطر امنیتی مشابهی که برای تشخیص بدافزارها طراحی‌شده بودند، بهتر عمل می‌کند.

ادامه این مقاله به شرح زیر سازمان‌دهی شده است. بخش بعدی مروری بر تحقیقات گذشته در زمینه شناسایی URL های مخرب و محاسبه خطر امنیتی ارائه می‌دهد. بخش سوم به تشریح مسئله تحقیق می‌پردازد. بخش چهارم، معیار پیشنهادی و الگوریتم مربوطه را توضیح می‌دهد. بخش پنجم برتری معیار پیشنهادی را ازنظر نرخ تشخیص URL مخرب، در مقایسه با معیارهای قبلی، با استفاده از دو مجموعه داده دنیای واقعی نشان می‌دهد. درنهایت، بخش ششم، مقاله را جمع‌بندی و نتیجه‌گیری می‌کند.

۲- پیشینه تحقیق

ایجاد و توزیع URL های مخرب یک تاکتیک پرکاربرد است که توسط مهاجمان برای نفوذ به سیستم قربانیان، استفاده می‌شود. چندین سال از استفاده اولیه از این روش می‌گذرد و رویکردهای مختلفی برای رفع آن ارائه‌شده است، اما همچنان یکی از دغدغه-های اصلی در حوزه امنیت سایبری است. این به دلیل استفاده مهاجمان از تکنیک‌های جدید برای غلبه بر این رویکردها است. تحقیقات گسترده‌ای برای شناسایی URL های مخرب با موفقیت نسبی انجام‌شده است. به‌طورکلی، تکنیک‌های شناسایی لینک‌های مخرب را می‌توان به سه دسته، طبقه‌بندی کرد: روش‌های مبتنی بر لیست سیاه [۳]، روش‌های مبتنی بر قانون [۴]، و مهم‌تر از همه، روش‌های مبتنی بر یادگیری ماشینی و یادگیری عمیق [۵]. باتوجه‌به محدودیت‌های دو گروه اول، یادگیری ماشینی و یادگیری عمیق توجه بیشتری را به خود جلب کرده است [۶، ۷]. برای یادگیری ماشین و روش‌های مبتنی بر یادگیری عمیق، ویژگی‌های استخراج‌شده از URL، شامل دامنه مربوطه به آن و میزبان‌های ارجاع شده، طول URL ازنظر تعداد کاراکترها، تعداد کاراکترهای متوالی، تعداد آدرس‌های IP مربوط به هر URL، رتبه الکسا آن، طول عمر URL و زمان انقضای مجوز SSL سایت مربوطه، هستند. پس از استخراج ویژگی‌های لازم، نمونه‌های متعلق به هر گروه از URLهای مخرب یا مفید، با مقادیر مربوط به این ویژگی‌ها، برای ساخت مدل استفاده می‌شوند. سپس مدل ساخته‌شده برای شناسایی برچسب URL های ناشناخته استفاده می‌شود. این مدل‌ها معمولاً طبقه‌بندی دو کلاسه هستند که می‌توانند هر URL ورودی را به‌عنوان نرمال(مفید) یا مخرب دسته‌بندی کنند. مدل‌های یادگیری ماشین مختلفی برای حل این مسئله پیشنهادشده‌اند که موفق‌ترین آن‌ها که دقت‌های بالایی را نشان می‌دهند، درخت‌های تصمیم [۸]، رگرسیون لجستیک [۹]، تقویت گرادیان [۱۰]، ماشین‌های بردار پشتیبان [۱۱]، جنگل‌های تصادفی [۱۲]، و بیز ساده [۱۳] هستند. روش‌های مبتنی بر یادگیری عمیق نیز عملکرد خوبی در این زمینه نشان داده‌اند [۶،۷،۱۴،۱۵]. برخی از محققان با استفاده از مدل‌های طبقه‌بندی گروهی و ترکیب طبقه‌بندی به عملکرد بهتری دست‌یافته‌اند [۱]. در [۱۶]، یک روش ترکیبی مبتنی بر هفت نوع طبقه‌بندی و یک تکنیک خوشه‌بندی برای فیلتر کردن URLهای فیشینگ ابداع شد که تصمیم نهایی با استفاده از تکنیک رأی‌گیری در میان کمیته‌ای از طبقه‌بندی کننده‌ها گرفته می‌شد. علاوه بر این، انواع مختلفی از درختان تصمیم، مانند C۴.۵، ADTree، CART، Random Tree و Random Forest، در یک ساختار یادگیری جمعی در [۱۷] به همراه استفاده از ویژگی‌های جدید برای افزایش دقت طبقه‌بندی، ارائه گردید. اگرچه محققان تلاش‌های قابل‌توجهی برای توسعه مدل‌های کارآمد برای شناسایی URL های مخرب انجام داده‌اند،

[1] Unified Resource Locator

اما هیچ تکنیک کاملی وجود ندارد که بتواند این مسئله را به‌طور کامل حل کند. درواقع، مدل‌ها مستعد خطاهای مثبت کاذب و منفی کاذب در طبقه‌بندی URL ها هستند [۱۸]. شناسایی نشانی‌های اینترنتی به‌طور غیرمستقیم بر امنیت وب تأثیرگذار است، زیرا شروع دسترسی به یک وب‌سایت، واردکردن و استفاده از آدرس URL آن است که خود باید آدرس معتبری باشد وگرنه می‌تواند خود آغازی بر یک حمله سایبری باشد. اگرچه شناسایی URL های مخرب نقش تعیین‌کننده‌ای در امنیت وب‌گردی برای کاربران اینترنت دارد؛ اما امنیت وب به‌طور خاص موردمطالعه بسیاری از محققین قرارگرفته است و روش‌های زیادی تاکنون برای مقابله با این حملات ارائه‌شده است] ۱۹،۲۰،۲۱،۲۲]بنابراین در اینجا به‌مرور برخی از تحقیقات در این زمینه نیز می‌پردازیم. شهیدی نژاد و همکارانش] ۱۹] به کمک مدل دسته‌بندی درخت تصمیم توانستند حملات تزریق SQL را در زمان اجرا تشخیص دهند. در] ۲۰] روشی مبتنی بر امضا به‌منظور شناسایی حملات اجرای کد با استفاده از سامانه تشخیص نفوذ وب در زبان PHP ایجاد شد. یادگاری و همکاران] ۲۱] توانستند به کمک مدل بردار پشتیبان و آنتروپی مربوط به هر آی پی در پنجره‌های زمانی، حملات منع خدمت وب را تشخیص دهند. در] ۲۲] یک سازوکار دفاعی پویا باقابلیت سفارشی‌سازی برای شناسایی ربات‌های وب بدخواه مشارکت‌کننده در حملات سایبری، با استفاده از تحلیل رفتار مرورگری آن‌ها، ارائه‌شده است.

آدرس‌های URL مختص وب‌سایت‌ها نیستند و در هر سیستم مبتنی بر شبکه و اینترنت برای دستیابی به سرورها و منابع می‌توانند استفاده شوند. ارزیابی خطرات امنیتی برای کلیه سیستم‌های فیزیکی سایبری مانند اینترنت اشیا، نیز موردبررسی پژوهشگران امنیت سایبری قرارگرفته است [۲۳]. بااین‌حال، یکی از حوزه‌هایی که ارتباط زیادی با پژوهش ما دارد، محاسبه خطرات امنیتی برای برنامه‌های کاربردی تلفن همراه است [۲۴]. مطالعات متعددی باهدف ارائه معیارهای ریسک موثر برای تشخیص بدافزار اندروید از برنامه‌های معمولی انجام‌شده است [۲۵،۲۶،۲۷،۲۸]. این معیارها در درجه اول بر اساس مجوزهای استفاده‌شده یا عملکرد برنامه‌های Android است. به‌عنوان‌مثال، گیتس و همکاران. تعدادی از معیارهای ریسک مبتنی بر مجوز [۲۴] را معرفی کرد که RSS (سیستم امتیاز ریسک) یکی از قابل‌توجه‌ترین آن‌هاست. در [۲۵] روشی بر اساس بهره اطلاعاتی به‌منظور محاسبه خطر امنیتی بدافزارهای اندروید ارائه شد. دی پیر و همکاران[۲۶] توانستند به کمک روش نزدیک‌ترین همسایه، رویکردی را برای شناسایی حملات و برنامه‌های مخرب اندروید ارائه دهند. در این روش از فاصله همینگ به‌منظور محاسبه نزدیکترین همسایه‌ها استفاده می‌شد. در [۲۷]، یک معیار ریسک با استفاده از نمونه‌های قبلی برنامه‌های مخرب و غیر مخرب اندروید پیشنهاد شد. در این معیار کل فضای ویژگی بدافزارها و برنامه‌های مفید به‌منظور محاسبه دقیق‌تر خطر امنیتی مورداستفاده قرار می‌گرفت. علاوه بر این، یک اندازه‌گیری امتیاز ریسک مبتنی بر آنتروپی در [۲۸] ابداع شد که از کسب اطلاعات مجوزها و همچنین ویژگی‌های دیگر برای تشخیص برنامه‌های پرخطر اندروید از برنامه‌های معمولی، استفاده می‌کند.

محاسبه امتیاز ریسک برای URL ها مستلزم استخراج ویژگی‌های مناسب برای به دست آوردن مقادیر ریسک منطقی و دقیق است. همان‌طور که گفته شد، از ویژگی‌های URL مختلف برای ساخت مدل طبقه‌بندی استفاده‌شده است. ما از این ویژگی‌ها برای محاسبه ریسک‌های امنیتی واقع‌بینانه استفاده کرده‌ایم. بسیاری از روش‌ها URL های آلوده را بر اساس ویژگی‌های متنی موجود در URL ها شناسایی می‌کنند و از تفاوت‌های متن URL های مخرب در مقایسه با URL های معمولی استفاده می‌کنند [۲۹]. برخی از مطالعات همچنین ویژگی‌های دامنه یا سرور مانند IP، SSL و اطلاعات Whois را علاوه بر ویژگی‌های متنی URL ها در نظر می‌گیرند [۳۰]. حجاج و همکاران [۲] سعی کرده‌اند ویژگی‌های قوی برای شناسایی URLهای مخرب ارائه دهند، به این معنی که تغییر آن‌ها برای متجاوزان دشوار یا غیرممکن است تا نتوانند از مدل‌های یادگیری ماشین یا سایر راه‌حل‌های امنیتی عبور کنند. هامون و همکاران [۳۱] مجموعه‌ای از ویژگی‌های متنی را با استفاده از تحلیل متنی برای دسته‌بندی URLها به چهار کلاس ارائه می‌کند: هرزنامه، فیشینگ، بدافزار و تغییر ظاهر[۱]. نویسندگان در [۳۲] یک روش مهندسی ویژگی را برای بهبود دقت طبقه‌بندی با استفاده از مدل‌های محبوبی مانند k-نزدیک‌ترین همسایه، ماشین بردار پشتیبان و پرسپترون چندلایه، ارائه کردند. برخی از مطالعات بر روی ویژگی‌های استخراج‌شده از سوابق DNS[۲]، وب، میزبان‌ها و دامنه‌ها برای دستیابی به‌دقت طبقه‌بندی بهتر، تمرکز می‌کنند [۳۳،۳۴]. اخیراً چندین روش تخصصی برای تشخیص URL مخرب معرفی‌شده است [۳۵-۴۰]. این روش‌ها شامل یک الگوریتم بهینه‌سازی [۳۵]، یک مدل مفصل عصبی موازی [۳۶]، یک رویکرد XGBoost حساس به هزینه برای داده‌های نامتعادل [۳۷]، یک طبقه‌بندی‌کننده انجمنی [۳۸]، یک شبکه کانولوشن برگشتی چندلایه بهبودیافته [۳۹] و یک تکنیک ترکیب یادگیری عمیق و فیلتر بلوم [۴۰] هستند. درحالی‌که خدمات آنلاین تجاری یا محصولاتی وجود دارند که سعی می‌کنند URL های مخرب را با استفاده از لیست سیاه یا یادگیری ماشینی شناسایی کنند، آن‌ها معمولاً مقدار ریسک را محاسبه نمی‌کنند که نشان‌دهنده میزان مشکوک بودن باشد. به‌عنوان‌مثال می‌توان به سرویس مستقل جداسازی تهدید ایمیل [۴۱] و ریسک وب گوگل [۴۲] اشاره کرد که سطوح خطر گسترده‌ای را بر اساس نمایه‌های سوابق URL یا فهرست‌های

[۱] Defacement

[۲] Domain Name Service

سیاه اختصاص می‌دهند. مشکل محاسبه ریسک امنیتی URL ها برای اولین بار در [۴۳] موردبررسی قرار گرفت، جایی که دو معیار برای تخمین ریسک URL های ناشناخته پیشنهاد شد. یکی از این معیارها مبتنی بر ویژگی و دیگری مبتنی بر نمونه بود که کارایی قابل‌قبولی در تخصیص مقدار خطر امنیتی به URL ها داشتند.

۳- بیان مسئله

همان‌طور که گفته شد روش‌های مختلفی برای طبقه‌بندی URL ها به دسته‌های مخرب و غیر مخرب ارائه‌شده است. بااین‌حال، این روش‌ها اغلب از کاستی‌هایی رنج می‌برند. موفق‌ترین رویکردهای طبقه‌بندی، مانند SVM[۱]، Random Forest، درختان تصمیم، Naïve Bayes و مدل‌های مبتنی بر یادگیری عمیق، علی‌رغم استفاده از طیف گسترده‌ای از ویژگی‌های متنی و مبتنی بر میزبان، هنوز هم خطاهای طبقه‌بندی قابل‌توجهی در مواجهه با داده‌های واقعی دارند. علاوه بر این، نفوذ گران می‌توانند عملکرد مدل‌ها را از طریق تجزیه‌وتحلیل جعبه سفید یا جعبه سیاه، شناسایی کنند و سپس الگوهای URL خود را برای دورزن تشخیص آن‌ها، تغییر دهند. از سوی دیگر، درحالی‌که ایجاد یک لیست سیاه، روشی ساده و آسان برای پیاده‌سازی است، اما راه‌حل مؤثری نیست. زیرا URL نای مخرب جدید روزانه ظاهر می‌شوند که در لیست وجود ندارند. این به دلیل مشکل در نگهداری فهرست جامع و به‌روز از URLهای مخرب است که منجر به مشکل منفی‌های نادرست بالا می‌شود. علاوه بر این، بسیاری از مهاجمان تغییرات جزئی در URL اصلی ایجاد می‌کنند، مانند جایگزینی دامنه سطح بالا (TLD[۲])، استفاده از معادل‌سازی آدرس IP، تغییر ساختار دایرکتوری، جایگزینی رشته‌های پرس‌وجو یا استفاده از معادل نام تجاری که می‌تواند باعث ایجاد URL مخرب ناشناخته شود. علاوه بر این، استفاده از مجموعه‌ای از قوانین در فایروال‌ها برای فیلتر کردن URL های مخرب نیز مؤثر نیست، زیرا نفوذ گران می‌توانند این قوانین را شناسایی و دور بزنند. مهم‌تر از همه، در همه انواع روش‌های تشخیص URL مخرب، یک مشکل اساسی وجود دارد: یک URL ممکن است به‌شدت مخرب یا غیر مخرب نباشد، یا ممکن است تا حدی مخرب باشد. به‌عنوان‌مثال، ممکن است یک URL وب‌سایت تبلیغاتی توسط برخی از کاربران مخرب تلقی نشود، اما ممکن است همچنان نامطلوب تلقی شود؛ بنابراین، انواع مختلف URL های نامطلوب، ازجمله هرزنامه، فیشینگ، بدافزار و حمله تغییر ظاهر، همه به یک اندازه مخرب نیستند. درحالی‌که یک دسته‌بندی URL های مخرب بر اساس نوع حمله آن‌ها در [۳۱] معرفی شد، میزان خطر مرتبط با هر نوع، هنوز برای کاربر نامشخص است؛ بنابراین، ارائه مقدار ریسک امنیتی به‌جای طبقه‌بندی دقیق می‌تواند رویکرد منطقی و مؤثرتری برای تصمیم‌گیری صحیح باشد. مفهوم محاسبه ریسک امنیتی قبلاً در زمینه‌های دیگری نیز مانند شناسایی برنامه‌های مخرب موردمطالعه قرارگرفته است و معیارهای مختلفی نیز ارائه‌شده است [۲۶-۲۸]. در این مطالعه، ما بر روی مشکل تخمین مقدار ریسک امنیتی برای URLهای ناآشنا برای مقابله با چالش‌های فوق تمرکز می‌کنیم. به‌طور رسمی می‌توان گفت که یک نشانی اینترنتی یا URL که با u مشخص‌شده است را می‌توان با یک بردار ویژگی نشان دهیم، به‌طوری‌که خواهیم داشت:

$$u = \{f_1, f_2, \ldots, f_n\} \quad (1)$$

این ویژگی‌ها می‌توانند ویژگی‌های متنی (واژه‌ای) درون URL، یا ویژگی‌های دامنه، میزبان و سروری باشند که URL به آن اشاره می‌کند، یا ترکیبی از انواع ویژگی‌های مختلف. هدف، ارائه معیاری است که سطح خطر امنیتی آدرس u را بر اساس اطلاعاتی که قبلاً در مورد URL های مخرب و غیر مخرب شناخته‌شده داریم، محاسبه می‌کند؛ بنابراین، ما به دنبال یافتن تابعی به شکل R(u) هستیم که بردار ویژگی u را می‌گیرد و یک مقدار عددی مثبت پیوسته که به‌عنوان امتیاز ریسک امنیتی u شناخته می‌شود، برمی‌گرداند. ارزش ریسک محاسبه‌شده باید برای URL های شناخته‌شده قبلی معقول باشد. به‌عبارت‌دیگر، مقدار ریسک به‌دست‌آمده باید برای یک URL مخرب نسبتاً بزرگ‌تر از یک URL نرمال(مفید) باشد.

۴- روش پیشنهادی

روش ما برای به دست آوردن یک معیار اندازه‌گیری مقدار ریسک برای URL ها شامل پنج مرحله اصلی ازجمله جمع‌آوری و پیش‌پردازش داده‌ها، مهندسی ویژگی، آموزش مدل، مکانیسم امتیاز ریسک و پیاده‌سازی سیستم است. در جمع‌آوری داده‌ها و پیش‌پردازش، مجموعه داده‌ای جامع از URLهای مخرب و نرمال قبلاً شناخته‌شده را جمع‌آوری می‌کنیم. ابتدا داده‌ها را برای استخراج ویژگی‌ها و برچسب‌های مربوطه پیش‌پردازش می‌کنیم. در مهندسی ویژگی، تکنیک‌هایی برای استخراج و نمایش ویژگی‌ها از URLها به کار می‌رود. در مرحله سوم برای ساخت و آموزش مدل، یک مدل ریسک مؤثر را بر روی مجموعه داده برچسب‌گذاری شده، آموزش می‌دهیم. در مکانیسم امتیازدهی ریسک، یک متریک یا مکانیزم امتیازدهی ریسک بر اساس ویژگی‌های استخراج‌شده از نمونه‌های قبلاً شناخته‌شده URL ها، توسعه می‌یابد. در آخرین مرحله برای پیاده‌سازی و ارزیابی سیستم، ما یک سیستم خودکار را پیاده‌سازی می‌کنیم که مدل آموزش‌دیده را برای تجزیه‌وتحلیل URL های جدید و برآورد میزان ریسک امنیتی آن‌ها به کار می‌برد.

[۱] Support Vector Machine

[۲] Top Level Domain

در این تحقیق از تحلیل تفکیک خطی[1] به‌منظور توسعه مدلی برای تخمین امتیاز ریسک URL ها استفاده‌شده است. این تکنیک که به‌اختصار بنام LDA شناخته می‌شود در دسته‌بندی‌های دو کلاسه یا به‌عنوان روشی برای کاهش بعد، کاربرد دارد. هدف از تکنیک LDA این است که ماتریس داده اصلی را در فضایی با ابعاد پایین‌تر نشان دهد. برای رسیدن به این هدف باید سه مرحله انجام شود[۴۴] اولین مرحله محاسبه تفکیک‌پذیری بین کلاس‌های مختلف (فاصله بین میانگین‌های کلاس‌های مختلف) است که به آن واریانس بین کلاس یا ماتریس بین کلاس می‌گویند. این واریانس با فرمول زیر محاسبه می‌شود:

$$S_W = \sum_{i=1}^{I}(\sum_{j=1}^{n_i}(x_j^{(i)} - m^{(i)})(x_j^{(i)} - m^{(i)})^T) \quad (۲)$$

مرحله دوم محاسبه فاصله بین میانگین و نمونه‌های هر کلاس است که به آن واریانس درون کلاسی یا ماتریس درون کلاسی می‌گویند که به‌صورت زیر محاسبه می‌شود:

$$S_B = \sum_{i=1}^{I} n_i(m^{(i)} - m)(m^{(i)} - m)^T \quad (۳)$$

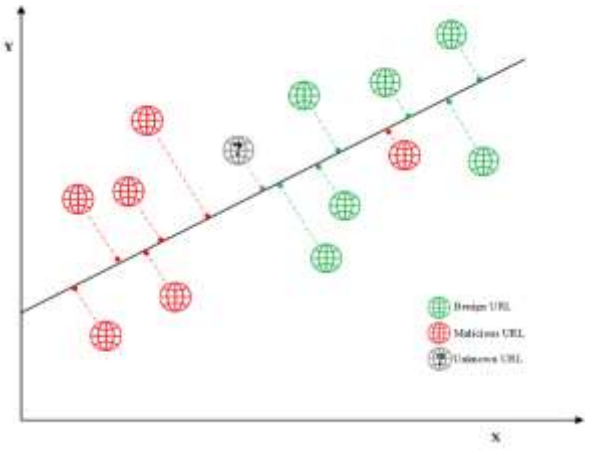
در فرمول‌های فوق I، n_i، x و m به ترتیب تعداد کلاس‌ها، تعداد نمونه‌های کلاس i ام، هر نمونه موجود و میانگین نمونه‌ها هستند. مرحله سوم ساخت فضا با ابعادی پایین‌تر است که واریانس بین کلاس را به حداکثر می‌رساند و هم‌زمان واریانس درون کلاس را به حداقل می‌رساند. به‌عبارت‌دیگر تابع هدف تحلیل تفکیکی خطی به‌صورت زیر است:

$$W_{opt} = argmax_w \frac{W^T S_B W}{W^T S_W W} \quad (۴)$$

البته در ما در اینجا از LDA به‌منظور دسته‌بندی یا کاهش بعد استفاده نمی‌کنیم؛ بلکه آن را برای تخمین خطر امنیتی نشانی‌های ناشناخته با استفاده از نمونه‌های مخرب و مفید قبلی به کار گرفته‌ایم. ایده اصلی ما در شکل (۱) نشان داده‌شده است. در این شکل برای سادگی، فضای دوبعدی نشان داده‌شده است اما در عمل می‌توان از بسیاری از ابعاد (ویژگی‌ها) استخراج‌شده از URLها استفاده کرد. همان‌طور که در شکل (۱) نشان داده‌شده است، با توجه به تجزیه‌وتحلیل تفکیک خطی، هر دو نمونه URL مخرب و مفید به فضای با بعد پایین‌تر نگاشت می‌شوند، به‌نحوی‌که در این فضا تفکیک‌پذیری افزایش می‌یابد. درواقع با توجه به فرمول‌های فوق، نگاشت نمونه‌ها به بعد پائین، سبب از دست رفتن اطلاعات مربوط به جدا پذیری نمونه‌ها نمی‌شود و عمل کاهش بعد با افزایش تفکیک‌پذیری نمونه‌های مخرب و مفید همراه است. معیار ارائه‌شده در اینجا پارامترهای لازم برای تخمین ریسک در فضای جدید را محاسبه می‌کند. این پارامترها مربوط به فاصله URL ناشناخته تا نمونه‌های مخرب و ایمن(نرمال)، در فضای نگاشت شده، است. این معیار درنهایت با کمک پارامترهای به‌دست‌آمده، میزان ریسک را تخمین می‌زند. درواقع، فضای متمایزکننده جدید با استفاده از تجزیه‌وتحلیل تفکیک خطی URLهای مخرب و ایمن موجود به دست می‌آید. تصویر نمونه‌های قبلی به فضای جدید به این دلیل است که در فضای جدید این دو نوع نمونه از هم متمایزتر هستند و می‌توان ارزش ریسک URL ناشناخته را با دقت بیشتری محاسبه کرد. به‌طورکلی، خطر امنیتی یک نمونه ناشناخته بافاصله آن از نمونه‌های URL مخرب رابطه معکوس دارد. زیرا هرچه فاصله با نمونه‌های مخرب کمتر باشد، خطر بالاتری دارد. از طرفی ریسک امنیتی این نمونه ارتباط مستقیمی بافاصله با نمونه‌های ایمن دارد، یعنی هر چه فاصله با نمونه‌های ایمن کمتر باشد، ریسک کمتری دارد. این ایده هم در فضای اصلی و هم در فضای پیش‌بینی‌شده، معتبر است. برای محاسبه ریسک در فضای جدید که تفکیک‌پذیری بیشتری دارد، پارامترهای موردنیاز یعنی فواصل، به‌دست‌آمده و درنهایت مقادیر به‌دست‌آمده با یک معادله ساده به‌صورت زیر برای محاسبه خطر امنیتی، به کار می‌روند:

$$Risk(u) = \frac{ds}{dm} \quad (۵)$$

معادله (۵) خطر امنیتی یک URL ناشناخته u را تخمین می‌زند. در این فرمول، ds و dm به ترتیب فاصله URLهای نگاشت شده u تا میانگین URL های ایمن و میانگین URLهای مخرب در این فضای جدید هستند.



شکل (۱): نحوه محاسبه امتیاز ریسک یک URL ناشناخته با استفاده از URLهای مفید و مخرب قبلی

شکل (۲) شبه کد الگوریتم پیشنهادی LRU (تحلیل تفکیکی خطی برای تخمین ریسک URLها) را نشان می‌دهد. این الگوریتم مجموعه URLهای شناخته‌شده قبلی (SUs) و URL ناشناخته u را به‌عنوان ورودی دریافت می‌کند. هدف تخمین ریسک امنیتی URL ناشناخته u است. جدول (۱) نمادهای استفاده‌شده در الگوریتم را نشان می‌دهد.

در خط ۱، داده‌های ورودی، نرمال شده‌اند. در خطوط ۲ و ۳ به ترتیب URLهای مخرب و ایمن از لیست کلی URLها استخراج می‌شوند. مقادیر میانگین این دو مجموعه بر این اساس در خطوط ۴ و ۵ محاسبه می‌شوند. در خطوط ۶ و ۷ به ترتیب ماتریس پراکندگی درون کلاسی، برای هر دو کلاس (مجموعه)

[1] Linear Discriminant Analysis (LDA)

URLهای مخرب و ایمن به‌دست‌آمده است. پس از محاسبه ماتریس پراکندگی بین کلاسی در خط ۸، ماتریس تبدیل LDA در خط ۹ محاسبه می‌شود. در خطوط ۱۰ و ۱۱، مجموعه‌های نگاشت شده از URLهای مخرب و ایمن، توسط تبدیل LDA به دست می‌آیند. سپس در خط ۱۲ نگاشت URL ورودی ناشناخته u نیز توسط همین تبدیل محاسبه می‌شود. پس‌ازآن، مقادیر میانگین دو مجموعه نگاشت شده به ترتیب در خطوط ۱۳ و ۱۴ محاسبه می‌شوند. در خطوط ۱۵ و ۱۶، فواصل نمونه ورودی u تا هر دو مقدار میانگین، محاسبه می‌شوند. این فواصل برای تخمین مقادیر ریسک URL ورودی ناشناخته u موردنیاز هستند. پس از به دست آوردن پارامترهای موردنیاز، درنهایت، در خط ۱۷، ریسک امنیتی URL ورودی u با استفاده از فرمول (۵) همان‌طور که قبلاً توضیح داده شد، محاسبه می‌شود.

۵- ارزیابی و مقایسه کارایی

ما الگوریتم مربوط به معیار پیشنهادی LRU خود را با استفاده از نرم‌افزار Matlab۲۰۲۳ و زبان برنامه‌نویسی آن پیاده‌سازی کرده‌ایم. برای افزایش سرعت محاسبات، به‌جای محاسبه ریسک برای هر URL به‌صورت جداگانه طبق رویکرد نشان داده‌شده در شکل (۲)، ما محاسبات مبتنی بر ماتریس را در Matlab انجام می‌دهیم تا امتیاز ریسک مجموعه‌ای از URLهای ناشناخته را به‌طور هم‌زمان محاسبه کنیم. ما از دو مجموعه داده در دسترس عموم برای ارزیابی کارایی معیار LRU پیشنهادی در سناریوهای مختلف استفاده می‌کنیم. جدول(۲) خصوصیات کلیدی این مجموعه داده‌ها را خلاصه می‌کند. مجموعه داده اول فقط شامل ویژگی‌های متنی(واژگانی) است ولی مجموعه داده دوم حاوی انواع ویژگی‌های واژگانی و مبتنی بر میزبان است.

***Procedure** LRU (SUs, u)*

begin

۱. $SUs = normalize\ (SUs);$
۲. $MS = \{\forall\ malicious\ \ su \in SUs\};$
۳. $SS = \{\forall\ safe\ \ su \in SUs\};$
۴. $mu = mean(MS);$
۵. $ms = mean(SS);$
۶. $SW_m = \sum_{x_i \in MS}(x_i - mu) \times (x_i - mu)^T;$
۷. $SW_S = \sum_{x_i \in SS}(x_i - ms) \times (x_i - ms)^T;$
۸. $SW = SW_m + SW_S;$
۹. $W = SW^{-1} \times (mu - ms);$
۱۰. $Ym = W^T \times MS;$
۱۱. $Ys = W^T \times SS;$
۱۲. $yu = W^T \times u;$
۱۳. $mYm = mean(Ym);$
۱۴. $mYs = mean(Ys);$
۱۵. $dm = |yu - mYm|;$
۱۶. $ds = |yu - mYs|;$
۱۷. ***return*** $\frac{ds}{dm}$;

end

شکل(۲): الگوریتم LRU

جدول(۱): نمادهای مورداستفاده و معنای آن‌ها در الگوریتم LRU.

نماد	مفهوم
SUs	مجموعه *URL*های شناخته‌شده موجود (مضر و ایمن)
MS	مجموعه *URL*های مخرب
SS	مجموعه *URL*های ایمن
mu	مقدار متوسط برای مجموعه*URL* های مخرب
ms	مقدار متوسط برای مجموعه*URL* های امن
SW_m	ماتریس پراکندگی درون کلاسی برای*URL* های مخرب
SW_s	ماتریس پراکندگی درون کلاسی برای*URL* های ایمن
SW	جمع ماتریس‌های پراکندگی درون کلاسی
W	ماتریس پراکندگی بین کلاسی
Ym	مجموعه‌ای از*URL* های مخرب نگاشت شده
Ys	مجموعه‌ای از*URL* های ایمن نگاشت شده
yu	نگاشت *URL* ورودی u
mYm	میانگین مقدار*URL* های مخرب نگاشت شده
mYs	میانگین مقدار*URL* های ایمن نگاشت شده
dm	فاصله بین *URL* نگاشت شده u وmYm
ds	فاصله بین *URL* نگاشت شده u وmYs

جدول(۲). ویژگی‌های مجموعه داده‌های مورداستفاده

مجموعه داده	Textual Kaggle	OpenINTEL
تعداد کل URLها	۴۵۰۱۷۶	۹۳۲۰۷۸
تعداد URLهای مخرب	۱۰۴۴۳۸	۵۸۷۲۵۷
تعداد URLهای ایمن	۳۴۵۷۳۸	۳۴۴۸۲۱
تعداد ویژگی‌ها	۱۸	۶
توصیف	ویژگی‌های متنی استخراج‌شده از URLهای مفید و مخرب موجود در سایت [Kaggle ۴۵]	تعداد بالایی از URLهای مخرب و مفید] ۴۶[

ما طیف گسترده‌ای از ویژگی‌های متنی (واژگانی) را با استفاده از یک اسکریپت پایتون برای اولین مجموعه داده‌ای که از وب‌سایت Kaggle به‌دست‌آمده است، پیش‌پردازش و استخراج کرده‌ایم. ویژگی‌های متنی یا واژگانی رایج‌ترین ویژگی‌هایی هستند که برای رویکردهای مختلف شناسایی URL مخرب استفاده می‌شوند. پس از حذف اطلاعات استفاده‌نشده یا اضافی از فایل ورودی، اسکریپت نوشته‌شده فایل ورودی را که شامل تعداد زیادی URL مخرب و غیر مخرب است را می‌گیرد و پردازش متنی را برای استخراج ویژگی‌ها انجام می‌دهد. این اسکریپت ابتدا داده‌ها را با حذف اطلاعات استفاده‌نشده؛ مانند شماره ردیف‌ها و فاصله‌های اضافی، پیش‌پردازش می‌کند. سپس، توکن‌سازی را برای تسهیل استخراج ویژگی انجام می‌دهد. در مرحله اول استخراج ویژگی، اسکریپت، طول URL کامل، طول نام میزبان، طول مسیر و طول اولین زیرشاخه را اندازه می‌گیرد.

همچنین کاراکترهای ویژه هر URL ازجمله کاراکترهای خاصی مثل -,@,#,?,.,/,:,= را می‌شمارد. علاوه بر این، تعداد ارقام، حروف و دایرکتوری‌های فرعی را می‌شمارد. متعاقباً، نوع پروتکل (یعنی http یا https)، وجود عبارت «www»، حاوی آدرس IP بودن و استفاده از سرویس کوتاه کننده را تشخیص می‌دهد. URLهایی که از سرویس‌های کوتاه کننده استفاده می‌کنند، عبارات خاصی در متن خوددارند که در اسکریپت فهرست شده‌اند. به‌طور خلاصه، هجده ویژگی از این مجموعه داده استخراج شد. ما دومین مجموعه داده [۴۶] را از GitHub[۱]، به دست آوردیم. این بزرگ‌ترین مجموعه داده مورداستفاده در آزمایش‌های ماست. برخلاف مورد اول، در این مجموعه داده تعداد URLهای مخرب بیشتر از تعداد URL های مفید است. مجموعه داده‌های استفاده‌شده و پیاده‌سازی‌های ما برای معیار پیشنهادی و معیارهای ریسک قبلی، به‌صورت رایگان برای محققین در دسترس هستند[۲].

ما آزمایش‌هایی را روی مجموعه داده‌های اول و دوم انجام دادیم تا عملکرد تشخیص متریک LRU پیشنهادی را با دو معیار پیشنهادی قبلی، FRU و NRU، برای اندازه‌گیری امتیاز ریسک URL مقایسه کنیم [۴۳]. علاوه بر این، برای یک تجزیه‌وتحلیل مقایسه‌ای جامع‌تر، ما همچنین سه معیار دیگر را که در ابتدا برای اندازه‌گیری امتیاز ریسک برنامه اندروید پیشنهادشده بود را در مقایسه‌ها استفاده کردیم. این معیارهای عبارت‌اند از: RSS [۲۶]، IRS [۲۷] و ERS [۲۸]. این سه معیار در این پژوهش برای محاسبه ریسک URL تطبیق داده شدند. تنها تفاوت این است که معیارهای امتیاز ریسک برنامه بر روی مقادیر ویژگی‌های دودویی کار می‌کنند، درحالی‌که در معیارهای پیشنهادی، مقادیر ویژگی برای URLها و دامنه‌ها می‌توانند به‌صورت پیوسته یا دودویی باشند. بنابراین، ما به‌سادگی این سه معیار را برای مطالعه تطبیقی، پیاده‌سازی و به کار گرفته شد . درمجموع می‌توانیم معیار پیشنهادی خود را با پنج معیار محاسبه ریسک قبلی مقایسه کنیم.

برای ارزیابی هر معیار ریسک، همه بردارهای URL مخرب و عادی را در یک لیست قرار می‌دهیم. ما از روش اعتبارسنجی متقابل ۱۰ بخشی[۳] برای ارزیابی استفاده می‌کنیم. به این صورت که در هر بخش، معیار امتیاز پیشنهادی را با استفاده از ۹۰٪ URLها به‌عنوان داده‌های آموزشی و ۱۰٪ باقی‌مانده به‌عنوان داده‌های آزمایشی به کار می‌بریم. با استفاده از هر معیار ریسک، مقادیر ریسک همه URLهای موجود در لیست آزمایشی را به‌طور جداگانه تخمین می‌زنیم و لیست را به ترتیب نزولی مقادیر ریسک امنیتی محاسبه‌شده، مرتب می‌کنیم. هرچه URLهای مخرب بیشتری در بالای لیست مرتب‌شده قرار گیرند، معیار خطر امنیتی، قوی‌تر است. برای اندازه‌گیری عملکرد برای مقادیر مختلف درصد، بالاترین URLها ازنظر امتیاز خطر امنیتی را از فهرست مرتب‌شده انتخاب می‌کنیم. پس‌ازآن، برای فهرست مرتب‌شده هر متریک، مقدار درصد URLهای مخرب موجود در قسمت انتخاب‌شده را تعیین می‌کنیم. درنتیجه، درصد انتخاب هر لیست مرتب و درصد شناسایی‌شده از URLنای مخرب به ترتیب به‌عنوان نرخ هشدار و نرخ تشخیص نام‌گذاری می‌شوند. تعداد موارد مثبت کاذب و مثبت واقعی به ترتیب با میزان هشدار و تشخیص نسبت مستقیم دارند. اگرچه محدوده‌های ارزش ریسک محاسبه‌شده برای معیارهای مقایسه شده متفاوت است، اما می‌توان آن‌ها را با این روش به‌صورت عادلانه مقایسه کرد زیرا آستانه نرخ هشدار مطلق در اینجا استفاده نمی‌شود. درنتیجه، فرآیند ارزیابی ما مستقل از محدوده ارزش ریسک معیارها است. واضح است که با معیار ریسک قوی‌تر، تعداد بیشتری از URLهای مخرب در بالای لیست مرتب‌شده قرار می‌گیرند و بنابراین معیار، نرخ تشخیص بیشتری دارد. به‌عبارت‌دیگر، کاربر نهایی انتظار دارد که URL هایی با بالاترین امتیاز ریسک، مخرب باشند و ایمن نباشند. ما معیارهای ابداع‌شده و پنج معیار قبلی را با استفاده از دو مجموعه داده نشان داده‌شده در جدول (۲) باهم مقایسه می‌کنیم. شکل (۳) نتایج به‌دست‌آمده را برای هر دو مجموعه داده، در قالب نمودارهای ROC نشان می‌دهد. همان‌طور که در این شکل نشان داده‌شده است، معیار LRU پیشنهادی ازنظر نرخ تشخیص بهتر از سایر معیارها است.

برای اولین مجموعه داده، شکاف عملکرد بین معیارهای پیشنهادی و موارد قبلی قابل‌توجه است. درواقع معیار پیشنهادی، هم در سطوح هشدار کم و هم سطوح هشدار بالاتر، دارای نرخ تشخیص بسیار بهتری است که نشان‌دهنده تخمین دقیق خطر امنیتی در این مجموعه داده است. استفاده از تحلیل تفکیکی خطی سبب می‌شود نمونه‌های مفید و مخرب قبلی جدا پذیری بیشتری پیدا کنند و مقدار واقع‌بینانه‌تری برای خطر امنیتی نمونه‌های ناشناخته به دست آید. در این مجموعه داده، FRU و NRU به ترتیب رتبه‌های دوم و سوم رادارند. از بین این دو، NRU پایداری بیشتری در سطوح هشدار مختلف نشان می‌دهد. در مجموعه داده دوم، پس از LRU پیشنهادی، NRU و RSS ازنظر عملکرد به ترتیب در رتبه‌های دوم و سوم قرار دارند. همچنین با دقت در این منحنی‌ها می‌توان دریافت که روش پیشنهادی علاوه بربرتری دارای عملکرد یکنواخت‌تری در سطوح هشدار مختلف است. نتایج به‌دست‌آمده برای نرخ تشخیص، روی هر دو مجموعه داده، نشان می‌دهد روش پیشنهادی LRU برای نمونه‌های مفید مقدار خطر امنیتی کمتر و برای نمونه‌های مخرب خطر امنیتی بالاتری را محاسبه می‌کند.

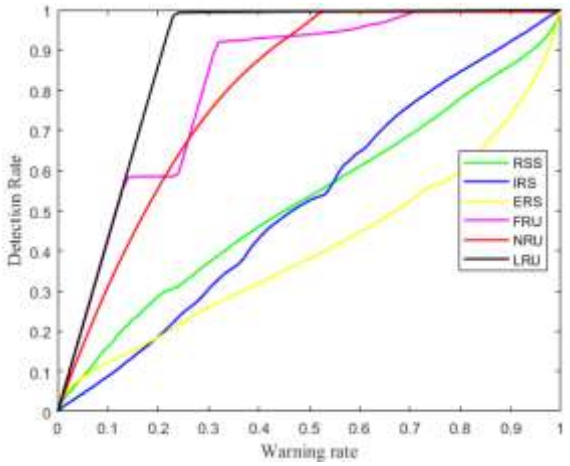
برای ارزیابی جامع‌تر در همه مقادیر سطح هشدار، شکل (۴) سطح کل زیر منحنی (AUC) همه معیارها را مقایسه می‌کند.

[۱] https://github.com/slrbl/malicious-urls-detection-with-autoencoder-neural-networks

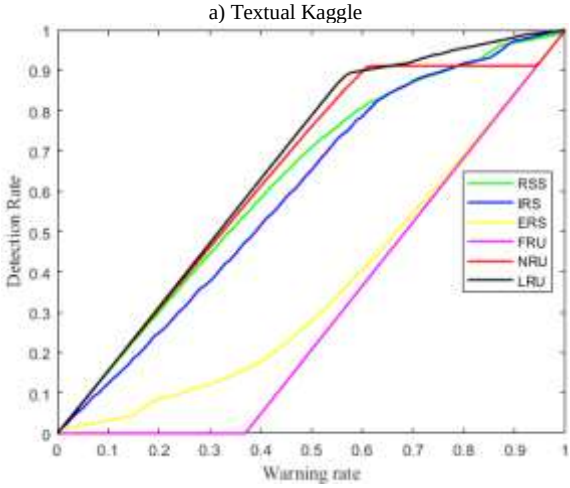
[۲] https://github.com/mdeypir/LRU

[۳] ۱۰-fold Cross Validation

به‌طور طبیعی، یک متریک با AUC نزدیک به یک، معیار خوبی به‌عنوان تخمین گر ریسک URL است.



a) Textual Kaggle

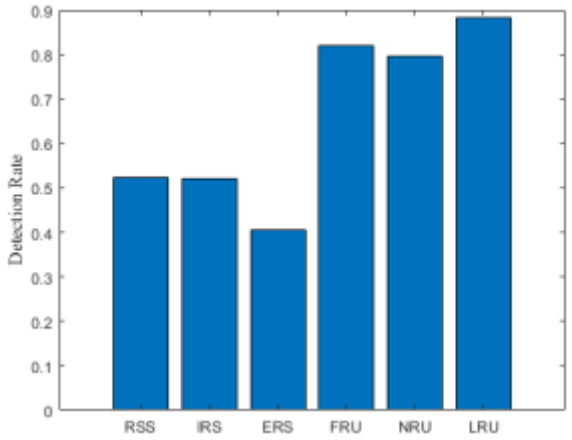


b) OpenINTEL

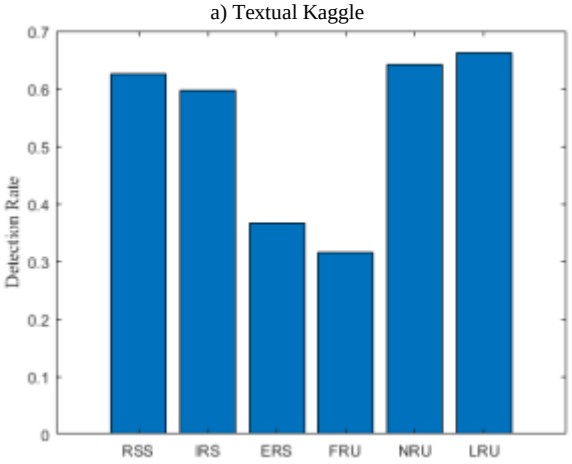
شکل (۳): مقایسه نرخ تشخیص معیارهای امتیاز ریسک پیشنهادی و معیارهای قبلی توسط نمودار ROC.

همان‌طور که در شکل (۴) نشان داده‌شده است، معیارهای LRU پیشنهادی دارای AUCهای بزرگ‌تری برای هر دو مجموعه داده است زیرا نرخ تشخیص بالاتری در مقادیر مختلف سطح هشدار دارد. در آزمایش بعدی، نمرات دقت، یادآوری و نمره F۱ تمام معیارهای محاسبه ریسک در هر دو مجموعه داده و برای همه معیارها محاسبه‌شده است. برای این منظور ابتدا باید مقادیر مثبت درست (True Positive)، منفی کاذب (False Negative) و مثبت کاذب (False Positive) محاسبه شوند. مثبت درست نشان‌دهنده URLهای مخربی است که با موفقیت شناسایی‌شده‌اند. مثبت کاذب نشان‌دهنده تعداد URLهای مفیدی است که به‌اشتباه به‌عنوان مخرب شناسایی‌شده‌اند. منفی‌های کاذب تعداد URLهای مخربی را نشان می‌دهد که به‌اشتباه به‌عنوان نشانی‌های مفید شناسایی‌شده‌اند. دقت (Precision) بیانگر نسبت تعداد نشانی‌های مخرب شناسایی‌شده درست به تعداد کل همه نشانی‌های مخرب شناسایی‌شده (درست یا غلط) است. معادله این نسبت به‌صورت زیر نشان داده‌شده است:

$$Percision = \frac{TP}{TP + FP} \quad (۵)$$



a) Textual Kaggle



b) OpenINTEL

شکل (۴): مقایسه مساحت زیر منحنی (AUC)

یادآوری(Recall) بیانگر نسبت تعداد نشانی‌های مخربی است که با موفقیت شناسایی شدند به تعداد همه نشانی‌های مخرب که با موفقیت شناسایی شدند یا به‌طور ناموفق به‌عنوان نشانی بی‌ضرر تشخیص داده شدند:

$$Recall = \frac{TP}{TP + FN} \quad (۶)$$

نمره F۱(F۱-Score) میانگین هارمونیک از دقت و یادآوری است. این مقدار نشان می‌دهد که معیار پیشنهادی ریسک چقدر توانایی عمل در تمایز بین دو کلاس مخرب و بی‌خطر را دارد:

$$\text{F۱-Score} = ۲ \times \frac{Percision \times Recall}{Percision + Recall} \quad (۷)$$

اگرچه فرمول‌های فوق اصولاً برای ارزیابی طبقه‌بندی‌ها به کار می‌روند، ما در اینجا از آن‌ها به‌منظور مقایسه توانایی معیارهای محاسبه خطر، استفاده کرده‌ایم. نتایج به‌دست‌آمده در جدول (۳) گزارش‌شده‌اند. همان‌طور که در این جدول مشاهده می‌شود، برای هر دو مجموعه داده، روش پیشنهادی LRU دقت، یادآوری و نمره F۱ خوبی دست می‌آورد. به‌ویژه در مجموعه داده اول برتری عددی روش پیشنهادی بسیار چشمگیر است. اما در مجموعه داده

دوم روش پیشنهادی با اختلاف کمی رتبه دوم را به دست آورده است. دلیل کارایی خوب روش پیشنهادی این است که مقدار ارزش ریسک واقع‌بینانه‌تری را برای نمونه‌های ناشناخته تخمین می‌زند. بنابراین در مقایسه با سایر معیارهای اندازه‌گیری ریسک، مثبت‌های واقعی بالاتر، مثبت‌های کاذب کمتر، منفی‌های واقعی بیشتر و منفی‌های کاذب کمتری به دست می‌آیند.

جدول (۳): مقایسه دقت، یادآوری، و نمره F۱

مجموعه داده	معیار	دقت	یادآوری	نمره *F۱*
Textual Kaggle	***RSS***	۰.۳۰۸۰	۰.۳۰۷۹	۰.۳۰۸۰
	IRS	۰.۲۲۵۶	۰.۲۲۵۵	۰.۲۲۵۶
	ERS	۰.۲۰۹۸	۰.۲۰۹۷	۰.۲۰۹۹
	FRU	۰.۵۸۵۸	۰.۵۸۵۷	۰.۵۸۵۸
	NRU	۰.۶۲۳۲	۰.۶۲۳۱	۰.۶۲۳۲
	LRU	۰.۹۹۱۰	۰.۹۹۰۹	۰.۹۹۱۰
OpenINTEL	***RSS***	۰.۸۲۸۱	۰.۸۲۸۱	۰.۸۲۸۱
	IRS	۰.۸۲۴۷	۰.۸۲۴۶	۰.۸۲۴۷
	ERS	۰.۴۴۵۱	۰.۴۴۵۰	۰.۴۴۵۱
	FRU	۰.۴۱۲۸	۰.۴۱۲۷	۰.۴۱۲۸
	NRU	۰.۹۱۱۴	۰.۹۱۱۳	۰.۹۱۱۴
	LRU	۰.۹۰۶۱	۰.۹۰۶۰	۰.۹۰۶۱

بر اساس آزمایش‌ها، نتیجه می‌گیریم که معیار برنده، معیار LRU پیشنهادی است، زیرا بهترین عملکرد را در همه موقعیت‌ها با هر دو مجموعه داده و تقریباً همه نرخ‌های هشدار دارد. دلیل برتری معیار پیشنهادی نسبت به سایر معیارها را می‌توان در تحلیل تفکیک خطی جستجو کرد که منجر به نگاشت نمونه‌ها در فضای جدید می‌شود. در این فضا، نمونه‌های مخرب و ایمن، بیشتر از هم تفکیک‌شده و درنهایت برآورد دقیق‌تری از ریسک امنیتی به دست می‌آید.

۶- بحث و نتیجه‌گیری

در این پژوهش، به‌جای طبقه‌بندی دقیق، تمرکز بر تخمین امتیاز ریسک امنیتی برای URLهای ناشناخته بود. معیاری به نام LRU (تحلیل تفکیکی خطی برای تخمین ریسک URL ها) برای حل این مسئله پیشنهاد شد. برای محاسبه معیار بر روی داده‌های واقعی، الگوریتم موردنیاز طراحی و پیاده‌سازی شد. این الگوریتم می‌تواند با محاسبه ریسک بالا برای URLهای مخرب و اعلام هشدار به کاربر، از کلاه‌برداری‌های اینترنتی پیشگیری کند. درصورتی‌که روش پیشنهادی به‌عنوان یک افزونه به مرورگرها اضافه شود و یا در شبکه‌های اجتماعی و سیستم‌های تبادل پیام به کار رود می‌تواند در اعلام هشدار به کاربران بسیار مؤثر باشد و جلوی بسیاری از نفوذها را بگیرد؛ زیرا بسیاری از روش‌های نفوذ با ارسال یک URL از سوی مهاجم شروع می‌شوند.

آزمایش‌های انجام‌شده، اثربخشی معیار پیشنهادی LRU را ازنظر میزان تشخیص نشان می‌دهند. نتایج نشان می‌دهد که LRU می‌تواند مقادیر بالای ریسک را به URLهای مخرب و مقادیر کم را به آدرس‌های غیرمخرب اختصاص دهد. این به دلیل مدل ریاضی بهتر آن است که از فضای ویژگی‌های نگاشت شده URL های مخرب و غیر مخرب قبلاً شناخته‌شده، برای بهبود تخمین ریسک URLهای ناشناخته استفاده می‌کند. همان‌طور که نتایج آزمایش‌ها نشان داد روش پیشنهادی بدون استفاده از یادگیری عمیق، کارایی خوبی از خود نشان داده است. از طرفی با استفاده از داده‌های کمی می‌توان مدل ساده پیشنهادی را آموزش داد. به‌طورکلی روش تحلیل تفکیکی خطی(LDA) که روش پیشنهادی ما مبتنی بر آن است دارای سه مشکل اصلی است که این مشکلات عبارت‌اند از مشکل اندازه نمونه کوچک، حساسیت به نویز و نقاط پرت، و ناتوانی در برخورد با داده‌های کلاس چندوجهی(Multi Modal classes) که می‌تواند در شرایطی سبب کاهش کارایی آن شود. البته برای حل یا تعدیل این مشکلات راه‌حل‌هایی نیز وجود دارد که تقسیم کلاس‌ها به زیر کلاس‌های مختلف از جمله این راه‌حل‌هاست که تفکیک‌پذیری داده‌ها در ساختارهای پیچیده را بیشتر می‌کند] ۴۶] روش‌های مبتنی بر یادگیری عمیق ممکن است کارایی بهتری داشته باشند؛ ولی علاوه بر نیاز به داده‌های زیاد، زمان آموزش بیشتری را نیاز خواهند داشت. در آینده قصد داریم محاسبه خطر امنیتی را با استفاده از یادگیری عمیق انجام دهیم و روشی را در این زمینه ارائه دهیم. همچنین استفاده از روش‌های متنوع‌تر استخراج ویژگی به‌منظور تخمین دقیق‌تر خطر امنیتی نشانی‌های اینترنتی از دیگر کارهای آینده است.

۷- مراجع

[۱] P. S. Pakhare, S. Krishnan, N. N. Charniya, "Malicious url detection using machine learning and ensemble modeling," In Computer Networks, Big Data and IoT: Proceedings of ICCBI ۲۰۲۰, pp. ۸۳۹-۸۵۰, Springer Singapore, ۲۰۲۰, doi: https://doi.org/۱۰.۱۰۰۷/۹۷۸-۹۸۱-۱۶-۰۹۶۵-۷_۶۵

[۲] C. Hajaj, N. Hason, and A. Dvir, "Less is more: Robust and novel features for malicious domain detection," Electronics, vol. ۱۱, no. ۶, p. ۹۶۹, ۲۰۲۲, https://doi.org/۱۰.۳۳۹۰/electronics۱۱۰۶۰۹۶۹.

[۳] S. Kim, J. Kim, and B. B. Kang, "Malicious URL protection based on attackers' habitual behavioral analysis," Computers & Security, vol. ۷۷, pp. ۷۹۰-۸۰۶, ۲۰۱۸, https://doi.org/۱۰.۱۰۱۶/j.cose.۲۰۱۸.۰۱.۰۱۳.

[۴] A. S. Raja, G. Pradeepa, and N. Arulkumar, "Mudhr: Malicious URL detection using heuristic rules based

approach," In AIP Conference Proceedings, vol. ۲۳۹۳, no. ۱, p. ۰۲۰۱۷۶, AIP Publishing LLC, ۲۰۲۲, https://doi.org/۱۰.۱۰۶۳/۵.۰۰۷۴۰۷۷

[۵] R. Madhubala, N. Rajesh, L. Shaheetha, and N. Arulkumar, "Survey on Malicious URL Detection Techniques," In ۲۰۲۲ ۶th International Conference on Trends in Electronics and Informatics (ICOEI), pp. ۷۷۸-۷۸۱, IEEE, ۲۰۲۲, doi: ۱۰.۱۱۰۹/ICOEI۵۳۵۵۶.۲۰۲۲.۹۷۷۷۲۲۱.

[۶] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust intelligent malware detection using deep learning," IEEE Access, vol. ۷, pp. ۴۶۷۱۷-۴۶۷۳۸, ۲۰۱۹, doi: ۱۰.۱۱۰۹/ACCESS.۲۰۱۹.۲۹۰۶۹۳۴.

[۷] Y. Liang, Q. Wang, K. Xiong, X. Zheng, Z. Yu, and D. Zeng, "Robust Detection of Malicious URLs With Self-Paced Wide & Deep Learning," IEEE Transactions on Dependable and Secure Computing, vol. ۱۹, no. ۲, pp. ۷۱۷-۷۳۰, ۲۰۲۱, doi: ۱۰.۱۱۰۹/TDSC.۲۰۲۱.۳۱۲۱۳۸۸.

[۸] R. Rakesh, S. Muthuraijkumar, L. Sairamesh, M. Vijayalakmi, and A. Kannan, "Detection of URL based attacks using reduced feature set and modified C۴. ۵ algorithm," Adv. Nat. Appl. Sci, vol. ۹, pp. ۳۰۴-۳۱۱, ۲۰۱۵.

[۹] F. A. Ghaleb, M. Alsaedi, F. Saeed, J. Ahmad, and M. Alasli, "Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning," Sensors, vol. ۲۲, no. ۹, p. ۳۳۷۲, ۲۰۲۲, doi: ۱۰.۳۳۹۰/s۲۲۰۹۳۳۷۲.

[۱۰] S. He, B. Li, H. Peng, J. Xin, and E. Zhang, "An effective cost-sensitive XGBoost method for malicious URLs detection in imbalanced dataset," IEEE Access, vol. ۹, pp. ۹۳۰۸۹-۹۳۰۹۶, ۲۰۲۱, doi: ۱۰.۱۱۰۹/access.۲۰۲۱.۳۰۹۳۰۹۴.

[۱۱] R. Patgiri, H. Katari, R. Kumar, and D. Sharma, "Empirical study on malicious URL detection using machine learning," In International Conference on Distributed Computing and Internet Technology, pp. ۳۸۰-۳۸۸, Springer, Cham, ۲۰۱۹, https://doi.org/۱۰.۱۰۰۷/۹۷۸-۳-۰۳۰-۰۵۳۶۶-۶_۳۱.

[۱۲] J. Chen, Z. Hu, and Z. Qian, "Research on malicious URL detection based on random forest," In ۲۰۲۲ ۱۴th International Conference on Computer Research and Development (ICCRD), pp. ۳۰-۳۶, IEEE, ۲۰۲۲, January, doi: ۱۰.۱۱۰۹/iccrd۵۴۴۰۹.۲۰۲۲.۹۷۳۰۴۵۱.

[۱۳] C. Ding, "Automatic detection of malicious urls using fine-tuned classification model," In ۲۰۲۰ ۵th International Conference on Information Science, Computer Technology and Transportation (ISCTT), pp. ۳۰۲-۳۲۰, IEEE, ۲۰۲۰, doi: ۱۰.۱۱۰۹/ISCTT۵۱۵۹۵.۲۰۲۰.۰۰۰۶۰.

[۱۴] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Evaluating deep learning approaches to characterize and classify malicious URL's," Journal of Intelligent & Fuzzy Systems, vol. ۳۴(۳), pp. ۱۳۳۳-۱۳۴۳, ۲۰۱۸, DOI: ۱۰.۳۲۳۳/JIFS-۱۶۹۴۲۹.

[۱۵] J. Yuan, Y. Liu, and L. Yu, "A novel approach for malicious url detection based on the joint model," Security and Communication Networks, p. ۴۹۱۷۰۱۶, ۲۰۲۱, https://doi.org/۱۰.۱۱۵۵/۲۰۲۱/۴۹۱۷۰۱۶.

[۱۶] P. L. Indrasiri, M. N. Halgamuge, and A. Mohammad, "Robust Ensemble Machine Learning Model for Filtering Phishing URLs: Expandable Random Gradient Stacked Voting Classifier," (ERG-SVC). IEEE Access, vol. ۹, pp. ۱۵۰۱۴۲-۱۵۰۱۶۱, ۲۰۲۱, doi: ۱۰.۱۱۰۹/access.۲۰۲۱.۳۱۲۴۶۲۸.

[۱۷] D. R. Patil, and J. B. Patil, "Malicious URLs detection using decision tree classifiers and majority voting technique," Cybernetics and Information Technologies, vol. ۱۸, no. ۱, pp. ۱۱-۲۹, ۲۰۱۸, doi: ۱۰.۲۴۷۸/cait-۲۰۱۸-۰۰۰۲.

[۱۸] D. K. Mondal, B. C. Singh, H. Hu, S. Biswas, Z. Alom, and M. A. Azim, "SeizeMaliciousURL: A novel learning approach to detect malicious URLs," Journal of Information Security and Applications, vol. ۶۲, ۱۰۲۹۶۷, ۲۰۲۱, https://doi.org/۱۰.۱۰۱۶/j.jisa.۲۰۲۱.۱۰۲۹۶۷.

[۱۹] A. Shahidinejad, M. Torabi, "Detection and Prevention of SQL Injection Attacks at Runtime Using Decision Tree Classification," Electronic and Cyber Defense, vol. ۸, no. ۴, pp. ۷۵-۹۳, ۲۰۲۱, dor: ۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۳۹۹.۸.۴.۷.۳ (in persian)

[۲۰] M. Maghale, M. Bagheri, "Detection of the Remote Code Execution Attacks Using the PHP Web Application Intrusion Detection System," Electronic and Cyber Defense, vol. ۱۰, no. ۲, pp. ۷۵-۸۵, ۲۰۲۲. dor: ۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۴۰۱.۱۰.۲.۷.۳ (in persian)

[۲۱] V. Yadegari, and A. R. Matinfar, "Detect Web Denial of Service Attacks Using Entropy and Support Vector Machine Algorithm," Electronic and Cyber Defense, vol. ۶, no. ۴, pp. ۷۹-۸۹, ۲۰۱۹, dor: ۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۳۹۷.۶.۴.۷.۹ (in persian)

[۲۲] M. Fathian, A. M. Abdollahi, and H. Dehghani, "Modeling Browsing Behavior Analysis for Malicious Robot Detection in Distributed Denial of Service Attacks," Electronic and Cyber Defense, vol. ۴, no. ۲,

[۳۲] T. Li, G. Kou, and Y. Peng, "Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods," Information Systems, vol. ۹۱, ۱۰۱۴۹۴, ۲۰۲۰, https://doi.org/۱۰.۱۰۱۶/j.is.۲۰۲۰.۱۰۱۴۹۴.

[۳۳] G. Palaniappan, S. Sangeetha, B. Rajendran, S. Goyal, and B. S. Bindhumadhava, "Malicious domain detection using machine learning on domain name features, host-based features and web-based features," Procedia Computer Science, vol. ۱۷۱, pp. ۶۵۴-۶۶۱, ۲۰۲۰, https://doi.org/۱۰.۱۰۱۶/j.procs.۲۰۲۰.۰۴.۰۷۱.

[۳۴] K. A. Messabi, M. Aldwairi, A. A. Yousif, A. Thoban, and F. Belqasmi, "Malware detection using dns records and domain name features," In Proceedings of the ۲nd International Conference on Future Networks and Distributed Systems, pp. ۱-۷, ۲۰۱۸, doi: ۱۰.۱۱۴۵/۳۲۳۱۰۵۳.۳۲۳۱۰۸۲.

[۳۵] W. Bo, Z. B. Fang, L. X. Wei, Z. F. Cheng, Z. X. Hua, "Malicious URLs detection based on a novel optimization algorithm," IEICE TRANSACTIONS on Information and Systems, vol. ۱۰۴(۴), pp. ۵۱۳-۵۱۶, ۲۰۲۱, doi: ۱۰.۱۵۸۷/transinf.۲۰۲۰EDL۸۱۴۷.

[۳۶] J. Yuan, G. Chen, S. Tian, and X. Pei, "Malicious URL detection based on a parallel neural joint model," IEEE Access, vol. ۹, pp. ۹۴۶۴-۹۴۷۲, ۲۰۲۱, doi: ۱۰.۱۱۰۹/access.۲۰۲۱.۳۰۴۹۶۲۵.

[۳۷] S. He, B. Li, H. Peng, J. Xin, and E. Zhang, "An effective cost-sensitive XGBoost method for malicious URLs detection in imbalanced dataset," IEEE Access, vol. ۹, pp. ۹۳۰۸۹-۹۳۰۹۶, ۲۰۲۱, doi: ۱۰.۱۱۰۹/access.۲۰۲۱.۳۰۹۳۰۹۴.

[۳۸] S. Kumi, C. Lim, S. G. Lee, "Malicious url detection based on associative classification," Entropy, vol ۲۳(۲), p. ۱۸۲, ۲۰۲۱, doi: ۱۰.۳۳۹۰/e۲۳۰۲۰۱۸۲.

[۳۹] Z. Chen, Y. Liu, C. Chen, M. Lu, and X. Zhang, "Malicious url detection based on improved multilayer recurrent convolutional neural network model," Security and Communication networks, no. ۱, p.۹۹۹۴۱۲۷, ۲۰۲۱, https://doi.org/۱۰.۱۱۵۵/۲۰۲۱/۹۹۹۴۱۲۷.

[۴۰] R. Patgiri, A. Biswas, S. Nayak, "deepBF: Malicious URL detection using learned bloom filter and evolutionary deep learning. Computer Communications," vol. ۲۰۰, pp. ۳۰-۴۱, ۲۰۲۳, https://doi.org/۱۰.۱۰۱۶/j.comcom.۲۰۲۲.۱۲.۰۲۷.

[۴۱] Broadcom, "URL Risk Levels," https://knowledge.broadcom.com/external/article/۱۷۵۵۸۹/url-risk-levels.html

pp. ۱-۱۳, ۲۰۱۶. dor: ۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۳۹۵.۴.۲.۱.۵ (in persian)

[۲۳] X. Lyu, Y. Ding, S. H. Yang, "Safety and security risk assessment in cyber- physical systems," IET Cyber- Physical Systems: Theory & Applications, vol. ۴, no. ۳, pp. ۲۲۱-۲۳۲, ۲۰۱۹, https://doi.org/۱۰.۱۰۴۹/iet-cps.۲۰۱۸.۵۰۶۸.

[۲۴] C. S. Gates, N. Li, H. Peng, B. Sarma, Y. Qi, R. Potharaju, C. Nita-Rotaru, and I. Molloy, "Generating summary risk scores for mobile applications," IEEE Transactions on dependable and secure computing, vol. ۱۱, no. ۳, pp. ۲۳۸-۲۵۱, ۲۰۱۴, doi: ۱۰.۱۱۰۹/tdsc.۲۰۱۴.۲۳۰۲۲۹۳.

[۲۵] M. Deypir, "Estimating Security Risks of Android Apps Using Information Gain," Electronic and Cyber Defense, vol. ۵, no. ۱, pp. ۷۳-۸۲, ۲۰۱۷. (in persian)

[۲۶] M. Deypir, "Presenting A Method Based on Nearest Neighbors and Hamming Distance in Order to Identify Malicious Applications," Scientific Journal of Electronical & Cyber Defence, vol. ۱۱, no. ۲, ۲۰۲۳, dor: ۲۰.۱۰۰۱.۱.۲۳۲۲۴۳۴۷.۱۴۰۲.۱۱.۲.۷.۰. (in persian)

[۲۷] M. Deypir, and A. Horri, "Instance based security risk value estimation for Android applications," Journal of information security and applications, vol. ۴۰, pp. ۲۰-۳۰, ۲۰۱۸, https://doi.org/۱۰.۱۰۱۶/j.jisa.۲۰۱۸.۰۲.۰۰۲.

[۲۸] M. Deypir, "Entropy-based security risk measurement for Android mobile applications," Soft Computing," vol. ۲۳, no. ۱۶, pp. ۷۳۰۳-۷۳۱۹, ۲۰۱۹, https://doi.org/۱۰.۱۰۰۷/s۰۰۵۰۰-۰۱۸-۳۳۷۷-۵.

[۲۹] A. S. Raja, R. Vinodini, and A. Kavitha, "Lexical features based malicious URL detection using machine learning techniques," Materials Today: Proceedings, vol. ۴۷, pp. ۱۶۳-۱۶۶, ۲۰۲۱, https://doi.org/۱۰.۱۰۱۶/j.matpr.۲۰۲۱.۰۴.۰۴۱.

[۳۰] M. Kuyama, Y. Kakizaki, R. Sasaki, "Method for detecting a malicious domain by using whois and dns features," In Proceedings of the Third International Conference on Digital Security and Forensics (DigitalSec۲۰۱۶), Kuala Lumpur, Malaysia, pp. ۶–۸, ۲۰۱۶.

[۳۱] M. S. I. Mamun, M. A. Rathore, A. H. Lashkari, N. Stakhanova, and A. A. Ghorbani, "Detecting malicious urls using lexical analysis," In International Conference on Network and System Security, pp. ۴۶۷-۴۸۲. Springer, Cham, ۲۰۱۶, doi: ۱۰.۱۰۰۷/۹۷۸-۳-۳۱۹-۴۶۲۹۸-۱_۳۰.

[۴۲] Github, "Google Web Risk," https://github.com/google/webrisk.

[۴۳] M. Deypir, T. Zoughi, "Novel Security Metrics for Identifying Risky Unified Resource Locators (URLs)," Iranian Journal of Science and Technology, Transactions of Electrical Engineering, pp. ۱-۱۹, ۲۰۲۴, doi: ۱۰.۱۰۰۷/s۴۰۹۹۸-۰۲۳-۰۰۶۹۰-x.

[۴۴] A.Tharwat, T. Gaber, A. Ibrahim, A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," AI communications, vol. ۳۰, no. ۲, pp. ۱۶۹-۱۹۰, ۲۰۱۷, doi: ۱۰.۳۲۳۳/aic-۱۷۰۷۲۹.

[۴۵] Kaggle, "Malicious URL Detection using MLP," https://www.kaggle.com/code/ashisharya۰۱/malicious-url-detection-using-mlp-۹۹-۶-accuracy/data?select=urldata.csv

[۴۶] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras, "A high-performance, scalable infrastructure for large-scale active DNS measurements," IEEE journal on selected areas in communications, vol. ۳۴, no. ۶, pp. ۱۸۷۷-۱۸۸۸, ۲۰۱۶, doi: ۱۰.۱۱۰۹/jsac.۲۰۱۶.۲۵۵۸۹۱۸.

[۴۷] L. Qu, Y. Pei, "A Comprehensive Review on Discriminant Analysis for Addressing Challenges of Class-Level Limitations, Small Sample Size, and Robustness," *Processes*, vol. ۱۲, no. ۷, p. ۱۳۸۲, ۲۰۲۴, https://doi.org/۱۰.۳۳۹۰/pr۱۲۰۷۱۳۸۲.