Scientific Journal of Electronic & Cyber Defense

Vol. 12, No. 3, Autumn 2024, Serial No. 47

ISSN: 2322-4347 , E-ISSN: 2980-8979

Malware Classification Based On Visualizing Binary Content of Samples

E. Khazai, M. R. Hassanzadeh*

Assistant professor, Babol Noshirvani University of Technology *

(Received:2024 /06/18, Revised: 2024/09/04, Accepted: 2024/10/14, Published: 2024/10/22)

DOR: DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.3.5.7

Abstract

Malware is one of the constant challenges of the modern world, which has particular importance due to the harm it causes to users. In the last decade, there has been a great increase in malware number and complexity that caused the current security tools and methods not able to defend against. Visualizing binary content of malware and searching for malicious elements among suspicious image patterns is one of the new methods that have achieved high progress and efficiency in the last decade thanks to deep learning algorithms. In this research, by combining various ideas that exist in the field of malware image analysis, a suitable algorithm has been presented for classifying malware into their corresponding families. Visualizing the binary content of the malware executable file, applying the GIST descriptor and classifying the extracted features using the SVM classifier forms the proposed algorithm of this research, which can achieve the same results as previous researches by using traditional machine learning methods and obtain average classification accuracy of 99.72 and 99.16% on Malimg and Microsoft datasets.

Keywords: Malware Family, Static Analysis, Visualization, Image Classification

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Publisher: Imam Hussein University © Authors

**Corresponding Author Email: m.hasanzadeh@nit.ac.ir*

«پدافند الکترونیکی و سایبری»
سال دوازدهم، شماره ۳، فصل پاییز ۱۴۰۳؛ ص ۴۷-۳۵

علمی – پژوهشی

دسته‌بندی بدافزارها بر اساس بصری‌سازی محتوای دودویی نمونه‌ها

اسماعیل خزائی [۱]، محمدرضا حسن‌زاده [۲*]

۱- کارشناسی ارشد، ۲- استادیار، دانشگاه صنعتی نوش یروانی بابل، بابل، ایران

(دریافت: ۱۴۰۳/۰۳/۲۹، بازنگری: ۱۴۰۳/۰۶/۱۴، پذیرش: ۱۴۰۳/۰۷/۲۳، انتشار: ۱۴۰۳/۰۸/۰۱)

DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.3.5.7

* این مقاله یک مقاله با دسترسی آزاد است که تحت شرایط و ضوابط مجوز Creative Commons Attribution (CC BY) توزیع شده است.
ناشر: دانشگاه جامع امام حسین (ع) نویسندگان

چکیده

بدافزارها از چالش‌های همیشگی دنیای مدرن محسوب می‌شوند که به علت آزار کاربران و خساراتی که به وجود می‌آورند، از اهمیت ویژه‌ای برخوردارند. در دهه اخیر، رشد سریع و پیچیده‌تر شدن سازوکار مخرب بدافزارها سبب شده است، ابزارها و روش‌های امنیتی فعلی نتوانند با این اندازه از وسعت و تنوع تهدیدها مقابله کنند. بصری ساختن محتوای دودویی بدافزار و جستجوی عناصر مخرب از بین الگوهای تصویری مشکوک، از روش‌های نوین محسوب می‌شود که در دهه گذشته، به لطف الگوریتم‌های پردازش تصویر به پیشرفت و کارایی بالایی دست پیدا کرده است. در این پژوهش با ترکیب ایده‌های متنوعی که در زمینه تحلیل تصویری بدافزارها وجود دارد، یک روش مناسب برای دسته‌بندی بدافزارها به خانواده‌های متناظرشان ارائه‌شده است. بصری‌سازی محتوای دودویی فایل اجرایی بدافزار، اعمال توصیفگر GIST[1] و دسته‌بندی ویژگی‌های استخراج‌شده با استفاده از طبقه‌بند SVM، روش پیشنهادی این پژوهش را تشکیل می‌دهد که می‌تواند تنها با به‌کارگیری روش‌های سنتی یادگیری ماشین، به نتایجی برابر با پژوهش‌های پیشین دست پیدا کند و در مجموعه‌داده‌های Malimg و مایکروسافت، به میانگین دقت طبقه‌بندی ۹۹/۷۲ و ۹۹/۱۶ درصد برسد.

کلیدواژه‌ها: خانواده بدافزار، تحلیل ایستا، بصری‌سازی، دسته‌بندی تصاویر

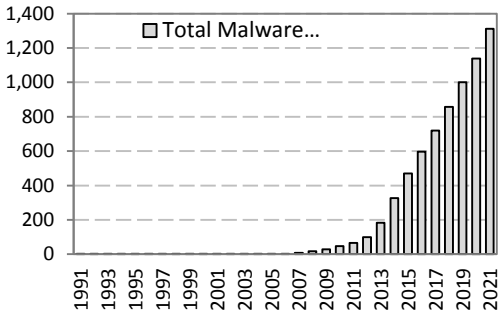
۱- مقدمه

به‌طورکلی هرگونه نرم‌افزاری که عامدانه برای آسیب‌زدن به یک رایانه یا شبکه رایانه‌ای طراحی‌شده باشد، بدافزار[2] گفته می‌شود [۱]. بدافزارها ازجمله چالش‌های جدی عصر حاضر هستند که انگیزه توسعه آن‌ها از یک مزاحمت ساده به دزدی، جاسوسی و خرابکاری گسترش‌یافته است و هم‌اکنون طیف وسیعی از افراد، کسب‌وکارها و حتی دولت‌ها را تهدید می‌کند. در گذشته به کلیه نرم‌افزارهای مخرب ویروس گفته می‌شد. ویروس‌های رایانه‌ای عمده تهدیداتی بوده‌اند که به شیوه ویروس‌های عادی در دنیای واقعی، با آلوده کردن حافظه‌های قابل‌حمل شیوع می‌یافتند. امروزه با پیشرفت فناوری‌های ارتباطی، سازوکار بدافزارها نیز متنوع‌تر و پیچیده‌تر شده است، به‌طوری‌که رفتار ویروس گونه بدافزار، تنها یک راهبرد از میان طیف وسیع تهدیدات محسوب می‌شود [۲].

طبق آمار ارائه‌شده از مؤسسه تحقیقاتی AV-Test، تعداد بدافزارهای کشف‌شده از ابتدای سال ۱۹۸۴ میلادی تاکنون در حدود ۱/۴ میلیارد برآورد شده است که نیمی از آن‌ها در ۵ سال اخیر تولیدشده‌اند. شکل (۱) نمودار تجمعی بدافزارهای کشف‌شده تا به امروز را نشان می‌دهد که رشدی نمایی را دنبال می‌کند [۳]. البته آمار واقعی بدافزارها از این میزان بیشتر بوده و نمونه‌های بسیاری وجود دارند تاکنون کشف نشده‌اند. در این شرایط روش‌های دستی تحلیل بدافزار، ازلحاظ تعداد نیروی متخصص و زمان موردنیاز برای یافتن راه‌حل، برای مقابله با رشد روزافزون تهدیدها کافی نیستند. ازاین‌رو امروزه اکثر فرایند تحلیل بدافزار به‌صورت خودکار پیگیری می‌شود و فقط درصورتی‌که تهدید موردنظر بسیار مهم باشد یا ابزارهای تحلیل خودکار توانایی مقابله با آن را نداشته باشند، متخصصان وارد عمل می‌شوند. انگیزه این پژوهش نیز ارائه الگوریتمی است که بتواند در خودکارسازی فرایند تحلیل بدافزار به متخصصان کمک کند.

* رایانامه نویسنده مسئول: m.hasanzadeh@nit.ac.ir

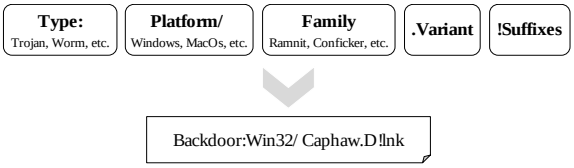
[1] Global Image Structure Tensor
[2] Malware



شکل (۱). نمودار تجمعی بدافزارهای کشف‌شده در سه دهه اخیر [۳]

۱-۱- دسته‌بندی بدافزارها

برای مقابله با طیف وسیع بدافزارها، متخصصان آن‌ها را به روش‌های متعدد دسته‌بندی می‌کنند. این امر به متخصص کمک می‌کند که هر نمونه جدید بر اساس انواع مشابه قبلی، قابل‌بررسی باشد. نوع رفتار، سکوی اجرا و خانواده بدافزار، معیارهای رایجی هستند که به دسته‌بندی بدافزارها کمک می‌کنند و همانند شکل (۲) چهارچوب دقیقی برای اشاره به هر بدافزار می‌سازند. کاربردی‌ترین معیار دسته‌بندی بدافزارها که در این پژوهش به آن پرداخته‌شده است، تفکیک نمونه‌ها به خانواده‌های مختلف است. هر خانواده به گروهی از بدافزارها اطلاق می‌شود که وجوه مشترک بسیاری در سازوکار مخرب خود داشته و اهداف یکسانی را دنبال می‌کنند. مستقل از اینکه هر بدافزار چه طیفی از رفتارها و سکوهای اجرایی داشته باشد، این معیار منحصربه‌فرد است و تنها به جامعه کوچکی از نمونه‌ها اشاره دارد. بدین ترتیب متخصصان می‌توانند تمرکز خود را بر روی نمونه‌های کاملاً مرتبط معطوف کرده و برداشت دقیق‌تری از سازوکارهای مخرب داشته باشند.



شکل (۲). چهارچوب نام‌گذاری بدافزارها، مورداستفاده در محصولات امنیتی شرکت مایکروسافت [۴]

۱-۲- تحلیل بدافزار

برای دسته‌بندی و مقایسه بدافزارها، نمونه‌های مربوطه مورد تحلیل قرار می‌گیرند. تحلیل بدافزار فرایندی است که در آن به روش‌ها و ابزارهای متعدد، حداکثر اطلاعات ممکن از یک نمونه مشکوک استخراج شود. اگر این فرایند بدون اجرای بدافزار و تنها بر اساس ویژگی‌های فایل آن انجام شود، آن را به‌اصطلاح «تحلیل ایستا» می‌نامند. مشخصات ظاهری فایل بدافزار اعم از نام و حجم نمونه موردنظر، مشخصات محتوایی و رشته‌های دودویی موجود در آن ازجمله اطلاعاتی هستند که در این تحلیل بکار گرفته می‌شوند [۵]. در کنار تحلیل ایستا، راهبردهای «تحلیل پویا» و «تحلیل پیشرفته» نیز وجود دارند که به ترتیب با نظارت بر اجرای بدافزار یا مهندسی معکوس الگوریتم آن به اطلاعات مهم دست پیدا می‌کنند. اما این تحلیل‌ها ازلحاظ زمانی نسبت به تحلیل ایستا طولانی‌تر بوده و نیاز به اجرای بدافزار در یک محیط قرنطینه یا تجربه کافی برای بررسی کد زبان ماشین دارند.

۱-۳- تحلیل اکتشافی[1]

تحلیل اکتشافی فرایند استخراج ویژگی‌های ایستا و پویا به‌منظور شناسایی و دسته‌بندی بدافزارها است. برخلاف تحلیل‌های قبلی، این ویژگی‌ها به مؤلفه‌های مخرب خاصی اشاره ندارند و وجود آن‌ها در یک نمونه مشکوک به معنای مخرب بودن قطعی آن نیست. ویژگی‌های استخراج‌شده تنها داده‌هایی خام هستند که متخصصان آن‌ها را از طیف وسیعی از نرم‌افزارهای مخرب و غیرمخرب استخراج کرده و طبقه‌بندها را بر اساس آن آموزش می‌دهند. هدف طبقه‌بند یادگیری الگوهایی است که به کشف بدافزارها و تشخیص نوع آن‌ها کمک کند. اگر این الگوها در نمونه غریبه‌ای یافت شود، آن نمونه به‌احتمال زیاد مخرب است و به خانواده‌ای تعلق دارد که بیشترین اشتراک را با آن داشته باشد. امروزه که با رشد سریع و پیچیده‌تر شدن بدافزارها، روش‌های پایه‌ای تحلیل به چالش کشیده شده‌اند، تحلیل اکتشافی می‌تواند با ارائه یک الگوریتم خودکار و یادگیرنده به این چالش‌ها پایان دهد.

۱-۴- تحلیل تصویری بدافزار

تحلیل تصویری بدافزار یک روش ایستای اکتشافی است که در آن نمونه‌ها به کمک تصاویرشان، شناسایی و دسته‌بندی می‌شوند. تصاویر بدافزارها از بصری‌سازی مستقیم محتوای فایل موردنظر به دست می‌آید به‌طوری‌که هر پیکسل تصویر، نماینده بخشی از داده‌های دودویی موجود در فایل است. از این طریق الگوریتم‌های اکتشافی می‌توانند الگوهای مخرب دودویی را از روی تصویر بیاموزند و متخصصان را در ریشه‌کن‌کردن بدافزارها یاری دهند. البته بصری‌سازی محتوای دودویی منحصربه‌فرد نیست و متخصصان می‌توانند بسته به اهداف مدنظرشان، روش‌های متعددی را بکار گیرند. اینکه تصاویر تولیدشده برای ناظر انسانی قابل‌درک یا برای استخراج ویژگی غنی باشد، می‌تواند ازجمله اهداف این تحلیل باشد.

۱-۵- پژوهش فعلی

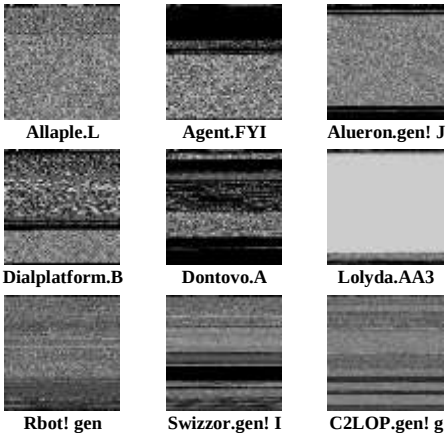
نوآوری‌های این پژوهش را می‌توان در ارتقا روش‌های تحلیل تصویری موجود و ترکیب آن‌ها با روش‌های جدیدتر و کارآمدتر دانست. استفاده از بردار بایتی برای ساخت بردارهای هیستوگرام و آنتروپی، استفاده چینش هیلبرت، ماتریس نقطه‌ای و ماتریس

[1] Heuristic

هیستوگرام برای نمایش فایل بدافزارها، بررسی عملکرد یکایک بصری‌سازی‌ها، استخراج محتوای دودویی بدافزارها از مجموعه Malimg، ترکیب مجموعه‌ها برای ساخت یک فضای نمونه بزرگ‌تر (۲۰ هزار بدافزار) و بالاتر این موارد، ساخت الگوریتمی که بتواند با استفاده از روش‌های رایج یادگیری ماشین با پژوهش‌های مبتنی بر یادگیری عمیق (CNN[1]) –که ازلحاظ محاسبات به‌مراتب سنگین‌تر هستند - برابری کند، به نوبه خود قدمی روبه‌جلو محسوب می‌شود.

۲- پیشینه پژوهش

تحلیل تصویری بدافزار مطابق آنچه امروز می‌شناسیم از پژوهش [۶] آغاز شد. در این پژوهش محتوای دودویی فایل بدافزار که به‌مثابه برداری از بایت‌های متوالی است، به برداری از پیکسل‌های خاکستری نگاشت داده می‌شود، به‌طوری‌که روشنایی هر پیکسل بر اساس عدد بایت متناظرش تعیین می‌شود. به عبارت دقیق‌تر، بایت صفر متناظر با پیکسل مشکی، بایت ۲۵۵ متناظر با پیکسل سفید و مقادیر بایتی بین این دو، ۲۵۵ درجه روشنایی را تعیین می‌کند. سپس با چینش سطر به سطر پیکسل‌ها، تصاویر دوبعدی از فایل موردنظر تهیه می‌گردد که می‌توان آن را به روش‌های متعدد موردبررسی قرارداد. این پژوهش نشان داد که بدافزارهای هم‌خانواده به دلیل استفاده از داده‌ها و کدهای مخرب مشابه، الگوهای تصویری نزدیکی دارند، همچنین تفاوت بین الگوهای تصویری خانواده‌های مختلف به حدی است که همانند شکل (۳) حتی با چشم غیر مجهز از یکدیگر قابل‌تشخیص‌اند. برای‌آنکه این ادعا به‌صورت علمی نیز تائید شود، با استخراج ویژگی GIST از تصاویر و به‌کارگیری طبقه‌بند k-NN، نشان داده شد که می‌توان خانواده‌های مختلف بدافزارها را با دقت ۹۸ درصد دسته‌بندی کرد. بدین ترتیب روش‌های مبتنی بر پردازش تصویر به‌عنوان راهکاری نوین درزمینه تحلیل ایستای بدافزارها مطرح شدند.



شکل (۳). بصری‌سازی خانواده‌های مختلف بدافزارها [۶]

در دهه گذشته، با موفقیت شبکه‌های عصبی عمیق در پردازش تصویر، استخراج ویژگی به عهده ماشین گذاشته شد تا خود بتواند طی یادگیری‌های متعدد بهترین راه‌حل را بیابد و محدود به یک الگوریتم خاص برای حل مسئله نباشد. همسو با این روند، در پژوهش‌های تحلیل تصویری بدافزار نیز توصیفگر GIST با روش‌های مبتنی بر شبکه‌های عصبی عمیق جایگزین شد، تا ازیک‌طرف دیگر نیازی به ویژگی‌های دست‌ساخته و غیر بهینه نباشد و از طرف دیگر سازندگان بدافزار نتوانند با دور زدن محاسبات ریاضی این ویژگی‌ها، بدافزارهای خود را بی‌خطر جلوه دهند [۷–۱۱]. پژوهش [۷] ازجمله موارد مهم در این زمینه محسوب می‌شود که در آن با همکاری مشترک شرکت‌های فناوری مایکروسافت و اینتل در سال ۲۰۲۰ میلادی، مجموعه‌ای متشکل از ۲ میلیون بدافزار به‌صورت بصری و با استفاده از شبکه عصبی Resnet موردبررسی قرارگرفته‌اند.

هم‌زمان با پژوهش‌های ذکرشده، برخی دیگر از محققان بر روی نحوه ساخت تصاویر متمرکزشده و سعی در غنی‌تر کردن ویژگی‌های بصری آن داشته‌اند. ازجمله تلاش‌های انجام‌شده در این زمینه می‌توان به آنتروپی اشاره کرد که از آن برای اندازه‌گیری میزان فشردگی بخش‌های مختلف فایل مشکوک استفاده‌شده است و می‌تواند برای تشخیص کُد و نوع داده‌ها و میزان فشردگی آن‌ها مورداستفاده قرار گیرد [۱۲, ۱۳]. هم‌اکنون این معیار در پژوهش‌های [۱۴, ۱۵] نه‌تنها به‌عنوان یک ویژگی تک-بُعدی بلکه به‌عنوان یک روش بصری‌سازی یا کانال رنگی برای نمایش بدافزارها استفاده‌شده است. منحنی فضا گیر هیلبرت[2] از دیگر ابزارهای مورداستفاده در خصوص بهبود بصری‌سازی است که به دلیل خاصیت آن در حفظ خصوصیات محلی برای نگاشت داده‌های یک-بُعدی به ابعاد بالاتر بکار گرفته می‌شود. در پژوهش [۱۶] نشان داده شد که چینش هیلبرت می‌تواند نسبت به چینش سطری به نتایج بهتری برای تشخیص/دسته‌بندی بدافزارها بیانجامد.

از آخرین پژوهش‌های انجام‌شده در این زمینه می‌توان به پژوهش [۱۷] اشاره کرد که در آن تصویر بدافزار و آنتروپی آن به‌عنوان ورودی‌های یک شبکه عصبی پیچشی برای دسته‌بندی بدافزارها را به کار گرفته‌شده‌اند. یا در پژوهش [۱۸] محتوای دودویی فایل بدافزار، به تصاویر رنگی تبدیل‌شده و با استفاده روش‌های مختلف یادگیری عمیق موردبررسی قرارگرفته‌اند.

۳- روش پژوهش

الگوریتم پیشنهادی این پژوهش متشکل از چهار فرایند اصلی است. ابتدا در مرحله پیش‌پردازش، سه بردار از فایل نمونه به دست می‌آید. سپس هریک از این بردارها به روش‌های متعدد

[1] Convolutional Neural Network

[2] Hilbert Space Filling Curve

بصری‌سازی شده و ویژگی‌های تصویری مطلوب از آن استخراج می‌شود. درنهایت با استفاده از طبقه‌بند SVM، هر نمونه به خانواده متناظرش دسته‌بندی‌شده و الگوریتم ارزیابی می‌شود.

شکل (۴). روند الگوریتم پیشنهادی

۳-۱- پیش‌پردازش فایل‌ها

در مرحله پیش‌پردازش از هر فایل سه بردار استخراج می‌شود: بردار بایتی، بردار آنتروپی و بردار هیستوگرام. بردار بایتی مستقیماً از محتوای دودویی فایل موردنظر به دست می‌آید. دو بردار دیگر نیز از روی بایتی محاسبه می‌شوند.

برای ساخت بردارهای آنتروپی و هیستوگرام، پنجره‌ای به‌اندازه ۲۵۶ بایت و با گام‌های ۲۵۶ بایت در طول بردار بایتی حرکت کرده و مقدار هر بایت را با آنتروپی آن پنجره یا تعداد دفعات تکرار بایت‌ها در آن پنجره جایگزین می‌کند. بدین ترتیب بردارهایی هم‌اندازه با بردار بایتی به دست خواهد آمد که می‌توانند نرخ تغییرات آنتروپی و تکرار بایت‌ها به‌موازات یکدیگر نشان دهند. نحوه ساخت این دو بردار در شکل‌های (۵) و (۶) نمایش داده است. در این اشکال b_n به معنای بایت n-ام و B_n بلوکی است که از بایت n-ام آغاز می‌شود. همچنین $E(B_n)$ آنتروپی بلوک B_n و $H(b_n)$ تعداد دفعات تکرار بایت n-ام در پنجره موردنظر است.

آنتروپی متغیر تصادفی X از رابطه (۱) به دست می‌آید که در آن p_i احتمال پیشامد i-ام از فضای پیشامد X است. ازآنجاکه هر بایت ۲۵۶ مقدار متفاوت (۰ الی ۲۵۵) می‌تواند داشته باشد، آنتروپی بایت‌ها مقداری حقیقی بین ۰ الی ۸ خواهد داشت.

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (۱)$$

اگر n تعداد بایت‌های بردار بایتی باشد، پیچیدگی زمانی محاسبه بردار هیستوگرام O(n) خواهد بود. همچنین ازآنجاکه می‌توان از این اطلاعات این بردار می‌توان برای محاسبه بردار آنتروپی استفاده کرد، در پیاده‌سازی الگوریتم این پژوهش پیچیدگی زمانی محاسبه بردار آنتروپی نیز برابر O(n) است.

...	$b_n \dots b_{n+255}$	$b_{n+256} \dots b_{n+511}$	...
	↓		
...	$\boldsymbol{E(B_n) \dots E(B_n)}$	$b_{n+256} \dots b_{n+511}$	...
		↓	
...	$E(B_n) \dots E(B_n)$	$\boldsymbol{E(B_{n+256}) \dots E(B_{n+256})}$	...

شکل (۵). نحوه ساخت بردار آنتروپی از روی بردار بایتی

...	$b_n \dots b_{n+255}$	$b_{n+256} \dots b_{n+511}$	...
	↓		
...	$\boldsymbol{H(b_n) \dots H(b_{n+255})}$	$b_{n+256} \dots b_{n+511}$	...
		↓	
...	$H(b_n) \dots H(b_{n+255})$	$\boldsymbol{H(b_{n+256}) \dots H(b_{n+511})}$	...

شکل (۶). نحوه ساخت بردار هیستوگرام از روی بردار بایتی

۳-۲- بصری‌سازی فایل‌ها

پس ازآنکه بردارهای سه‌گانه از فایل موردنظر تهیه شد، درایه‌های بردارها تغییر مقیاس داده شده و به ماتریس‌های دوبعدی نگاشت داده می‌شوند تا تصاویر مطلوب را بسازند. به‌طورکلی ۵ نوع بصری‌سازی مختلف برای هر بردار در نظر گرفته‌شده است.

۳-۲-۱- نمایش مستقیم

در نمایش مستقیم یک بردار، هر درایه پیکسل متناظری در تصویر دارد. به‌طوری‌که می‌توان هر بافت یا الگوی تصویر را در بخش متناظری از بردار جستجو کرد. مراحل نمایش مستقیم هر بردار دلخواه به شرح زیر است:

۱. درایه‌های بردار به بازه ۰ الی ۱ تغییر مقیاس داده می‌شوند.
۲. بردار موردنظر به روش درون‌یابی خطی بسط داده می‌شود تا اندازه بردار به توانی از چهار برسد. پیچیدگی زمانی این الگوریتم در بدترین حالت O(n) است که n تعداد بایت‌های بردار بایتی است.
۳. درایه‌های بردار بر مبنای چینش هیلبرت در یک ماتریس مربعی جانمایی می‌شوند. (بسط‌دادن بردار در مرحله پیشین کمک می‌کند تا در این مقطع ماتریس مربعی کاملی ساخته شود)
۴. ماتریس به‌دست‌آمده می‌تواند به‌عنوان یک تصویر خاکستری مورداستفاده قرار گیرد که در آن موقعیت و روشنایی هر پیکسل بر اساس جانمایی و عدد درایه متناظرش تعیین می‌شود.

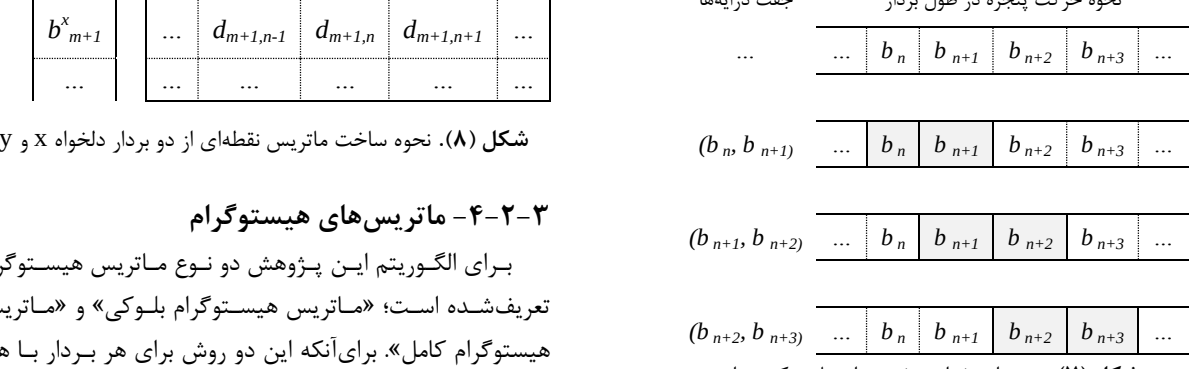
۳-۲-۲- نمایش غیرمستقیم

در نمایش غیرمستقیم، ماتریس‌های ویژگی با تغییر مقیاس به بازه ۰ الی ۱ به‌صورت یک تصویر خاکستری مورداستفاده قرار می‌گیرند. وجه مشترک این ماتریس‌ها، استفاده از مشخصات آماری مرتبه دو و اطلاعات همسایگی درایه‌ها برای ساخت آن‌ها است. البته در بعضی موارد تغییر مقیاس می‌تواند جزئیات مربوط به درایه‌های کوچک‌تر را از بین برده و تصویر را کاملاً تیره کند. این موضوع معمولاً زمانی رخ می‌دهد که تعدادی درایه بسیار بزرگ که جزو نمونه‌های پرت به شمار می‌آیند، در ماتریس وجود داشته باشند. برای عبور از این مشکل، پیش از تغییر مقیاس، مقدار بیشینه خاصی برای درایه‌های ماتریس در نظر گرفته می‌شود و درایه‌های بالاتر از این محدوده با مقدار بیشینه

جایگزین می‌شوند. در این پژوهش، حد بیشینه دو برابر میانگین درایه‌ها در نظر گرفته‌شده است، تا تضاد بین روشنایی و تاریکی تصاویر متعادل باشد.

۳-۲-۲-۱- ماتریس هم رخداد[1]

ماتریس هم رخداد ازجمله روش‌های تحلیل بردارها و تصاویر محسوب می‌شود که در آن تعداد دفعات رخدادن زوج درایه‌ها به‌صورت یک ماتریس هیستوگرام دوبُعدی به نمایش درمی‌آید. در این پژوهش برای‌آنکه از هر بردار دلخواه، ماتریس هم رخداد یکسانی تهیه شود، ابتدا درایه‌های بردار به بازه ۱ الی ۲۵۶ تغییر مقیاس داده‌شده و به پایین گرد می‌شوند تا مقدار صحیحی داشته باشند. سپس پنجره‌ای به‌اندازه دو درایه و با گام یک درایه، در طول بردار به حرکت درآمده و همانند شکل (۷) درایه‌ها را به‌صورت دوتایی از آن استخراج می‌کند. مقادیر هر جفت درایه به مختصات سطر و ستون متناظری از ماتریس هم رخداد اشاره دارد. ازآنجاکه بازه مقادیر درایه‌ها، اعداد صحیح بین ۱ الی ۲۵۶ هستند، ماتریس هم رخداد مطلوب نیز ابعاد ۲۵۶×۲۵۶ خواهد داشت. درنهایت تعداد دفعات تکرار هر جفت درایه منحصربه‌فرد در ماتریس هم رخداد ذکر می‌شود. پیچیدگی زمانی این الگوریتم $O(n)$ است که n تعداد بایت‌های بردار بایتی است.

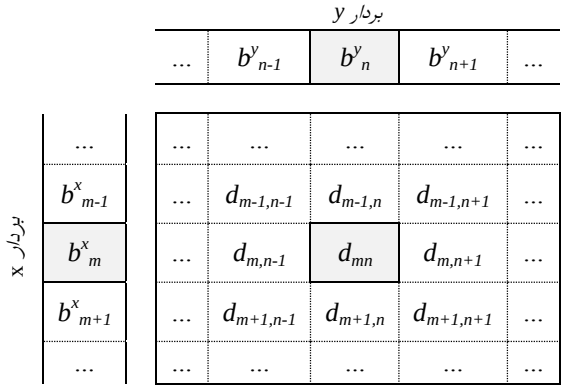


شکل (۷). نحوه استخراج جفت درایه‌های یک بردار

۳-۲-۳- ماتریس نقطه‌ای

ماتریس نقطه‌ای ازجمله روش‌های مقایسه دو بردار محسوب می‌شود که در آن شباهت‌ها و تفاوت‌های یکایک درایه‌ها، به‌صورت ماتریسی از صفرها و یک‌ها نمایش داده می‌شود. به‌عنوان‌مثال برای دو بردار x و y اگر درایه x_m از بردار اول با درایه y_n از بردار دوم برابر باشد، درایه (m,n) از ماتریس نقطه‌ای برابر یک و در غیر این صورت برابر صفر خواهد بود. البته ازآنجاکه مرتبه حافظه برای این الگوریتم $O(n_x.n_y)$ است، این موضوع سبب می‌شود برای دو بردار یک مگابایتی، ماتریسی در حدود یک ترابایت تولید شود. ازاین‌رو برای کاهش حافظه مورد نیاز، هر بردار به بلوک‌های هم‌اندازه تفکیک‌شده و بلوک به بلوک مقایسه می‌شود. به‌عنوان‌مثال برای دو بردار x و y، اگر بلوک b^x_m از بردار اول با بلوک b^y_n از بردار دوم، d درایه مشترک داشته باشند، درایه (m,n) از ماتریس نقطه‌ای برابر d خواهد بود که این فرایند در شکل (۸) نشان‌داده‌شده است. پیچیدگی زمانی و مرتبه حافظه این الگوریتم $O(n_x.n_y)$ است؛ اما با تغییر نحوه محاسبه آن می‌توان مرتبه حافظه موردنیاز را به $O(b_x.b_y)$ رساند.

در این پژوهش از ماتریس نقطه‌ای برای مقایسه هر بردار با خودش استفاده شود. برای‌آنکه از هر بردار، ماتریس نقطه هم‌اندازه‌ای به دست آید، ابتدا درایه‌ها به بازه ۰ الی ۲۵۵ تغییر مقیاس داده‌شده و به پایین گرد می‌شوند تا مقادیر صحیحی داشته باشند. سپس بردار به ۲۵۶ بلوک هم‌اندازه تفکیک‌شده و با خودش مقایسه می‌شود. در این حالت ماتریس نقطه‌ای به‌دست‌آمده همواره ابعاد ۲۵۶×۲۵۶ خواهد داشت.



شکل (۸). نحوه ساخت ماتریس نقطه‌ای از دو بردار دلخواه x و y

۳-۲-۴- ماتریس‌های هیستوگرام

برای الگوریتم این پژوهش دو نوع ماتریس هیستوگرام تعریف‌شده است؛ «ماتریس هیستوگرام بلوکی» و «ماتریس هیستوگرام کامل». برای‌آنکه این دو روش برای هر بردار با هر اندازه‌ای قابل‌استفاده باشد، ابتدا بردارهای ورودی به بازه ۰ الی ۲۵۵ تغییر مقیاس داده‌شده و به پایین گرد می‌شوند تا درایه‌ها مقادیر صحیحی داشته باشند. از این طریق بردارهای هیستوگرام محاسبه‌شده همواره ۲۵۶ درایه‌ای خواهند بود.

برای ساخت ماتریس هیستوگرام بلوکی، بردار پس از تغییر مقیاس، به ۲۵۶ بلوک هم‌اندازه تفکیک‌شده و هیستوگرام هریک محاسبه می‌شود. با چیدن سطر به سطر بردارهای به‌دست‌آمده، یک ماتریس ۲۵۶×۲۵۶ به دست خواهد آمد که ساختار آن در شکل (۹) نشان داده‌شده است. در این شکل B_n بیانگر بلوک n-ام و $F_n(m)$ بیانگر تعداد تکرار عدد m در بلوک n-ام است.

فرایند ذکرشده برای ساخت ماتریس هیستوگرام کامل یکسان است، با این تفاوت که در آن تنها هیستوگرام کل بردار محاسبه می‌شود. بردار هیستوگرام به‌دست‌آمده را می‌توان آن در

[1] Co-Occurrence Matrix

ماتریسی مربعی به ابعاد ۱۶×۱۶ جایگذاری کرد که نحوه چینش آن در شکل (۱۰) نشان داده‌شده است. در این شکل F(m) بیانگر تعداد تکرار عدد m در بردار موردنظر است.

ماتریس هیستوگرام بلوکی

بردار ورودی						
	B_1	$F_1(0)$	$F_1(1)$	$F_1(2)$	...	$F_1(255)$
	B_2	$F_2(0)$	$F_2(1)$	$F_2(2)$	...	$F_2(255)$
	B_3	$F_3(0)$	$F_3(1)$	$F_3(2)$	...	$F_3(255)$
	...	...	...	...	...	...
	B_{256}	$F_{256}(0)$	$F_{256}(1)$	$F_{256}(2)$	...	$F_{256}(255)$

شکل (۹). چینش درایه‌ها در ماتریس هیستوگرام بلوکی یک بردار

ماتریس هیستوگرام کامل یک بردار دلخواه

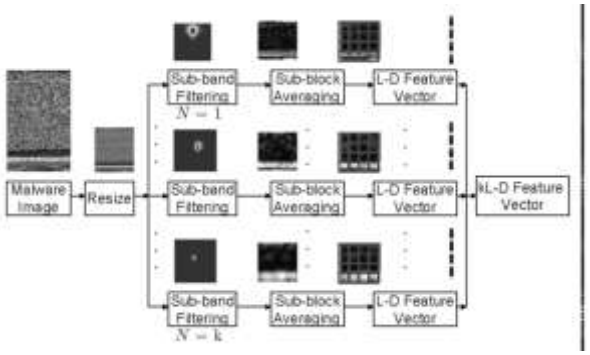
$F(0)$	$F(1)$	$F(2)$	...	$F(15)$
$F(16)$	$F(17)$	$F(18)$	...	$F(31)$
$F(32)$	$F(33)$	$F(34)$	...	$F(47)$
...	...	...	...	...
$F(240)$	$F(241)$	$F(242)$	...	$F(255)$

شکل (۱۰). چینش درایه‌ها در ماتریس هیستوگرام

۳-۳- استخراج ویژگی

پس ازآنکه فایل بدافزار به روش‌های مختلف بصری‌سازی شد، برای تحلیل تصاویر به‌دست‌آمده از توصیفگر GIST استفاده می‌شود. این توصیفگر نخستین‌بار در پژوهش [۱۹] معرفی‌شده است که می‌تواند بدون نیاز به تشخیص اشیاء، ویژگی‌های کافی برای تشخیص محیط را فراهم کند. در این الگوریتم بانکی از فیلترهای گابور در مقیاس‌ها و چرخش‌های مختلف به تصویر موردنظر اعمال می‌شوند. تصاویر به‌دست‌آمده هم‌اندازه با تصویر ورودی خواهند بود که در آن‌ها مطابق با فیلتر اعمال‌شده، برخی الگوهای تصویر برجسته شده است. در ادامه هریک از این تصاویر به بلوک‌های هم‌اندازه تفکیک‌شده و میانگین هر بلوک محاسبه می‌شود. مجموعه میانگین‌های به‌دست‌آمده، برداری از ویژگی‌ها را تشکیل می‌دهد که به آن توصیف GIST از تصویر موردنظر گفته می‌شود. البته باید به این نکته اشاره کرد که تعداد جهت‌ها و مقیاس‌ها در هر کاربرد متغیر است و انتخاب دقیق آن می‌تواند تأثیر زیادی در نتایج داشته باشد. بااین‌حال روش نظام‌مندی برای انتخاب تعداد جهت‌ها و مقیاس‌ها وجود ندارد و بهترین تنظیمات با انجام آزمون‌های متعدد به دست می‌آید.

در این پژوهش از توصیفگر GIST برای استخراج ویژگی تصاویر استفاده می‌شود. فیلترها در یک‌جهت و در ۸ مقیاس مختلف به تصویر اعمال می‌شوند. سپس تصاویر فیلتر شده به ۱۶ بلوک هم‌اندازه تفکیک‌شده و میانگین هریک محاسبه می‌شود. بدین ترتیب برداری حاوی ۱۲۸ ویژگی از هر تصویر به دست می‌آید. شکل (۱۱) روند کلی استخراج بردار ویژگی از تصویر یک بدافزار را نشان می‌دهد که مشابه آن در پژوهش [۱۸] استفاده‌شده است.



شکل (۱۱). روند استخراج ویژگی‌های GIST از تصویر بدافزار [۱۸]

۴- ارزیابی پژوهش

الگوریتم پیشنهادی این پایان‌نامه بر اساس سه معیار دقت[1]، پوشش[2] و صحت[3] مورد ارزیابی قرار می‌گیرد. منظور از معیار دقت، نسبت تعداد پاسخ‌های صحیح به کل مجموعه است که عملکرد الگوریتم را در دسته‌بندی انواع بدافزارها مشخص می‌کند. معیار پوشش، دقت الگوریتم را در یادآوری نمونه‌های مخصوص به هر دسته می‌سنجد. معیار صحت نیز میزان درستی پیش‌بینی الگوریتم در معرفی اعضای هر دسته را نشان می‌دهد. تعاریف ریاضی این معیارها به ترتیب در روابط (۲) الی (۴) آورده شده است. برای ارائه دقیق‌تر نتایج، معیارهای موردنظر به همراه جدول سردرگمی[4] آورده می‌شوند تا قوت و ضعف الگوریتم در تفکیک انواع مختلف خانواده‌ها مشخص شود.

البته فارغ از اینکه نقاط ضعف و قوت الگوریتم مربوط کدام نوع بدافزارها باشد، دقت دسته‌بندی به‌عنوان معیار اصلی این پژوهش همواره اولویت دارد. دلیل اهمیت معیار دقت آن است که الگوریتم این پژوهش و پژوهش‌های مرتبط، برای خودکارسازی فرایند تحلیل بدافزار به‌خصوص دسته‌بندی آن‌ها طراحی‌شده است. دسته‌بندی اشتباه یک بدافزار به نوبه خود عواقبی ندارد و تنها زمان تحلیل را برای فرد متخصص طولانی می‌کند. اما اگر دقت دسته‌بندی الگوریتم برای تعداد بسیاری از نمونه‌ها کافی نباشد، به همان اندازه بهره‌وری تحلیل‌گران نیز کاهش پیدا خواهد کرد. البته می‌توان استدلال کرد که تشخیص بعضی

[1] Accuracy
[2] Recall
[3] Precision
[4] Confusion Matrix

خانواده‌های بدافزار به دلیل خطرناک‌تر بودن اولویت دارد، اما ازآنجاکه در مجموعه‌داده‌های فعلی به میزان اهمیت و خطرناک بودن آن‌ها اشاره نشده است، به همه آن‌ها با یک اولویت نگاه می‌شود.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۲)$$

$$Recall = \frac{TP}{TP + FN} \quad (۳)$$

$$Precision = \frac{TP}{TP + FP} \quad (۴)$$

۴-۱- مجموعه‌داده

در این پژوهش از دو مجموعه‌داده Malimg و مایکروسافت برای ارزیابی الگوریتم پیشنهادی استفاده‌شده است.

مجموعه‌داده Malimg دربردارنده ۹۳۳۹ بدافزار از ۲۵ خانواده مختلف است که در پژوهش [۶] ارائه‌شده است و یکی از پراستفاده‌ترین مجموعه‌ها در حوزه تحلیل تصویری بدافزار به شمار می‌آید. نمونه‌های مخرب با استفاده از روش بصری‌سازی مربوط به آن پژوهش به قالب تصویری png درآمده‌اند. البته ازآنجاکه هیچ دخل و تصرفی در محتوای آن‌ها داده نشده است، محتوای دودویی بدافزارها از تصویرشان قابل‌استخراج است. نکته دیگر در خصوص این مجموعه آن است که در ابتدا، تعداد نمونه‌های این مجموعه ۹۳۴۲ عدد بوده است که ۳ نمونه اضافه‌تر از خانواده Adialer.C داشته است. بااین‌حال به دلیل از دسترس خارج شدن لینک دریافت از دانشگاه کالیفرنیا، مجموعه‌داده موردنظر از تارنمای شخصی گروه مربوطه دریافت شده است که این سه نمونه بدافزار را ندارد [۲۰]. در حال حاضر پژوهش‌هایی که با مجموعه‌داده Malimg کار می‌کنند، همین شرایط را داشته و نبود این سه نمونه را نادیده می‌گیرند. جدول (۱) توزیع بدافزارها در این مجموعه را نشان می‌دهد.

مجموعه‌داده شرکت مایکروسافت دربردارنده بیش از ۲۰ هزار بدافزار از ۹ خانواده مختلف است که طی برگزاری یک رقابت در سال ۲۰۱۵ میلادی به شرکت‌کنندگان ارائه‌شده است [۲۱]. در این مجموعه ۱۰۸۶۸ نمونه برای آموزش و ۱۰۸۷۳ نمونه برای ارزیابی در نظر گرفته‌شده است. البته بنا به دلایل نامعلوم هیچ‌گاه برچسب نمونه‌های ارزیابی به‌صورت عمومی اعلام نشد به همین دلیل در پژوهش‌های مختلف تنها از نمونه‌های آموزش این مجموعه برای ارزیابی الگوریتم‌ها استفاده می‌شود. از برتری‌های این مجموعه که آن را به یک انتخاب ایدئال تبدیل می‌کند، آن است که به همراه فایل دودویی بدافزار، محتوای اسمبلی معکوس آن نیز قرار داده‌شده است. ازاین‌رو برای محققانی که بر روی تحلیل پیشرفته بدافزارها کار می‌کنند، بسیار مناسب است. جدول (۲) توزیع بدافزارها در این مجموعه را نشان می‌دهد.

جدول (۱). توزیع بدافزارها در مجموعه Malimg

تعداد	بدافزار		تعداد	بدافزار
۱۸۴	*Lolyda.AA2*		۱۲۲	*Adialer.C*
۱۲۳	*Lolyda.AA3*		۱۱۶	*Agent.FYI*
۱۵۹	*Lolyda.AT*		۲۹۴۹	*Allaple.A*
۱۳۶	*Malex.gen! J*		۱۵۹۱	*Allaple.L*
۱۴۲	*Obfuscator.AD*		۱۹۸	*Alureon.gen! J*
۱۵۸	*Rbot! gen*		۱۰۶	*Autorun.K*
۸۰	*Skintrim.N*		۲۰۰	*C2LOP.gen! g*
۱۲۸	*Swizzor.gen! E*		۱۴۶	*C2LOP.P*
۱۳۲	*Swizzor.gen! I*		۱۷۷	*Dialplatform.B*
۴۰۸	*VB.AT*		۱۶۲	*Dontovo.A*
۹۷	*Wintrim.BX*		۳۸۱	*Fakerean*
۸۰۰	*Yuner.A*		۴۳۱	*Instantaccess*
۹۳۳۹	**تعداد کل**		۲۱۳	*Lolyda.AA1*

جدول (۲). توزیع بدافزارهای آموزش در مجموعه مایکروسافت

تعداد	بدافزار
۱۵۴۱	*Ramnit*
۲۴۷۸	*Lollipop*
۲۹۴۲	*Kelihos_ver3*
۴۷۵	*Vundo*
۴۲	*Simda*
۷۵۱	*Tracur*
۳۹۸	*Kelihos_ver1*
۱۲۲۸	*Obfuscator.ACY*
۱۰۱۳	*Gatak*
۱۰۸۶۸	**تعداد کل**

۵- یافته‌های پژوهش

پارامترهای تنظیم‌شده برای ارزیابی روش پیشنهادی این پژوهش به شرح زیر است.

۱. **پیش‌پردازش:** از نمونه بدافزارهای موردنظر، سه بردار بایتی، آنتروپی و هیستوگرام تهیه می‌شود.
۲. **بصری‌سازی:** بردارهای سه‌گانه هر بدافزار به روش‌های مختلف بصری‌سازی شده و ۱۵ تصویر مختلف از آن‌ها به دست آورده می‌شود. سپس همه تصاویر به‌دست‌آمده به روش درون‌یابی دوخطی به ابعاد ۲۵۶×۲۵۶ تغییر سایز داده می‌شوند.
۳. **استخراج ویژگی:** به‌وسیله توصیفگر GIST از هریک از

تصاویر به‌دست‌آمده، ۱۲۸ ویژگی استخراج می‌شود که با کنار هم قرار دادن آن‌ها، برداری به طول ۱۹۲۰ ویژگی به دست خواهد آمد.

۴. **طبقه‌بندی:** پس از به‌دست‌آوردن بردار ویژگی برای همه نمونه‌ها، ستون‌های مربوط به هر ویژگی نرمال‌سازی می‌شود. درنهایت طبقه‌بندی SVM بر اساس اعتبارسنجی متقابل ۱۰ لایه مورد آموزش و ارزیابی قرار می‌گیرد.

۵-۱- بررسی جداگانه بردارهای ویژگی

در این پژوهش برای‌آنکه خصوصیات بصری‌سازی‌های مختلف سنجیده شود، بردارهای ویژگی هر بصری‌سازی، جداگانه به طبقه‌بند ارائه‌شده و بر اساس معیار دقت مورد ارزیابی قرار می‌گیرند. جدول (۳) و (۴) دقت دسته‌بندی الگوریتم به ازای بصری‌سازی‌های مختلف را نشان می‌دهد. با مقایسه این دو جدول که مربوط به مجموعه بدافزارهای متفاوتی است، می‌توان مشاهده کرد که لزوماً بصری‌سازی برتر و کامل‌تری بین آن‌ها وجود ندارد. هرکدام از آن‌ها به سهم خود می‌تواند تا حد قابل قبولی (معمولاً ۹۰ درصد) بدافزارها را به خانواده‌های متناظرشان دسته‌بندی کنند.

۵-۲- ترکیب بردارهای ویژگی

هریک از بصری‌سازی‌های معرفی‌شده در این پژوهش، در تشخیص بدافزارهای به خصوصی موفق‌اند که انتظار می‌رود ترکیب بردارهای ویژگی آن‌ها بتواند نتایج بهتری را فراهم کند. ترکیب ویژگی‌های ۱۵ بصری‌سازی مختلف، برداری شامل ۱۹۲۰ ویژگی می‌سازد (۱۵×۱۲۸) که می‌توان آن را مستقیماً به طبقه‌بند ارائه داده و موردبررسی قرارداد. آزمون‌های متعدد نشان داده است که تنها با استفاده از بردار ویژگی ۱۹۲۰ بُعدی، می‌توان بهترین نتیجه را به دست آورد.

شکل (۱۲) جدول سردرگمی الگوریتم برای مجموعه‌داده Malimg را نشان می‌دهد. در این دسته‌بندی، دقت الگوریتم ۹۹/۷۲ درصد به‌دست‌آمده است که مطابق جدول (۵) به‌جز در پژوهش [۲۲] از تمام پژوهش‌های مربوطه جلوتر است. البته همان پژوهش نتوانسته است در ارزیابی مجموعه مایکروسافت با نتایج این پایان‌نامه رقابت کند؛ بنابراین می‌توان استدلال کرد که الگوریتم پیشنهادی این پایان‌نامه درک کامل‌تری از بدافزارها دارد که سبب می‌شود در هردو مجموعه آزمون بسیار موفق باشد. نکته نادیده گرفته‌شده در مجموعه‌داده Malimg آن است که در این مجموعه علاوه بر خانواده‌های مختلف بدافزارها، نسل آن‌ها نیز به‌صورت جداگانه آورده شده است. بااین‌حال تشخیص صحیح نوع خانواده یک بدافزار باید اولویت بیشتری نسبت به تشخیص نسل آن داشته باشد و عواقب انتخاب اشتباه آن نباید در الگوریتم یکسان در نظر گرفته شود. به همین دلیل در ارزیابی دیگری که از مجموعه Malimg انجام‌شده است بدافزارهای هم‌خانواده در یک دسته‌بندی تجمیع شده‌اند. در این حالت میزان دقت الگوریتم پیشنهادی به ۹۹/۸۶ درصد می‌رسد که مطابق شکل (۱۳) عموم اشتباهات الگوریتم در شناسایی نسل‌های خانواده Swizzor و Lolyda برطرف شده است.

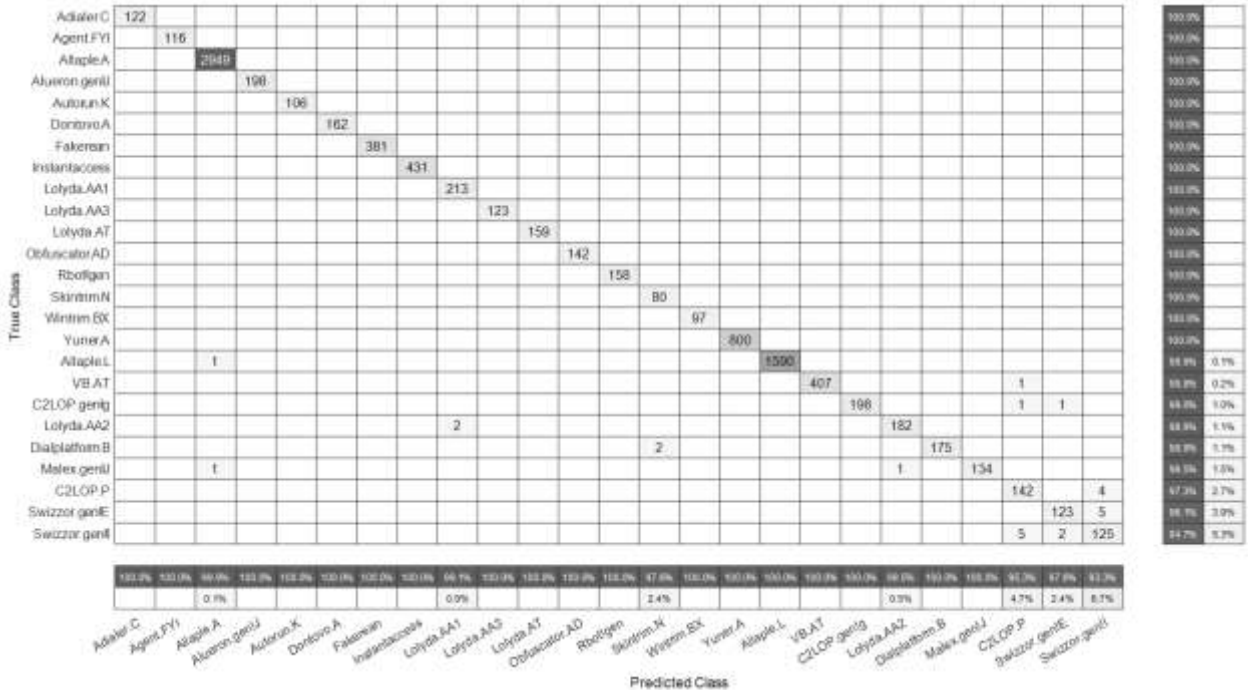
شکل (۱۴) جدول سردرگمی الگوریتم برای مجموعه بدافزارهای مایکروسافت را نشان می‌دهد. در این ارزیابی، دقت الگوریتم ۹۹/۱۶ درصد به‌دست‌آمده است که مطابق جدول (۶) از پژوهش‌های مشابه تحلیل تصویری بدافزار جلوتر است.

جدول (۳). «دقت» دسته‌بندی الگوریتم (به درصد) برای مجموعه بدافزارهای Malimg

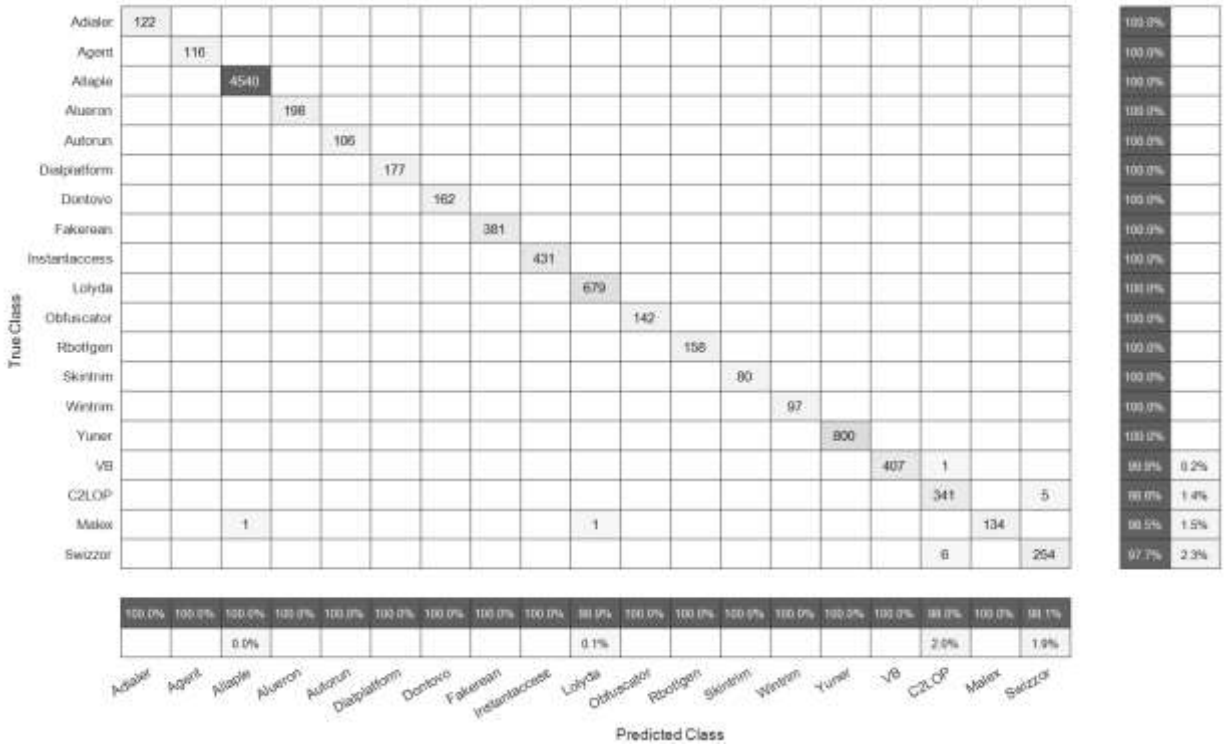
بردار / بصری‌سازی	بایتی	آنتروپی	هیستوگرام
مستقیم	۹۹/۱۰	۹۸/۸۸	۹۹/۳۵
هم رخدادی	۹۰/۵۸	۹۳/۵۰	۹۵/۶۲
نقطه‌ای	۹۸/۸۹	۹۹/۳۳	۹۸/۸۴
هیستوگرام بلوکی	۹۹/۱۹	۹۹/۶۰	۹۹/۲۱
هیستوگرام کامل	۸۰/۸۸	۹۳/۲۵	۹۵/۱۴

جدول (۴). «دقت» دسته‌بندی الگوریتم (به درصد) برای مجموعه بدافزارهای مایکروسافت

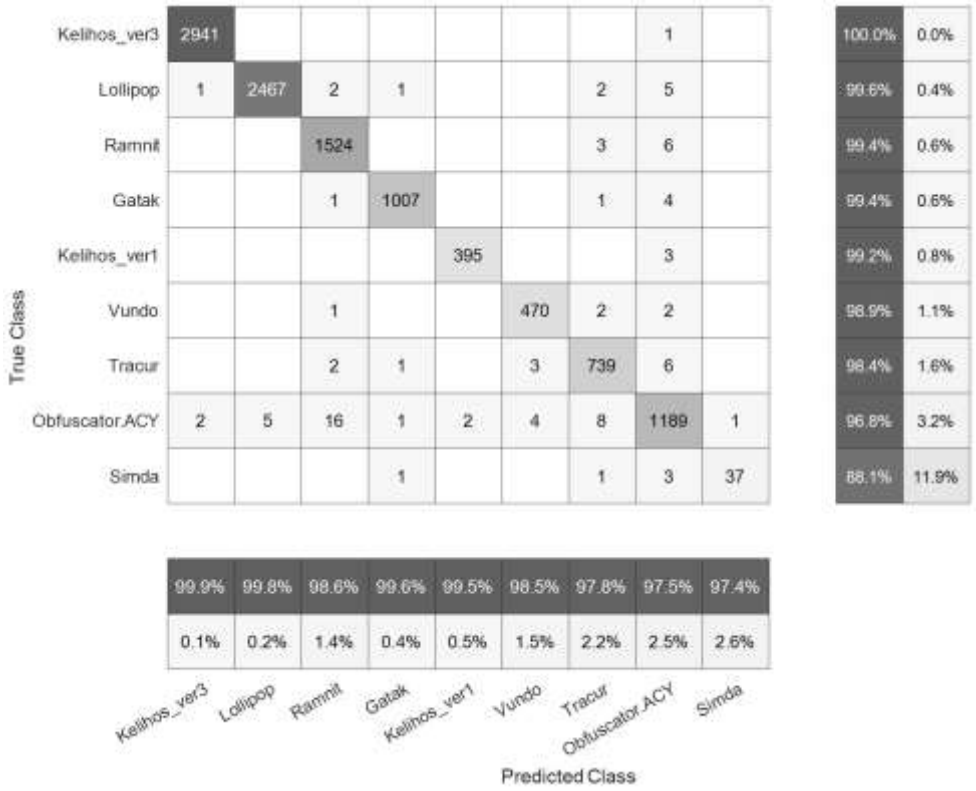
بردار / بصری‌سازی	بایتی	آنتروپی	هیستوگرام
مستقیم	۹۷/۳۱	۹۷/۳۷	۹۷/۹۷
هم رخدادی	۹۶/۶۷	۹۸/۱۸	۹۷/۲۵
نقطه‌ای	۹۷/۱۲	۹۷/۳۱	۹۶/۷۳
هیستوگرام بلوکی	۹۶/۶۳	۹۸/۲۷	۹۷/۲۳
هیستوگرام کامل	۹۶/۶۶	۹۸/۱۱	۹۷/۱۵



شکل (۱۲). جدول سردرگمی الگوریتم برای مجموعه بدافزارهای Malimg



شکل (۱۳). جدول سردرگمی الگوریتم برای مجموعه بدافزارهای Malimg (بدون درنظرگرفتن نسل بدافزارها)



شکل (۱۴). جدول سردرگمی الگوریتم برای مجموعه بدافزارهای Microsoft

جدول (۵). دقت الگوریتم‌های مختلف بر اساس مجموعه Malimg

دقت الگوریتم	پژوهش‌های مرتبط
۹۸/۰۸	Nataraj L et al. 2011 [6]
۹۸/۶۲	Rezende E et al. 2017 [23]
۹۸/۶۳	Yue S. 2017 [24]
۹۴/۵۰	Cui Z et al. 2018 [10]
۹۸/۴۰	Gibert D et al. 2019 [25]
۹۸/۴۰	Naeem H et al. 2019 [26]
۹۸/۸۲	Vasan D et al. 2020 [27]
۹۸/۵۰	Wang C et al. 2021 [28]
۹۹/۵۰	Vasan D et al. 2020 [8]
۹۹/۸۳	Guo H et al. 2020 [22]
۹۹/۷۲	**روش پیشنهادی**

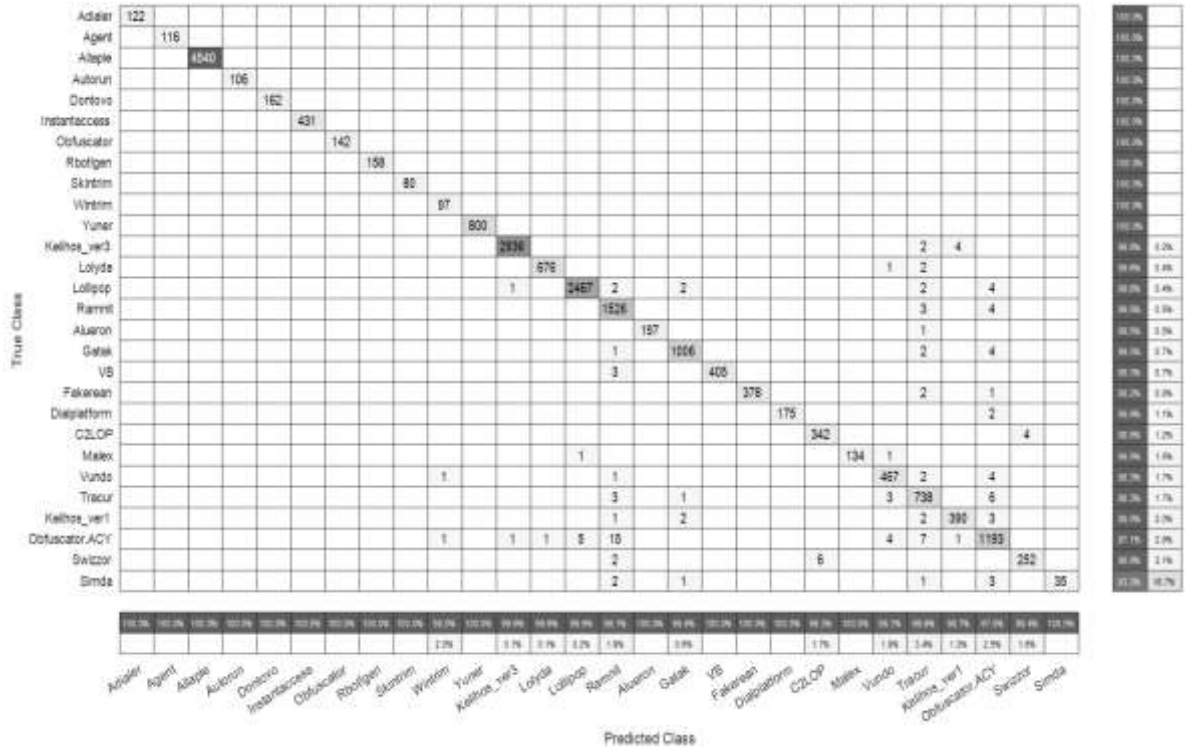
جدول (۶). دقت الگوریتم‌های مختلف بر اساس مجموعه مایکروسافت

دقت الگوریتم	پژوهش‌های مرتبط
۹۶/۶۰	Narayanan BN et al. 2016 [29]
۹۸/۴۰	Gibert D et al. 2019 [25]
۹۷/۳۰	Wang C et al. 2021 [28]
۹۷،۸۸	Guo H et al. 2020 [22]
۹۸،۷۴	Santacroce ML et al. 2020 [30]
۹۹،۱۶	**روش پیشنهادی**

۵-۳- ترکیب مجموعه‌ها

ترکیب مجموعه‌ها برای ساخت فضای نمونه گسترده‌تر، آخرین مرحله ارزیابی این پژوهش محسوب می‌شود که تاکنون در هیچ پژوهش دیگری انجام‌نشده است. ازیک‌طرف به دلیل نبود مجموعه‌داده‌های متنوع، پژوهش‌های سالیان گذشته تنها به دو مجموعه Malimg و Microsoft محدود بوده و از طرف دیگر مجموعه Malimg به دلیل بصری‌سازی پیش‌فرض خود، هیچ‌گاه

قابل‌ترکیب با مجموعه مایکروسافت نبوده است. بااین‌حال در این پژوهش نشان داده شد که می‌توان در مجموعه Malimg، محتوای دودویی نمونه‌ها را از تصاویرشان قابل‌استخراج کرد که این موضوع کمک می‌کند آن‌ها را در کنار مجموعه مایکروسافت با هر بصری‌سازی دلخواه مورداستفاده قرارداد. بدین ترتیب نه‌تنها توانایی الگوریتم در طیف وسیع‌تری از بدافزارها مورد آزمایش قرار می‌گیرد؛ بلکه مسیر جدیدی برای پژوهش‌های بعدی فراهم می‌شود. شکل (۱۵) جدول سردرگمی الگوریتم برای مجموعه ترکیبی جدید را نشان می‌دهد. این مجموعه ۲۰۱۹۹ نمونه از ۲۸ خانواده متفاوت دارد که دقت دسته‌بندی الگوریتم برای آن ۹۹/۳۷ درصد ارزیابی‌شده است. با مقایسه شکل (۱۵) با شکل‌های (۱۲، ۱۳ و ۱۴) می‌توان مشاهده کرد که مشخصه «پوشش» نمونه‌ها بدون تغییر مانده است و الگوریتم با بزرگ‌تر شدن فضای نمونه‌ای، یادگیری قبلی خود را از دست نداده است.

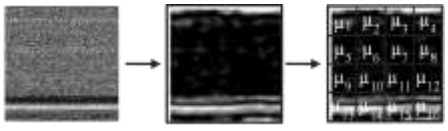


شکل (۱۵). جدول سردرگمی الگوریتم از ترکیب مجموعه بدافزارهای Malimg و Microsoft

۵-۴- اعتبار و مزایای روش پیشنهادی

روش پیشنهادی این پژوهش بااینکه فرایند آموزش و ارزیابی ساده‌تری نسبت به پژوهش‌های مشابه دارد؛ اما در هنگام ارزیابی می‌تواند هم‌تراز آن‌ها عمل کند. اکثر پژوهش‌های مشابه از روش‌های یادگیری عمیق استفاده کرده‌اند که به‌مراتب پیچیدگی زمانی بیشتری نسبت به روش این پژوهش دارند و ازلحاظ عملکرد در بسیاری از زمینه‌های علمی نیز پیشتاز هستند. با این تفاصیل اگر بتوان تشخیص داد که چه ویژگی‌هایی در تشخیص بدافزارها مؤثرند می‌توان تاحدامکان، روش‌های ساده‌تر و بهینه‌تری را برای دسته‌بندی بدافزارها به کار گرفت.

شکل (۱۶) نمونه‌ای از فرایند استخراج ویژگی از تصویر یک بدافزار دلخواه را نشان می‌دهد. مطابق شکل، تصویر بدافزار بعد از اعمال فیلتر به ۱۶ بلوک تفکیک‌شده و از هریک میانگین گرفته می‌شود تا برداری متشکل از ۱۶ ویژگی بسازد. در این فرایند ویژگی‌های استخراج‌شده به دستورات به‌کاررفته در فایل اشاره ندارند بلکه ساختار فایل اعم از نوع داده‌های به‌کاررفته و نحوه قرارگیری آن‌ها برای الگوریتم قابل‌تشخیص می‌شود. این موضوع نشان می‌دهد که الگوریتم‌های تحلیل تصویری بدافزار، برخلاف ادعاهای مطرح‌شده در پژوهش‌های پیشین، الگوهای مخرب را جستجو نمی‌کنند، بلکه الگوهای انسانی را دنبال می‌کنند که سازندگان بدافزار از روی عادت مرتکب می‌شوند. الگوها و اشتباهات انسانی که می‌توانند به دلیل استفاده از ابزارهای برنامه‌نویسی یکسان یا رونوشت کُدهای مخربی باشد که خود فرد آن را توسعه داده و به‌عنوان امضای کاری خود بکار می‌گیرد.



شکل (۱۶). مراحل استخراج ویژگی از تصویربردار بایتی یک بدافزار

استفاده از روش یادگیری عمیق که امروز در کانون توجهات قرار دارد، سبب شده که پژوهشگران از تحلیل جزئیات مسئله بازبمانند و با سپردن استخراج ویژگی به این جعبه سیاه، تحلیل‌هایی را به آن نسبت دهند که لزوماً با واقعیات مسئله تطابق نداشته باشد. البته این موضوع، موفقیت شبکه‌های عصبی

عمیق را انکار نمی‌کند و چه‌بسا در مجموعه‌داده‌های وسیع‌تر نیاز به یک استخراج گر ویژگی عمومی‌تر مثل شبکه‌های عصبی عمیق احساس شود. درهرصورت تا بدین جای کار، نتایج این پژوهش نشان می‌دهد که حتی برای مجموعه‌ای به وسعت ۲۰ هزار بدافزار می‌توان با بهبود بصری‌سازی بدافزارها به نتیجه یکسان دست‌یافت. نکته نگران‌کننده دراین‌باره آن است که الگوریتم این پایان‌نامه و پژوهش‌های پیشین توانسته است بدافزارها را نه بر اساس الگوهای مخرب، بلکه بر اساس اشتباهات انسانی شناسایی کند. ازآنجاکه سازندگان بدافزار همچون محققان درزمینه کاری خود بسیار پیگیر هستند، درصورتی‌که از این نتایج مطلع شوند، روش‌های پیشگیرانه‌ای را برای محافظت از بدافزارهای خود ترتیب خواهند داد که این موضوع سبب می‌شود، در آینده، استفاده از بصری‌سازی‌های این پژوهش و توصیفگر ساده‌ای همچون GIST برای جستجوی الگوها کافی نباشد.

۶- نتیجه‌گیری

دقت دسته‌بندی بالا و سبک بودن محاسبات الگوریتم، از نقاط قوت روش پیشنهادی محسوب می‌شود که آن را هم‌تراز با روش‌های مشابه خود قرار می‌دهد. ارزیابی روش این پژوهش با مجموعه بدافزارهای Malimg و Microsoft نشان می‌دهد که نه‌تنها ازلحاظ دقت، توانایی روش فعلی با روش یادگیری عمیق برابری می‌کند؛ بلکه حتی با ترکیب این مجموعه بدافزارها نیز توانایی آن همچنان قابل‌قبول است، بدون آنکه نیاز به بازنگری پارامترها یا آموزش‌های طولانی مجدد باشد. در کنار مزایای ذکرشده، سادگی استفاده از توصیفگر GIST و طبقه‌بند SVM نیز باعث شده است که نگاه دقیق‌تری به سازوکار داخلی آن شود و اعتبار تحلیل تصویری بدافزار بیش‌ازپیش مورد بررسی قرار گیرد.

تحلیل تصویری بدافزار باوجود گذشت یک دهه، همچنان در ابتدای راه خود به سر می‌برد. علی‌رغم طیف وسیع خلاقیت‌ها و ابتکاراتی که می‌توان در این مسیر بکار برد، همچنان پژوهش‌های کمی وجود دارند که به طور اساسی به این مقوله پرداخته باشند. در اغلب موارد، چالش مربوطه بدون تحلیل دقیق‌تر و تنها با به‌کارگیری یک طبقه‌بند پیشرفته پشت سر گذاشته شده است. این موضوع سبب شد که در ابتدای مسیر این پژوهش، به یافته‌های پژوهش‌های قبلی با دیده شک‌وتردید نگاه شود. چراکه اکثر آن‌ها همان پژوهش قبلی اما با یک الگوریتم بهبودیافته‌تر بوده‌اند بدون آنکه تفسیری از سازوکار آن داشته باشد. استفاده از توصیفگر GIST به‌عنوان استخراج گر ویژگی این پژوهش، درنتیجه تلاش‌هایی بوده است که در ابتدا برای تولید نتایج پژوهش‌های قبلی انجام‌گرفته است؛ اما درنهایت به دلیل کاربردی‌بودن آن، به‌عنوان الگوریتم اصلی پژوهش حفظ‌شده است. به لطف بار پردازشی کمی که توصیفگر GIST به همراه داشته است، این فرصت فراهم شد که تمرکز پژوهش بر روی ساخت انواع بصری‌سازی‌ها قرار گیرد. بصری‌سازی‌های مختلف طی آزمون‌وخطاهای بسیار مورد بررسی قرار گرفتند تا به آنچه اکنون در اختیار است، برسند البته در ابتدای امر تمام تلاش‌ها به این منظور صورت‌گرفته است که با کشف یک بصری‌سازی مناسب، الگوریتم نهایی از توصیفگر GIST به یک شبکه عصبی پیچشی مهاجرت داده شود؛ اما به‌مرورزمان مشخص شد که حتی اگر بهترین طبقه‌بندی‌ها بکار گرفته شود، یک بصری‌سازی تنها نمی‌تواند راه‌حلی برای همه مشکلات باشد؛ بنابراین رویکرد این پژوهش ناخواسته به سمت ترکیب ویژگی‌ها سوق داده شد که توانست به نتایج بسیار بهتری ختم شود.

الگوریتم فعلی این پژوهش این استعداد را دارد که با جایگزینی استخراج گر GIST و طبقه‌بند SVM با شبکه‌های عصبی عمیق، بهبود یابد. اکنون با در اختیار داشتن مجموعه‌ای از ۲۰ هزار بدافزار، این فرصت فراهم است که بدون نگرانی از بابت نامتوازن بودن مجموعه‌ها، با آزادی عمل بیشتری به آموزش الگوریتم‌های یادگیری عمیق پرداخته شود. البته مطالعه مقالات مختلف نیز نشان می‌دهد که طراحی یک شبکه عصبی پیچشی از پایه، لزوماً راهبرد سریعی نیست. روش‌های یادگیری عمیق همه‌روزه در حال به‌روزرسانی هستند و متخصصان بسیاری از سراسر دنیا در توسعه آن نقش دارند. استفاده از شبکه‌های عصبی پیچشی که به‌صورت انتقالی آموزش می‌بینند نه‌تنها فرایند آموزش آن سرعت می‌بخشد بلکه اجازه می‌دهد که تمرکز پژوهش بر روی بهبود بصری سازی‌ها و روش‌های بصری‌سازی بهتر معطوف شود. بدین ترتیب با معرفی هر شبکه جدیدتر و تواناتر، الگوریتم مربوطه می‌تواند به ساختارهای جدید مهاجرت کند.

۷- مراجع

[1] R. Moir, "Defining Malware: FAQ," 2009. [Online]. Available: https://technet.microsoft.com/en-us/library/dd632948.aspx

[2] C. C. Elisan, *Advanced malware analysis*. McGraw Hill Professional, 2015.

[3] AV-TEST - The Independent IT-Security Institute, "Malware Statistics & Trends Report," 2022. [Online]. Available: https://www.av-test.org/en/statistics/malware/

[4] D. Simpson, "Malware names." [Online]. Available: https://docs.microsoft.com/en-us/microsoft-365/security/intelligence/malware-naming

[5] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," *J. Netw. Comput. Appl.*, vol. 153, p. 102526, 2020, doi: 10.1016/j.jnca.2019.102526.

[19] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001, doi: 10.1023/A:1011139631724.

[20] SARVAM Team, “Supervised Classification with k-fold Cross Validation on a Multi Family Malware Dataset.” [Online]. Available: https://sarvamblog.blogspot.com/2014/08/supervised-classification-with-k-fold.html

[21] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, and M. Ahmadi, “Microsoft malware classification challenge,” *arXiv Prepr. arXiv1802.10135*, 2018.

[22] H. Guo, S. Huang, C. Huang, Z. Pan, M. Zhang, and F. Shi, “File entropy signal analysis combined with wavelet decomposition for malware classification,” *IEEE Access*, vol. 8, pp. 158961–158971, 2020, doi: 10.1109/ACCESS.2020.3020330.

[23] E. Rezende, G. Ruppert, T. Carvalho, F. Ramos, and P. De Geus, “Malicious software classification using transfer learning of resnet-50 deep neural network,” in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, pp. 1011–1014. doi: 10.1109/ICMLA.2017.00-19.

[24] S. Yue, “Imbalanced malware images classification: a CNN based approach,” *arXiv Prepr. arXiv1708.08042*, 2017.

[25] D. Gibert, C. Mateu, J. Planes, and R. Vicens, “Using convolutional neural networks for classification of malware represented as images,” *J. Comput. Virol. Hacking Tech.*, vol. 15, no. 1, pp. 15–28, 2019, doi: 10.1007/s11416-018-0323-0.

[26] H. Naeem, B. Guo, M. R. Naeem, F. Ullah, H. Aldabbas, and M. S. Javed, “Identification of malicious code variants based on image visualization,” *Comput. \& Electr. Eng.*, vol. 76, pp. 225–237, 2019, doi: 10.1016/j.compeleceng.2019.03.015.

[27] D. Vasan, M. Alazab, S. Wassan, H. Naeem, B. Safaei, and Q. Zheng, “IMCFN: Image-based malware classification using fine-tuned convolutional neural network architecture,” *Comput. Networks*, vol. 171, p. 107138, 2020, doi: 10.1016/j.comnet.2020.107138.

[28] C. Wang, Z. Zhao, F. Wang, and Q. Li, “A novel malware detection and family classification scheme for IoT based on DEAM and DenseNet,” *Secur. Commun. Networks*, vol. 2021, 2021, doi: 10.1155/2021/6658842.

[29] B. N. Narayanan, O. Djaneye-Boundjou, and T. M. Kebede, “Performance analysis of machine learning and pattern recognition algorithms for malware classification,” in *2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS)*, 2016, pp. 338–342. doi: 10.1109/NAECON.2016.7856826.

[30] M. L. Santacroce, D. Koranek, and R. Jha, “Detecting malware code as video with compressed, time-distributed neural networks,” *IEEE Access*, vol. 8, pp. 132748–132760, 2020, doi: 10.1109/ACCESS.2020.3010706.

[6] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, “Malware images: visualization and automatic classification,” in *Proceedings of the 8th international symposium on visualization for cyber security*, 2011, pp. 1–7. doi: 10.1145/2016904.2016908.

[7] L. Chen, R. Sahita, J. Parikh, and M. Marino, “STAMINA: Scalable Deep Learning Approach for Malware Classification,” 2020. [Online]. Available: https://www.intel.com/content/dam/www/public/us/en/ai/documents/stamina-scalable-deep-learning-whitepaper.pdf

[8] D. Vasan, M. Alazab, S. Wassan, B. Safaei, and Q. Zheng, “Image-Based malware classification using ensemble of CNN architectures (IMCEC),” *Comput. \& Secur.*, vol. 92, p. 101748, 2020, doi: 10.1016/j.cose.2020.101748.

[9] L. Chen, “Deep transfer learning for static malware classification,” *arXiv Prepr. arXiv1812.07606*, 2018.

[10] Z. Cui, F. Xue, X. Cai, Y. Cao, G. Wang, and J. Chen, “Detection of malicious code variants based on deep learning,” *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3187–3196, 2018, doi: 10.1109/TII.2018.2822680.

[11] E. Bastami, P. Keshavarzi, M. Rahmanimanesh, and H. Soltanizadeh, “A Malware Classification Method Using visualization and Word Embedding Features,” *Electron. Cyber Def.*, 2023, https://dor.isc.ac/dor/20.1001.1.23224347.1402.11.1.1.2

[12] R. Lyda and J. Hamrock, “Using entropy analysis to find encrypted and packed malware,” *IEEE Secur. \& Priv.*, vol. 5, no. 2, pp. 40–45, 2007, doi: 10.1109/MSP.2007.48.

[13] X. Ugarte-Pedrero, I. Santos, B. Sanz, C. Laorden, and P. G. Bringas, “Countering entropy measure attacks on packed software detection,” in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, 2012, pp. 164–168. doi: 10.1109/CCNC.2012.6181079.

[14] D.-L. Vu, T.-K. Nguyen, T. V Nguyen, T. N. Nguyen, F. Massacci, and P. H. Phung, “HIT4Mal: Hybrid image transformation for malware classification,” *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 11, p. e3789, 2020, doi: 10.1002/ett.3789.

[15] J. Fu, J. Xue, Y. Wang, Z. Liu, and C. Shan, “Malware visualization for fine-grained classification,” *IEEE Access*, vol. 6, pp. 14510–14523, 2018, doi: 10.1109/ACCESS.2018.2805301.

[16] Z. Ren, G. Chen, and W. Lu, “Malware visualization methods based on deep convolution neural networks,” *Multimed. Tools Appl.*, vol. 79, no. 15, pp. 10975–10993, 2020.

[17] J. Kim, J.-Y. Paik, and E.-S. Cho, “Attention-Based Cross-Modal CNN Using Non-Disassembled Files for Malware Classification,” *IEEE Access*, vol. 11, pp. 22889–22903, 2023, doi: 10.1109/ACCESS.2023.3253770.

[18] K. Shaukat, S. Luo, and V. Varadharajan, “A novel deep learning-based approach for malware detection,” *Eng. Appl. Artif. Intell.*, vol. 122, p. 106030, 2023, doi: 10.1016/j.engappai.2023.106030.