تشخیص فایل‌های مخرب غیراجرایی به کمک روش‌های یادگیری ماشین

رسول رضوانی جلال
کارشناسی ارشد، دانشکده مهندسی کامپیوتر
دانشگاه علم و صنعت ایران
تهران، ایران
(رایانامه: rasoul_rezvanijalal@comp.iust.ac.ir)

مرتضی ذاکری
استادیار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر
پژوهشگر، پژوهشکده مهندسی کامپیوتر، پژوهشگاه دانش‌های بنیادی
تهران، ایران
(رایانامه: zakeri@aut.ac.ir)

چکیده

با افزایش استفاده کاربران از فایل‌های آفیس و پی‌دی‌اف و استفاده این نوع فایل‌ها در مراکز امنیتی جهت انتقال اطلاعات، توجه طراحان بدافزار به این فایل‌ها جلب شده است. فعالیت‌ها و مطالعات گوناگونی جهت تشخیص این نوع فایل‌ها با انتخاب ویژگی‌های تعیین کننده هر نوع فایل صورت گرفته است. در این پژوهش سعی شده است تا با تهیه مجموعه داده تقویت‌شده و همچنین ارائه ویژگی‌های مؤثر و جامع برای انواع فایل‌های ذکر شده، به‌طوری که حملات مربوط به هر نوع فایل را پوشش دهند، نرخ تشخیص بدافزارهای مربوطه افزایش داده شود. در این راستا، این مطالعه توانست با بکارگیری مدل‌های طبقه‌بند دودویی، در مقایسه با برترین مطالعات انجام شده، به بهبود ۲ درصدی تشخیص بدافزارهای پی دی اف با مدل گرادیان افزایشی و همچنین بهبود ۱/۹ درصدی تشخیص بدافزارهای آفیس با مدل جنگل تصادفی دست پیدا کند. به‌طور دقیق‌تر، این مطالعه با اعمال مدل گرادیان افزایشی بر روی فایل‌های پی دی اف، توانست به دقت ۹۹/۳ درصدی تشخیص بدافزارها دست پیدا کند در حالی که برای فایل‌های آفیس با اعمال مدل جنگل تصادفی، به نرخ ۹۹/۴ درصدی در تشخیص بدافزار رسیده شد.

کلیدواژه‌ها

تشخیص بدافزار، فایل‌های پیچیده و غیراجرایی، یادگیری ماشین، فایل‌های پی دی اف، فایل‌های آفیس، طبقه‌بندی دودویی

۱. مقدمه

تهدیداتی که بدافزارها در سراسر جهان ایجاد می‌کنند به سرعت در حال افزایش است. بدافزار نرم‌افزاری است که بدون اطلاع به صورت مخفیانه به سیستم قربانی نفوذ می‌کند و قصد بدی برای ایجاد اختلال در عملکرد رایانه قربانی دارد. از طرفی سرویس‌های آفیس از شرکت مایکروسافت پرکاربردترین مجموعه برای پردازش اسناد، صفحات گسترده و ارائه‌ها است که به دلیل محبوبیت آن، به طور مداوم برای انجام فعالیت های های مخرب مورد استفاده قرار می‌گیرد. خراب کاران با سوء استفاده از ویژگی های پویای این سرویس‌ها، از آن برای راه‌اندازی حملات خود و نفوذ به میلیون‌ها میزبان در فعالیت‌های خراب‌کارانه خود استفاده می‌کنند. همان‌طور که توسط منابع مختلف مانند Avira و Eset [۳، ۴] گزارش شده است اسناد آفیس دومین قالب از میان چهار قالب پرکاربرد است که توسط بدافزارها در محیط ویندوز استفاده میشود [۲]. همچنین در طول سالیان اخیر، فایل‌ها با فرمت پی دی اف به دلیل انعطاف‌پذیری و ویژگی های کار آسان، به محبوبترین فرمت ارائه محتوا در بین کاربران تبدیل شده است که بر اساس آن‌چه در مورد فایل‌های آفیس توضیح داده شد، مورد سوء استفاده خراب‌کاران قرار می‌گیرند. لذا وجود یک ساز و کار جهت شناسایی این نوع بدافزارها بیش از پیش ضروری است.

به طور کلی، عامل اصلی در دقت عملکرد و در نتیجه موفقیت مدل های یادگیری، فضای ویژگی است که برای جستجوی هدف

ششمین کنفرانس ملی انفورماتیک ایران، ۸ و ۹ اسفند ۱۴۰۳

مورد استفاده قرار می گیرد .بنابراین، اثربخشی شناسایی فعالیت‌های مخرب در فایل‌های غیرقابل اجرا، مانند اسناد آفیس، به شناسایی الگوهای شامل اسکریپت‌ها، پیوندها، تصاویر و تعامل بین کدهای جاوا اسکریپت و وی بی اسکریپت بستگی دارد. هدف این مقاله ارتقای قابلیت‌های تشخیص بدافزار پی دی اف و آفیس است. این کار با جمع‌آوری مجموعه داده از انواع فایل های مختلف، از جمله pdf، doc، docx، xls، xlsx، ppt و pptx و همچنین معرفی ویژگی‌های جدید برای پوشش طیف گسترده‌ای از حملات انجام می‌شود. در ادامه این فرآیند با ایجاد مدل‌های طبقه‌بندی دودویی از این مجموعه داده و تنظیم دقیق فراپارامترها برای رسیدن به دقت تشخیص حداکثری ادامه می‌یابد. به‌طور خلاصه نوآوری‌های اصلی این مطالعه به شرح زیر است:

۱. با مطالعه فرمت فایل‌های پیچیده و مهندسی ویژگی‌های این نوع فایل ها، ویژگی های متمایز و متنوعی را برای هر نوع فایل انتخاب کردیم و تشخیص بدافزار مرتبط را در مقایسه با تلاش‌های قبلی افزایش دادیم.

۲. با انجام آزمایش های خودکار بر روی طیف وسیعی از فراپارامترها در مدل‌های مختلف و استفاده از اعتبارسنجی k-fold، ما مؤثرترین پیکربندی‌ها و فراپارامتر را مشخص کردیم و قابلیت اطمینان خروجی‌های مدل خود را تضمین می‌کنیم.

۳. ما مجموعه ای بزرگ از فایل های مخرب و سالم را برای هر دو فرمت پی دی اف و آفیس گردآوری کردیم. با بهره‌گیری از منابع گوناگون، ما مجموعه داده‌های موجود را از لحاظ کمی و کیفی ارتقاع دادیم. به این ترتیب آموزش بر روی یک مجموعه داده گسترده انجام می‌شود و نتایج مدل‌های طبقه‌بندی دودویی قابل اعتمادتر است.

در بخش بعدی، مروری بر تحقیقات موجود در مورد بدافزار ارائه می‌کنیم که به صراحت بر اسناد آفیس و فایل‌های پی دی اف تمرکز دارد. این مطالعه به بررسی عناصر ساختاری آن ها و تکنیک‌های تشخیص مورد استفاده می‌پردازد. متعاقباً، ما به روش به کار گرفته شده در این مطالعه می پردازیم و جزئیات جمع آوری داده ها، اعتبار سنجی، استخراج ویژگی، و فرآیندهای یادگیری را در بخش ۳ بیان می کنیم. بخش ۴ یک نمای کلی از مجموعه داده پیشنهادی ما، از جمله تجزیه و تحلیل آماری آن و بحثی درباره آمار مجموعه داده ارائه می‌کند. بخش ۵ یافته ها و آزمایش های ما را تحلیل می‌کند. در نهایت، بخش ۶ یک نتیجه‌گیری و جمع‌بندی از مطالعه انجام شده را، به‌طور خلاصه بیان می‌کند.

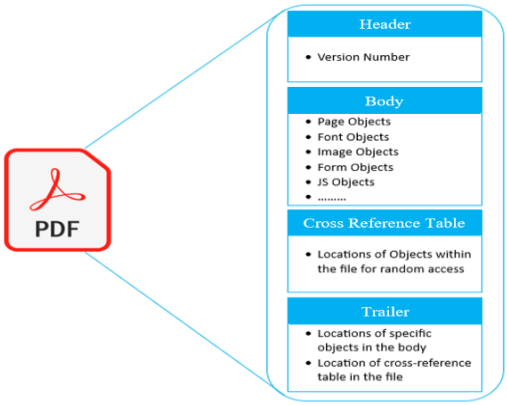
۲. کارهای مرتبط

این بخش یک مرور کلی از کار مرتبطی که زیربنای مطالعه ما است ارائه می‌دهد. در ابتدا، ساختار اسناد آفیس و فایل‌های پی دی اف را بررسی می‌کنیم و یک پیش‌زمینه جامع ارائه می‌کنیم. متعاقباً، روش‌های دفاعی مورد استفاده در برابر بدافزارهای این دسته را تشریح می‌کنیم.

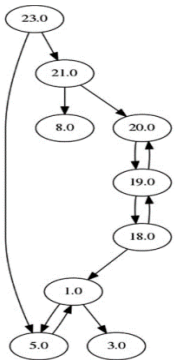
۲-۱. فایل‌های پی دی اف

قبل از پرداختن به بررسی موارد مربوط به شناسایی بدافزارها، ضروری است که با فایل‌های پی دی اف آشنا شویم. ساختار کلی فایل های پی دی اف در شکل ۱ نشان داده شده است. این مطالعه بر پیچیدگی این نوع فایل ها تاکید می‌کند. فایل های پی‌دی‌اف، ساختار سلسله مراتبی دارند، که می‌توان آن‌ها را به صورت یک گراف جهت‌دار نشان داد. همان‌طور که در شکل ۲ نشان داده شده است، هر گره در این نمودار با یک مقدار عددی که شماره شی در ترتیب اشیاء موجود در فایل را نشان می‌دهد، شناسایی می‌شود. این مفهوم در مطالعه‌ای [۵] توضیح داده شده است، که فرض می‌کند هر گره در این نمودار نماد یک شی منفرد در فایل پی دی اف است. برای بررسی دقیق تر این نمودار، شکل ۳ مربوط به مطالعه Šrndic و همکاران [۶] را می‌توان در نظر گرفت که تجسم مناسبی از ساختار فایل پی دی اف ارائه می‌دهد.

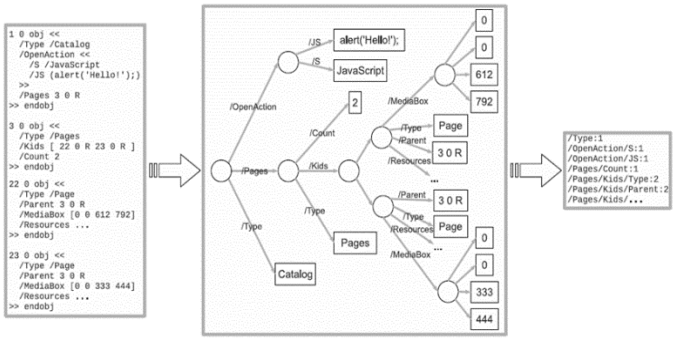
اسنادی مانند فایل‌های پی دی اف به دلیل ظرفیت آن‌ها برای ترکیب انواع محتوا، به عنوان فایل های پیچیده طبقه‌بندی شوند. مؤلفه قابل توجه محتوای ذکر شده کد جاوا اسکریپت است که قابلیت این نوع فایل ها را افزایش می دهد. در نتیجه، مطالعات متعددی بر روی شناسایی فایل‌های پی دی اف مخرب با شناسایی کد جاوا اسکریپت متمرکز شده‌اند. مطالعه ای توسط جیاکسیانگ گو و همکاران مجموعه ای از توکن های مشتق شده از متغیرهای کد جاوا اسکریپت را به عنوان ویژگی های فایل معرفی می‌کند و با استفاده از مدل SVM به دقت ۹۶/۹۳ درصد دست می‌یابد[۷].



شکل ۱: نمای کلی فایل پی دی اف



شکل ۲: فایل پی دی اف به شکل گراف جهت‌دار [۵]



شکل ۳: ساختار فیزیکی در مقابل ساختار سلسله مراتبی فایل پی دی اف

به طور مشابه، یک مطالعه مرتبط [۸] از الگوریتمی به نام PJscanبرای ترجمه کد جاوا اسکریپت به توکن‌ها استفاده کرد که دقت ۸۵٪ را به دست آورد. علاوه بر این، قدرت تشخیص با ترکیب ویژگی‌های عمومی اضافی مرتبط با جاوا اسکریپت افزایش یافته است، و با استفاده از تکنیک یادگیری پشته‌ای با رگرسیون لجستیک فرا یادگیرنده (LR) به قدرت تشخیص ۹۹/۸۹ درصد دست می‌یابد [۱]. این رویکرد با پرداختن به نظارت بر پیش پردازش در مطالعات قبلی برتری خود را نشان می‌دهد.

مطالعه اخیر [۹] بر روی مجموعه داده CIC-Evasive-PDFMal2022، با استفاده از الگوریتم درخت تصمیم بهینه سازی (O_DT) انجام شده است که به دقت تشخیص ۹۸/۸۴ درصد دست یافته است. علاوه بر این، مجموعه‌های مختلفی از ویژگی‌ها برای فایل های پی دی اف پیشنهاد شده‌اند [۶]، که هر فایل را به مجموعه‌ای از مسیرها یا مجموعه‌ای از مسیرها تبدیل می‌کند، جایی که هر مسیر نشان‌دهنده یک ویژگی از فایل است. این روش، با استفاده از مدل‌های SVM و DT، به دقت بالایی در تشخیص بدافزار دست یافته است.

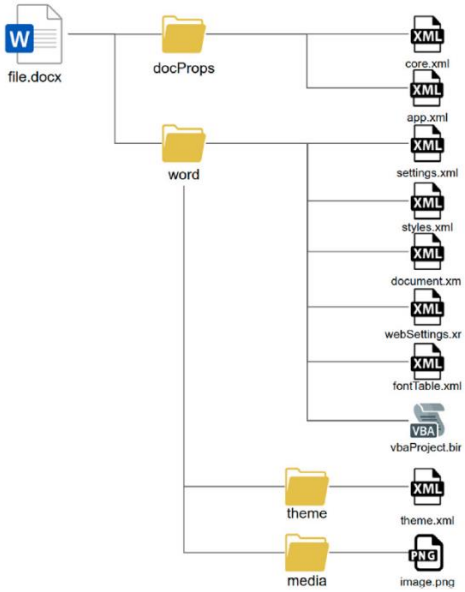
۲-۲. اسناد آفیس

درک عمیق از ترکیب ساختاری اسناد اداری در تجزیه و تحلیل اسناد ضروری است. این ضرورت توسط تحقیقات علمی در این حوزه، مانند تحقیقهای انجام شده توسط کوتسوکوستاس و همکاران [۲]، و سینگ و همکاران. [۵] که توضیحات مفصلی از سازماندهی سلسله مراتبی این انواع فایل ارائه کرده است، تأکید می شود. همانطور که در شکل ۴ نشان داده شده است، واضح است که اسناد آفیس ساختاری درخت مانند شبیه فایل های پی دی اف را نشان می دهند. این شباهت ساختاری بر پیچیدگی ذاتی در هر دو نوع سند تأکید می کند.

مطالعات و فعالیت های زیادی برای شناسایی و تحلیل اسناد اداری با تاکید بر بررسی کدهای ماکرو انجام شده است. به عنوان مثال، در مطالعه ای [۱۰] کد VBA با استفاده از کتابخانه Olevba در پایتون کاوش شده است اما مدل مورد استفاده بر اساس مدل‌های یادگیری ماشین ارزیابی خاصی را انجام نداده است. پژوهش دیگری [۱۱] از الگوریتم‌های مبتنی بر NLP، برای ساخت بردارهای ویژگی برای هر نمونه استفاده کرده است و با مدل SVM به F-score ۹۳ درصدی دست یافت. چالش شناسایی ماکروهای VBA مبهم در یک مطالعه [۱۲]، با استفاده از ویژگی‌های استاتیک مختلف با دستیابی به نرخ دقت ۹۰ درصد مورد بررسی قرار گرفته است. راوی و همکاران در مطالعه خود [۱۳] ماکروهای مبهم را با استفاده از روشی به نام obfuscated_word2vec کاوش کردند. آن‌ها ویژگی‌هایی مانند فراخوانی های API و متغیرهای مبهم را استخراج کردند و به نرخ دقت ۸۲/۶۵ درصد دست یافتند. کوتسوکوستاس و همکاران [۲] مطالعه خود را در مورد تجزیه و تحلیل ماکروهای VBA گسترش دادند و از تبادل پویای داده DDE استفاده کردند،

ویژگی ای که تبادل داده بین اسناد اداری مختلف را تسهیل می‌کند. این مطالعه به دقت تشخیص ۹۷/۵ درصد با مدل جنگل تصادفی RF دست یافت.

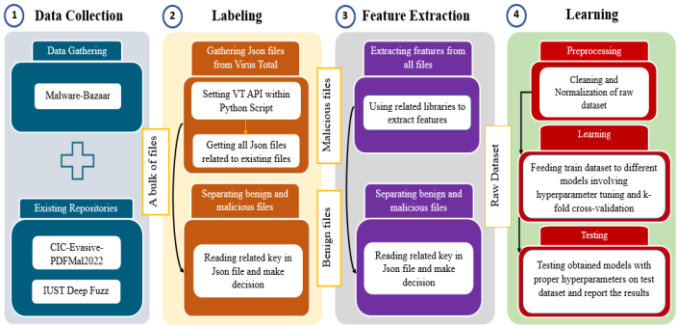
علاوه بر این، روشی توسعه یافته است [۱۴] که تجزیه و تحلیل تصویر را با تجزیه و تحلیل کد VBA برای تشخیص اسناد ترکیب می‌کند. در حالی که بسیار موثر است، این رویکرد محدود به نمونه های حاوی تصاویر است که ممکن است در همه موارد وجود نداشته باشد. مطالعه انجام شده توسط Nissim و همکاران [۱۵] با تمرکز بر مسیرهای فایل به عنوان ویژگی های فایل، و معرفی ویژگی‌هایی که نشان دهنده مسیر از ریشه تا هر برگ در ساختار فایل هستند، از مطالعات قبلی فاصله گرفت. با استفاده از مدل SVM، این روش با موفقیت فایل های docx مخرب را با قدرت تشخیص بالا شناسایی می‌کند. با این حال، حدود ۹۸٪ از فایل ها سالم بودند، و تنها حدود ۲٪ مخرب بودند، که می‌تواند بر یادگیری تاثیر بگذارد.



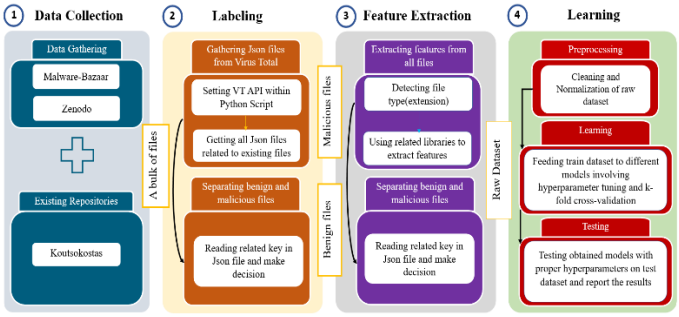
شکل ٤: ساختار اسناد آفیس (docx).

۳. روش پیشنهادی

در این تحقیق، هدف ما بهبود تشخیص فایل‌های مخرب غیرقابل اجرا و پیچیده با استفاده از طبقه‌بندی‌کننده‌ها برای دسته‌بندی فایل‌ها به عنوان مخرب یا سالم است. در زمینه طبقه بندی فایل های مخرب از انواع پیچیده غیر قابل اجرا، توسعه مدل های یادگیری یک فرآیند حیاتی است. این فرآیند در شکل ۵ و شکل ۶ نشان داده شده است که روش شناسی فایل های پی دی اف و آفیس را مشخص می‌کند. این رویکرد شامل چهار مرحله مجزا است: جمع‌آوری داده‌ها، برچسب‌گذاری قابل اعتماد، استخراج ویژگی، و پیش پردازش و به دنبال آن ایجاد مدل‌های طبقه‌بندی‌کننده. نوآوری‌های این مطالعه شامل ایجاد مجموعه داده جدید و تقویت‌شده، ارائه ویژگی‌های جدید جهت تشخیص بدافزارها و همچنین تنظیم هدفمند هایپرپارامترها جهت رسیدن به نتایج قابل اعتماد است که به ترتیب در بخش‌های جمع‌آوری داده، استخراج ویژگی و یادگیری در فرآیندهای نشان داده شده توضیح داده خواهند شد.



شکل ٥: فرآیند تشخیص فایل‌های مخرب پی دی اف



شکل ٦: فرآیند تشخیص فایل‌های مخرب آفیس

۳-۱. جمع‌آوری داده

این بخش مجموعه قابل توجهی از فایل های مرتبط را برای کمک به یادگیری بهتر و دقیق تر در مراحل بعدی جمع آوری میکند. در ابتدا، ما به منبعی به نام MalwareBazaar [۲۳]، دسترسی داشتیم که فایل‌های بدافزار را از سال ۲۰۲۰ تا کنون در بسته‌های روزانه جمع آوری می‌کند. به طور خاص، ما بسته‌های روزانه را برای سال‌های ۲۰۲۱، ۲۰۲۲ و ۲۰۲۳ به‌دست آوردیم. با توجه به تمرکز پروژه بر فایل‌های آفیس و پی دی اف، ما از یک اسکریپت پایتون برای شناسایی و جداسازی فایل‌های مربوطه از بقیه انواع موجود در بسته های دریافت شده استفاده کردیم.

برای تقویت مجموعه داده برای این پروژه، نمونه‌هایی از فایل های مرتبط از تحقیقات قبلی را ترکیب کردیم. برای فایل های پی دی اف، از مجموعه داده‌های CIC-Evasive-PDFMal2022 که شامل فایل‌های مخرب و سالم است و مجموعه داده‌های IUST Deep Fuzz [۱۶] که تنها شامل فایل‌های سالم است، همانطور که در شکل ۵ نشان داده شده است، استفاده کردیم. برای تقویت

مجموعه داده اسناد آفیس، ما یک crawler برای جمع آوری داده های مناسب، از جمله فرمت های doc، docx، xls، xlsx، ppt و pptx ایجاد کردیم. علاوه بر این، ما مجموعه داده‌هایی را که در کار مرتبط توضیح داده شده است [۲] را به مجموعه جمع آوری شده اضافه کردیم. در نهایت یک مجموعه داده شامل ۵۲,۳۰٤ فایل پی دی اف و ۳۸,۵۷٦ سند آفیس با تنوع بالا، با افزودن قابل توجهی از ۲,۲۳۷ سند آفیس حاوی اشیاء RTF که در مجموعه داده مرتبط وجود ندارد، به دست آمد. در بخش چهارم به تفصیل آمار مربوط به فایل های جمع آوری شده توضیح داده خواهد شد.

۲-۳. برچسب‌گذاری

در مرحله دوم، وظیفه حیاتی تعیین مخرب بودن فایل‌هایی که در ابتدا جمع‌آوری شده‌اند، بسیار مهم است. برای این منظور، یک فایل JSON، مربوط به هر فایل جمع‌آوری شده، از ویروس توتال [۱۷] گرفته می شود. برای اختصاص دادن یک برچسب ۰ یا ۱ به هر نمونه برای آموزش مدل های طبقه بندی دودویی، فایل JSON به دست آمده از ویروس توتال تجزیه و تحلیل شده و سپس برچسب مناسب به فایل متناظر تعلق میگیرد.

۳-۳. استخراج ویژگی

در مرحله سوم تجزیه و تحلیل، مجموعه‌ای از ویژگی‌ها را با دقت استخراج کردیم. با توجه به تفاوت‌های ذاتی فایل‌های پی دی اف و اسناد آفیس در ساختار و محتوا، همان‌طور که در شکل ۵ و شکل ۶ نشان داده شده است، استفاده از روش های متفاوت جهت استخراج ویژگی های مربوطه ضروری است.

در این مطالعه ۳٤ ویژگی برای فایل های آفیس و ٦٦ ویژگی برای فایل های پی دی اف استخراج شده است. دسته‌بندی کلی ویژگی‌های استخراج شده برای فایل پی دی اف شامل ویژگی‌های محتوایی، ساختاری، متاداده‌ای و مبتنی بر اشیاء می‌باشد. این دسته‌بندی ها برای فایل های آفیس شامل ویژگی های محتوایی، ساختاری، متاداده‌ای و مبتنی بر کد VBA می‌باشد.

۴-۳. یادگیری

این مطالعه بر بهینه‌سازی مدل‌های طبقه‌بندی دودویی از طریق پیش پردازش داده‌ها، تنظیم فراپارامتر و ارزیابی مدل تأکید دارد. این فرایند با تقسیم مجموعه داده‌ها به مجموعه‌های آموزشی و آزمایشی (۷۵% و ۲۵%) شروع می‌شود و مراحل پیش‌پردازش مانند حذف موارد تکراری و عادی‌سازی ویژگی‌های عددی را اعمال می‌کند. بهینه‌سازی فراپارامترها از رویکرد kfold cross validation با مقدار k=5 و GridSearchCV استفاده می‌کند و پارامترهای بهینه را برای مدل‌های مختلف شناسایی می‌کند. این مطالعه از مدل های یادگیری دودویی مانند تقویت گرادیان [۱۸]، پرسپترون چندلایه [۱۹]، جنگل تصادفی [۲۰]، و روش های یادگیری گروهی مانند Voting Classifiers- model3و Voting Classifier-model5 استفاده می‌کند. هدف این مطالعه به کارگیری همه موارد لازم جهت تعیین دقیق‌ترین مدل‌ها برای وظایف طبقه‌بندی دودویی فایل‌های مورد نظر است.

۴. مجموعه داده

همان طور که اشاره شد هدف این پروژه ایجاد مدل‌هایی برای تمایز بین فایل‌های مخرب و سالم، با تمرکز بر استخراج ویژگی‌های منحصر به فرد از نسخه‌های فایل‌های فیزیکی فایل‌های مد نظر است.

جدول ۱: آمار مربوط به فایل‌های پی دی اف جمع‌آوری شده

Maliciousness	Resource			
	CIC-Evasive-PDFMal2022 [22]	Malware bazaar [23]	IUST Deep Fuzz [16]	Overall
Benign	9107	38	6109	15254
Malicious	21898	1424	0	23322
Total	31005	1462	6109	38576

فایل‌ها از منابع مختلف، از جمله Zenodo، پایگاه داده بدافزار Malwarebazaar و فایل‌های سالم که توسط پروژه Zenodo و ذاکری نصرآبادی و همکاران تهیه شده بودند، جمع‌آوری شد. مجموعه نهایی با ترکیب این منابع و ارائه یک مجموعه داده جامع برای توسعه و ارزیابی مدل های دودویی اشاره شده شکل گرفت. آمار مربوط به فایل‌های جمع‌آوری شده برای فایل‌های پی دی اف و آفیس در جدول ۱ و جدول ۲ نشان داده شده است.

جدول ۲: آمار مربوط به اسناد آفیس جمع‌آوری شده

Maliciousness	Resource			
	Malware Bazaar [23]	Koutsokostas [2]	Zenodo [21]	Overall
Benign	87	2968	3998	7053
Malicious	30529	14965	0	45504
Total	30616	17933	3998	52557

۵. ارزیابی

پس از توضیح روش پیشنهادی، اجزای تشکیل دهنده معماری و ویژگی‌های مجموعه داده، این بخش به فاز آزمایشی می پردازد. این

بخش آزمون‌های انجام‌شده، نحوه انجام آن‌ها، نتایج به‌دست‌آمده و معیارهای ارزیابی مورد استفاده را تشریح می‌کند.

۵-۱. سؤالات پژوهشی

به منظور درک بهتر زمینه مطالعه، تدوین سؤالات پژوهشی مهم امری ضروری است. بر این اساس، ما سه سؤال اصلی تحقیق را به عنوان دستورالعملی برای آزمایشات خود به شرح زیر طراحی کرده‌ایم:

- پرسش ۱: کدام مدل طبقه‌بندی‌کننده، بدافزار پیچیده غیر قابل اجرا را به‌طور مؤثرتر شناسایی می‌کند؟
- پرسش ۲: مهم ترین ویژگی‌ها برای تشخیص فایل های مخرب پیچیده غیر قابل اجرا چیست؟
- پرسش ۳: اثربخشی رویکرد ما در مدل‌های طبقه‌بندی در مقایسه با سایر کارهای مرتبط چیست؟

۵-۲. معیارهای ارزیابی

ارزیابی عملکرد مدل‌های طبقه‌بندی دودویی شامل معیارهایی مانند accuracy, precision, recall و F score است که بینشی در مورد توانایی مدل برای طبقه‌بندی صحیح نمونه‌ها و حساسیت آن به موارد مثبت و منفی کاذب ارائه می‌کند [۲۴]. بر اساس این تعاریف، معیارهای بررسی شده برای مدل های دودویی به شرح زیر است:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (۱)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (۲)$$

$$\text{F-score} = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (۳)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (۴)$$

۵-۳. نتایج

در این بخش، با به ارائه شواهد و نتایجی که از آزمایش‌های خود پیدا کرده‌ایم، به سؤالات تحقیق که در بخش قبل بیان شد پاسخ می‌دهیم. در پایان هر بخش، پرسش و پاسخ تحقیق مرتبط را بر اساس نتایج نشان داده شده مطرح می‌کنیم.

۵-۳-۱. عملکرد مدل های تشخیص بدافزار

در مورد پرسش اول، اثربخشی هر مدل آموخته شده با معیارهای استاندارد مورد استفاده برای ارزیابی عملکرد مدل‌های طبقه‌بندی، از جمله دقت، صحت، یادآوری و در نهایت، امتیاز F اندازه‌گیری شد. جدول ۳ معیارهای ارزیابی را برای همه مدل‌های آموخته شده با استفاده از مجموعه داده‌های ارائه شده از فایل‌های پی دی اف نشان می‌دهد. بهترین مقدار به دست آمده برای هر معیار ارزیابی در این جدول پررنگ شده است. همچنین جدول ۴ نتایج آزمایشات برای مجموعه داده های آفیس را نشان می دهد.

جدول ۳: نتایج مدل‌های دودویی تشخیص بدافزارهای پی دی اف

Model name	Evaluation criteria			
	Accuracy	F-score	Precision	Recall
Random Forest	0.991	0.989	0.992	0.987
Gradient Boosting	**0.993**	**0.992**	**0.994**	**0.990**
Decision Tree	0.960	0.952	0.993	0.913
MLP	0.989	0.987	0.991	0.984
KNN	0.980	0.976	0.991	0.963
AdaBoost	0.986	0.983	0.985	0.981
Logistic Regression	0.964	0.958	0.979	0.938
SVM	0.977	0.973	0.987	0.960
Voting Classifier-model3	**0.993**	0.991	**0.994**	0.989
Voting Classifier-model5	0.991	0.989	0.993	0.985

جدول ٤: نتایج مدل‌های دودویی تشخیص بدافزارهای آفیس

Model name	Evaluation criteria			
	Accuracy	F-score	Precision	Recall
Random Forest	**0.994**	**0.996**	**0.997**	**0.995**
Gradient Boosting	0.992	0.995	**0.997**	0.994
Decision Tree	0.982	0.989	0.993	0.986
MLP	0.989	0.994	0.992	**0.995**
KNN	0.986	0.992	0.992	0.991
AdaBoost	0.987	0.992	0.991	0.993
Logistic Regression	0.969	0.982	0.990	0.973
SVM	0.984	0.991	0.989	0.993
Voting Classifier-model3	**0.994**	**0.996**	**0.997**	**0.995**
Voting Classifier-model5	0.990	0.994	0.993	**0.995**

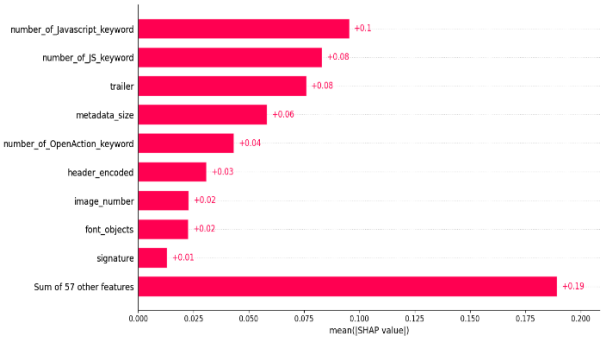
پرسش۱: کدام مدل طبقه‌بندی‌کننده، بدافزار پیچیده غیر قابل اجرا را به‌طور مؤثرتر شناسایی می‌کند؟

پاسخ به پرسش اول: مدل‌ها با استفاده از بهترین تنظیمات برای هر مدل، ارزیابی می‌شوند. مدل گرادیان تقویتی بهترین عملکرد را در مجموعه داده پی دی اف در چندین معیار از جمله دقت، امتیاز F، صحت و فراخوانی دارد. برعکس، جنگل تصادفی و طبقه‌بند رأی۳ بهترین عملکرد را در مجموعه داده آفیس در تمام معیارهای تعریف شده دارند.

۵-۳-۲. مهمترین ویژگی‌ها

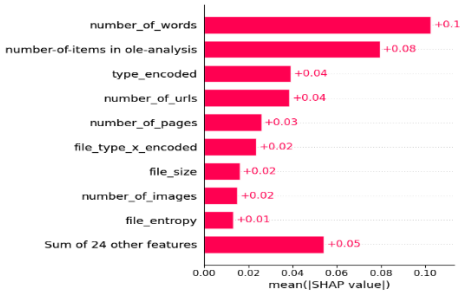
با توجه به اینکه مدل‌های یادگیرنده اغلب به‌عنوان «جعبه‌های سیاه» عمل می‌کنند و درک عملکرد درونی آن‌ها با چالش مواجه است، ساز و کاری لازم است تا عملکرد درونی آن‌ها مشخص شود. روش‌های

یادگیری توصیفی، مانند روش SHAP، در نمایش تأثیر ویژگی‌های مختلف بر خروجی مدل مؤثر هستند [۲۵]. SHAP روشی است که ریشه در تئوری بازی های مشارکتی دارد، به ویژه در افزایش شفافیت و تفسیرپذیری مدل های یادگیری ماشین عملکرد خوبی دارد.



شکل ۷: مهم‌ترین ویژگی‌ها در فایل‌های پی دی اف

با توجه به توضیحات ارائه شده، لازم است در این تحقیق به ویژگی‌هایی اشاره شود که بیشترین تاثیر را در یادگیری بهترین مدل داشته اند. همان‌طور که اشاره شد گرادیان تقویتی برای مجموعه داده پی دی اف و همچنین جنگل تصادفی برای مجموعه داده آفیس بهترین عملکرد را داشته‌اند. به همین دلیل نمودار SHAP برای این دو مدل به ترتیب در شکل ۷ و شکل ۸ ارائه شده است.



شکل ۸: مهم‌ترین ویژگی‌ها در فایل‌های آفیس

۳-۳-۵. مقایسه عملکرد روش‌های مختلف تشخیص بدافزار

مقایسه روش پیشنهادی ما با کارهای مختلف مرتبط، چالش بزرگی را به دلیل عدم وجود اندازه‌گیری‌های کامل و دقیق و مشکل در دسترسی به مجموعه داده‌های دقیق مورد استفاده برای آموزش و آزمایش مدل‌ها، در بر دارد. با این وجود، ما برخی از مقالاتی را که نتایج قابل اندازه‌گیری گزارش کرده‌اند، خلاصه کرده‌ایم و یافته‌های آن‌ها را در جدول ۵ مقایسه کرده‌ایم. رویکرد ما دقت آزمون برتر را در مقایسه با آزمایش‌های دیگر نشان داده است.

پرسش۲: مهم ترین ویژگی‌ها برای تشخیص فایل‌های مخرب پیچیده غیر قابل اجرا چیست؟

پاسخ به پرسش دوم: همانطور که در شکل ۷ نشان داده شده است، موثرترین و مهم‌ترین ویژگی‌ها در تشخیص بدافزار پی دی اف به ترتیب عبارتند از: number_of_javascript، number_of_js_keyword، trailer و metadata_size که نشان دهنده اهمیت کدهای جاوا اسکریپت برای تجزیه و تحلیل در فایل های پی دی اف است. از سوی دیگر، بر اساس آنچه در شکل ۸ نشان داده شده است، ضروری ترین ویژگی‌ها در تشخیص بدافزار آفیس به ترتیب تعداد_واژه‌ها، تعداد_آیتم ها در آنالیز OLE، نوع فایل و تعداد urls است که اهمیت تحلیل لینک های خطرناک و تعداد کلمات مشکوک در کد را نشان می دهد.

اگرچه تحقیق انجام شده توسط عیسی خانی و همکاران [۱] نتایج بهتری را نشان داد، ما آزمایشی را برای اثبات برتری تحقیق خود با توجه به تفاوت های مجموعه داده انجام دادیم. جدول۶ نشان می‌دهد که مدل ما نتایج بهتری را در هر دو مجموعه داده به دست می‌آورد و استحکام و مقاومت در برابر تغییرات اندازه مجموعه دارد. برعکس، مدل ارائه شده در مطالعه مذکور [۱] نتایج کاهش یافته را بر روی یک مجموعه داده بزرگتر با بهترین فراآموز[۱] خود که رگرسیون خطی است نشان داد، که حساسیت آن را به اندازه مجموعه داده برجسته می کند.

جدول ۵: مقایسه کارهای مرتبط با تشخیص بدافزارهای پی دی اف

Reference	Dataset	Features	Outcomes
[8]	Collection of PDF files obtained from Virus Total for a total of 65942	A set of JavaScript-related features	Best Accuracy 85%
[7]	Collection of 22196 malware and 40441 benign files	A set of JavaScript-related features	Best Accuracy 96.93%
[1]	A refined version of the Contagio dataset [26] integrated with Virus Total is	File structure, File content, and JavaScript	The Best F-score with LR is 99.86%, recall is

[۱] Meta Learner

Reference	Dataset	Features	Outcomes
	called the Evasive-PDFMal2022 dataset, which has 5557 malicious and 4468 benign entries.	for a total of 28 features	99.88%, precision is 99.84%, and accuracy is 99.89%.

جدول ٦: مقایسه بین استحکام و پایداری نتایج دو مطالعه انجام شده

Method	Dataset	
	Evasive-PDFMal2022	Our dataset
The best model of Issakhani et al. study [1](Linear Regression)	Accuracy of 99.89%, F-score of 99.86%, Precision of 99.84%, and Recall of 99.88%	Accuracy of 97.26%, F-score of 94.11%, precision of 95.68%, and recall of 92.58%
Best model of our study (Gradient Boosting)	Accuracy of 99.95%, F-score of 99.95%, precision of 100%, and recall of 99.91%	Accuracy of 99.3%, F-score of 99.2%, Precision of 99.4%, and Recall of 99%

جدول ۷: مقایسه کارهای مرتبط با تشخیص بدافزارهای آفیس

Reference	Dataset	Features	Outcomes
[12]	A random collection of office files, collected according to VirusTotal analysis, included 2,537 files and 4,212 macros (sometimes more than one per file), of which 877 were obfuscated.	A set of 15 lexicographical and function call features	The different machine-learning approaches report accuracies of around 90% in the task of identifying obfuscated macros. MLP was the most prominent with a 92% *F*2 score.
[11]	7,145 samples, including macros retrieved from VirusTotal.	Different language processing-related features, including SCDV, LSI, Doc2vec, Bag-of-words	The best *F*1 score reported is 93%.
[2]	Benign samples with macros collected from official sites and malicious samples collected from VirusTotal AppAny, Virusign, and Malshare, for a total of 2736 benign and 15,571 malicious	Lexicographical, VBA statistics, and function call analysis (LOLBAS, PowerShell, PSDecode)	The best F1 score was above 98%, and best accuracy score was 97.5% with RF, and recall was above 97.5% with RF. In the F2 score, 98% with RF was obtained.

علاوه بر این، تحقیقات ما به دلیل شفافیت آن، از جمله در دسترس بودن منبع مجموعه داده و مشخص بودن فراپارامترهای انتخاب شده برای مدل‌های یادگیری، در کنار جزئیات پیاده‌سازی، متمایز است. این عناصر بر وضوح آزمایش‌های ما تأکید می‌کنند. همانطورکه در جدول ۷ نشان داده شده است، این مطالعه مزایای متمایز را نسبت به کارهای مرتبط در شناسایی بدافزار آفیس ارائه می کند. تحقیق انجام شده در این مطالعه نه تنها مجموعه داده‌های جامع‌تر و متنوع‌تری را ارائه می‌کند که هر دو فایل ppt و pptx را در بر می‌گیرد، بلکه به نتایج برتر در هر چهار معیار معرفی‌شده یعنی دقت، صحت، امتیاز F و یادآوری، دست می‌یابد. علاوه بر این، این مطالعه تلاش می‌کند تا پارامترهای بهینه هر مدل یادگیرنده را شناسایی کند و تنظیم دقیق را پیاده‌سازی کند، ویژگی که معمولاً در مطالعات دیگر مشاهده نمی‌شود. علاوه بر این، در دسترس بودن پیاده‌سازی‌های اجرا شده و مجموعه داده‌های جمع‌آوری‌شده، این تحقیق را از همتایان خود متمایز می‌کند.

پرسش ۳: اثربخشی رویکرد ما در مدل‌های طبقه‌بندی در مقایسه با سایر کارهای مرتبط چیست؟

پاسخ به پرسش سوم: همانطورکه در چندین جدول نشان داده شده است، مدل های یادگیری معرفی شده در این مطالعه به طور موثر بدافزار پیچیده غیر قابل اجرا را شناسایی کردند. برای تشخیص بدافزار پی دی اف، مدل تقویت گرادیان ما به دقت ۹۹/۳ درصد دست یافت که از عملکرد مدل جنگل تصادفی با فراآموز رگرسیون خطی مورد استفاده در مطالعات مرتبط پیشی گرفت. به طور مشابه، در شناسایی بدافزارهای آفیس، مدل ما به دقت ۹۹/۴ درصد دست یافت که با مدل جنگل تصادفی در این زمینه بهتر از سایرین عمل کرد.

۶. نتیجه‌گیری

در این مطالعه به اهمیت تشخیص بدافزارهای پیچیده غیراجرایی اشاره شد. با توجه به ساختار متمایز فایل‌های پی دی اف و آفیس، دو ساز و کار کلی برای تشخیص آن‌ها ارائه شد. این مطالعه توانست با بهره‌گیری از مجموعه‌داده تقویت شده و بهره‌گیری از ویژگی‌های مناسب برای هر نوع فایل، مدل‌های دودویی ارائه دهد که عملکرد آن‌ها نسبت به مطالعات پیشین بهبود یافته است. به‌طور جزئی‌تر، این مطالعه توانست با ارائه مدل تقویت گرادیان برای پی دی اف و جنگل تصادفی برای آفیس به ترتیب به دقت ۹۹/۳ و ۹۹/۴ درصد دست پیدا کند و نسبت به مطالعات مشابه عملکرد بهتری را به ثبت برساند.

مراجع

[1] Issakhani M, Victor P, Tekeoglu A, and Lashkari AH. PDF Malware Detection based on Stacking Learning, in *International Conference on Information Systems Security and Privacy*, Science and Technology Publications, Lda, 2022, pp. 562–570. doi: 10.5220/0010908400003120.

[2] Koutsokostas V, Lykousas N, Apostolopoulos T, Orazi G, Ghosal A, Casino F, Conti M, Patsakis C. Invoice# 31415 attached: Automated analysis of malicious Microsoft Office documents. Computers & Security. 2022 Mar 1;114:102582.

[3] Kaspersky, "https://www.kaspersky.com/about/press-releases/2023_rising-threats-cybercriminals-unleash-411000-malicious-files-daily-in-2023."

[4] Avira, "https://www.avira.com/en/blog/malware-threat-report-q3-2020-statistics-and-trends."

[5] Singh P, Tapaswi S, Gupta S. Malware detection in pdf and office documents: A survey. Information Security Journal: A Global Perspective. 2020 May 3;29(3):134-53.

[6] Šrndic N, Laskov P. Detection of malicious pdf files based on hierarchical document structure. InProceedings of the 20th annual network & distributed system security symposium 2013 Feb (pp. 1-16). Citeseer.

[7] Gu J, Kong R, Sun H, Zhuang H, Pan F, Lin Z. A novel detection technique based on benign samples and one-class algorithm for malicious PDF documents containing JavaScript. In International Conference on Computer Application and Information Security (ICCAIS 2021) 2022 May 24 (Vol. 12260, pp. 599-607). SPIE.

[8] Laskov P, Šrndić N. Static detection of malicious JavaScript-bearing PDF documents. In Proceedings of the 27th annual computer security applications conference 2011 Dec 5 (pp. 373-382).

[9] Abu Al-Haija Q, Odeh A, Qattous H. PDF malware detection based on optimizable decision trees. Electronics. 2022 Sep 30;11(19):3142.

[10] Sohail B. Macro Based Malware Detection System. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 2021 Apr 11;12(3):5776-87.

[11] Liu JK, Huang X, editors. Network and System Security: 13th International Conference, NSS 2019, Sapporo, Japan, December 15–18, 2019, Proceedings. Springer Nature; 2019 Dec 10.

[12] Kim S, Hong S, Oh J, Lee H. Obfuscated VBA macro detection using machine learning. In2018 48th annual ieee/ifip international conference on dependable systems and networks (dsn) 2018 Jun 25 (pp. 490-501). IEEE.

[13] Ravi V, Gururaj SP, Vedamurthy HK, Nirmala MB. Analysing corpus of office documents for macro-based attacks using machine learning. Global Transitions Proceedings. 2022 Jun 1;3(1):20-4.

[14] Casino F, Totosis N, Apostolopoulos T, Lykousas N, Patsakis C. Analysis and correlation of visual evidence in campaigns of malicious office documents. Digital Threats: Research and Practice. 2023 Aug 10;4(2):1-9.

[15] Nissim N, Cohen A, Elovici Y. ALDOCX: detection of unknown malicious microsoft office documents using designated active learning methods based on new structural feature extraction methodology. IEEE Transactions on Information Forensics and Security. 2016 Dec 1;12(3):631-46.

[16] Zakeri-Nasrabadi M, Parsa S, Kalaee A. Format-aware learn&fuzz: deep test data generation for efficient fuzzing. Neural Computing and Applications. 2021 Mar;33(5):1497-513.

[17] Virus Total, https://www.virustotal.com/gui/home/upload., 2023.

[18] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 2013 Dec 4;7:21.

[19] Popescu MC, Balas VE, Perescu-Popescu L, Mastorakis N. Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 2009 Jul 1;8(7):579-88.

[20] Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.

[21] Zenodo, https://zenodo.org/., 2023.

[22] Issakhani, "CIC-Evasive-PDFMal2022 dataset." Accessed: Apr. 12, 2024. [Online]. Available: https://www.unb.ca/cic/datasets/pdfmal-2022.html

[23] Malwarebazaar, https://bazaar.abuse.ch/., 2023.

[24] Canbek G, Sagiroglu S, Temizel TT, Baykal N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In2017 International Conference on Computer Science and Engineering (UBMK) 2017 Oct 5 (pp. 821-826). IEEE.

[25] Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of computer-aided molecular design. 2020 Oct;34(10):1013-26.

[26] Contagio,"https://contagiodump.blogspot.com/2013/03/16800-clean-and-11960-malicious-files.html."