DOR: 20.1001.1.2538161.1404.10.3.2.2

DOI: 10.22034/csj.2025.233053

پاییز ۱۴۰۴، دوره ۱۰، شماره ۳، پیاپی ۳۸

دریافت مقاله: ۱۴۰۴/۰۴/۰۹

پذیرش مقاله: ۱۴۰۴/۰۷/۰۶

نوع مقاله: پژوهشی

شناسایی فایل‌های آلوده غیراجرایی به‌کمک مدل‌های یادگیری ماشین

رسول رضوانی جلال

کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

پست الکترونیکی: rasoul_rezvanijalal@alumni.iust.ac.ir

مرتضی ذاکری*

استادیار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران

پست الکترونیکی: zakeri@aut.ac.ir

چکیده

با افزایش استفاده کاربران از اسناد آفیس و پی‌دی‌اف و استفاده این نوع فایل‌ها در مراکز امنیتی جهت انتقال اطلاعات، توجه طراحان بدافزار به این فایل‌ها جلب شده است. فعالیت‌های گوناگونی جهت تشخیص این نوع فایل‌ها با انتخاب ویژگی‌های تعیین‌کننده هر نوع فایل صورت گرفته است. در این پژوهش سعی شده است تا با تهیه مجموعه داده تقویت شده 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 به کارگیری مدل‌های طبقه‌بند دودویی، در مقایسه با برترین مطالعات انجام شده، به بهبود ۲ درصدی تشخیص بدافزارهای پی‌دی‌اف 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 پی‌دی‌اف ، توانست به دقت ۹۹/۳ درصدی تشخیص فایل‌های آلوده دست پیدا کند در حالی که برای فایل‌های آفیس با اعمال مدل جنگل تصادفی، به نرخ ۹۹/٤ درصدی در تشخیص بدافزارها رسیده شد.

کلیدواژه‌ها: تشخیص بدافزار، فایل‌های پیچیده و غیراجرایی، یادگیری ماشین، فایل‌های پی‌دی‌اف، فایل‌های آفیس

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 تا در عملکرد رایانه قربانی اختلال ایجاد کند یا اطلاعات قربانی را به سرقت ببرد. از طرفی سرویس‌های آفیس از شرکت مایکروسافت پرکاربردترین مجموعه برای

* نویسندهٔ مسئول

پردازش اسناد، صفحات گسترده و ارائه‌ها است که به دلیل محبوبیت آن، به طور مداوم بـــرای انجام فعالیت‌های‌های مخـــرب مورد اســـتفاده قـــرار می‌گیـــرد. خـــراب‌کاران با ســـوء استفاده از ویژگی‌های این ســـرویس‌ها، از آن برای راه‌انـــدازی حملات خـــود و نفوذ به میلیون‌هـــا میزبان در فعالیت‌های خراب‌کارانه خود اســـتفاده می‌کنند. همان‌طور که توسط منابع مختلف مانند Avira و Eset [۳، ٤] گزارش شده اســـت اســـناد آفیس دومین قالب از میان چهار قالب پرکاربرد اســـت که توســـط بدافزارهـــا در محیط ویندوز استفاده می‌شود [۲]. همچنین در طول سالیان اخیر، فایل‌ها بـــا قالب پی‌دی‌اف به دلیل انعطاف‌پذیری و ویژگی‌های کار آســـان، به محبوب‌ترین قالب ارائه محتوا در بین کاربران تبدیل شـــده است که بر اســـاس آن‌چه در مورد فایل‌های آفیس توضیح داده شد، مورد ســـوء استفاده خراب‌کاران قرار می‌گیرند. لذا وجود یک ساز و کار جهت شناسایی این نوع بدافزارها بیش از پیش ضروری است.

به طورکلی، عامل اصـــلی در دقت عملکرد و در نتیجه موفقیت مدل‌هـــای یادگیری، مجموعه ویژگی‌هایی اســـت که برای جســـتجوی هدف مورد اســـتفاده قـــرار می‌گیرد. بنابرایـــن، اثربخـــشی شناســـایی فعالیت‌هـــای مخرب در فایل‌های غیرقابل اجرا، مانند اســـناد آفیس، به شناســـایی الگوهای شـــامل اســـکریپت‌ها، پیوندها، تصاویر و تعامل بین کدهای جاوا اســـکریپت و وی‌بی‌اســـکریپت و ســـایر ویژگی‌های مهمی بســـتگی دارد که در بخش‌های آینده به آنها اشاره خواهد شـــد. هدف این مقاله ارتقای قابلیت‌های تشـــخیص بدافـــزار پی‌دی‌اف و آفیس اســـت. ایـــن‌کار با جمع‌آوری مجموعه داده از انواع فایل‌های مختلف، از جمله ppt، xlsx، xls، docx، doc، pdf و pptx و همچنیـــن معرفی ویژگی‌های جدید برای پوشش طیف گسترده‌ای از حملات انجـــام می‌شـــود. در ادامه ایـــن فرآیند با ایجـــاد مدل‌های طبقه‌بنـــدی دودویی از ایـــن مجموعـــه داده و تنظیم دقیق فراپارامترها برای رسیدن به دقت تشخیص حداکثری ادامه می‌یابـــد. به‌طور خلاصه نوآوری‌های اصلی این مطالعه به شرح زیر است:

با مطالعه قالب فایل‌های پیچیده و مهندسی ویژگی‌های این نوع فایل‌هـــا، ویژگی‌های متمایز و متنوعی را برای هر نوع فایل انتخاب کردیم و تشـــخیص بدافزار مرتبط را در مقایسه با تلاش‌های قبلی افزایش دادیم.

با انجام آزمایش‌های خودکار بر روی طیف وسیعی از فراپارامترها در مدل‌های مختلف و استفاده از اعتبارسنجی k-fold، مؤثرتریـــن پیکربندی‌ها و فراپارامتر را مشـــخص کردیم و قابلیت اطمینان خروجی‌های مدل خود را تضمین می‌کنیم.

ما مجموعه‌ای بـــزرگ از فایل‌های مخرب و ســـالم را بـــرای هر دو قالـــب پی‌دی‌اف و آفیس گردآوری کردیم. با بهره‌گیری از منابع گوناگون، مجموعه داده‌های موجود را از لحاظ کمّی و کیفی ارتقـــاء دادیم. به این ترتیب آموزش بر روی یک مجموعه داده گســـترده انجام می‌شود و نتایج مدل‌های طبقه‌بندی دودویی قابل اعتمادتر است.

در بخـــش بعدی، مروری بر تحقیقات موجود در مورد بدافـــزار ارائـــه می‌کنیم که به صراحت بر اســـناد آفیس و فایل‌هـــای پی‌دی‌اف تمرکـــز دارد. ایـــن مطالعه به بررسی عناصـــر ســـاختاری آنهـــا و تکنیک‌های تشـــخیص مورد استفاده می‌پردازد. متعاقباً، ما به روش به کار گرفته شده در ایـــن مطالعه می‌پردازیم و جزئیـــات جمع‌آوری داده‌ها، اعتبارسنجی، اســـتخراج ویژگی، و فرآیندهای یادگیری را در بخش ۳ بیان می‌کنیم. بخش ٤ یک نمای کلی از مجموعه داده پیشنهادی ما، از جمله تجزیه‌وتحلیل آماری آن و بحثی دربـــاره آمار مجموعه داده ارائه می‌کند. بخش ٥ یافته‌ها و آزمایش‌هـــای ما را تحلیل می‌کنـــد. در نهایت، بخش ٦ یک نتیجه‌گیری و جمع‌بندی از مطالعه انجام شـــده را، به‌طور خلاصه بیان می‌کند.

۲- کارهای مرتبط

ایـــن بخش یک مـــرور کلی از کار مرتبطی که زیربنای مطالعه ما است ارائه می‌دهد. در ابتدا، ساختار اسناد آفیس

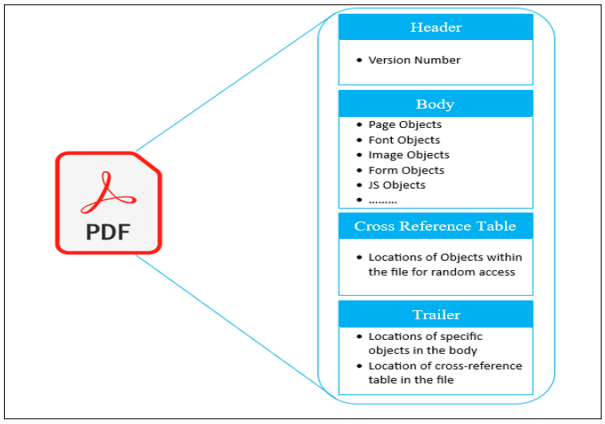
و فایل‌هـــای پی‌دی‌اف را بررسی می‌کنیم و یک پیش‌زمینه جامع ارائه می‌کنیم. متعاقباً، روش‌های دفاعی مورد استفاده در برابر بدافزارهای این دسته را تشریح می‌کنیم.

۱-۲- فایل‌های پی‌دی‌اف

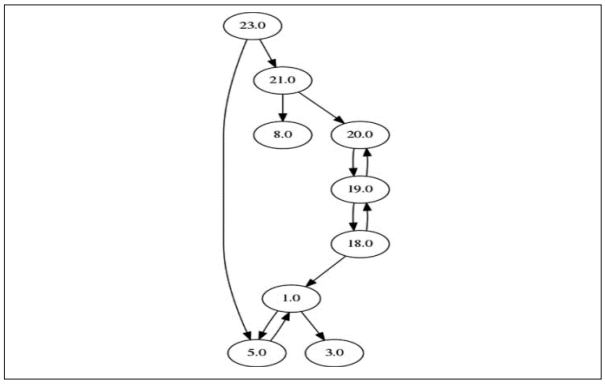
قبل از پرداختن به بررسی موارد مربوط به شناسایی بدافزارها، ضروری اســت که با فایل‌های پی‌دی‌اف آشــنا شویم. ســاختار کلی فایل‌های پی‌دی‌اف در شکل ۱ نشان داده شــده اســت. این مطالعه بر پیچیدگی این نوع فایل‌ها تاکید می‌کند. فایل‌های پی‌دی‌اف، ســاختار سلسله مراتبی دارنــد که می‌توان آن‌ها را بــه صورت یک گراف جهت‌دار نشــان داد. همان‌طور که در شکل ۲ نشان داده شده است، هر گره در این نمودار با یک مقدار عددی که شــماره شیء در ترتیب اشیاء موجود در فایل را نشان می‌دهد، شناسایی می‌شــود. این مفهوم در مطالعه‌ای [۵] توضیح داده شــده است، که فرض می‌کند هر گره در این نمودار نماد یک شیء منفرد در فایل پی‌دی‌اف اســت. بــرای بررسی دقیق‌تر این نمودار، شکل ۳ مربوط به مطالعه Šrndic و همکاران [۶] را می‌توان در نظر گرفت که تجســم مناسبی از ساختار فایل پی‌دی‌اف ارائه می‌دهد.

اســنادی ماننــد فایل‌هــای پی‌دی‌اف بــه دلیل ظرفیت آنها برای ترکیب انواع محتوا، بــه عنوان فایل‌های پیچیده طبقه‌بندی می‌شوند. مؤلفه قابل توجه محتوای ذکر شده کد جاوا اسکریپت اســت که قابلیت این نوع فایل‌ها را افزایش می‌دهــد. در نتیجه، مطالعات متعددی بر روی شناســایی فایل‌های پی‌دی‌اف مخرب با شناســایی کد جاوا اسکریپت متمرکــز شــده‌اند. مطالعــه‌ای توســط جیاکسیانگ گو و همکاران مجموعه‌ای از توکن‌های مشتق شده از متغیرهای کد جاوا اســکریپت را به عنوان ویژگی‌هــای فایل معرفی می‌کند و با اســتفاده از مــدل SVM به دقت ۹۶/۹۳ درصد دست می‌یابد[۷].

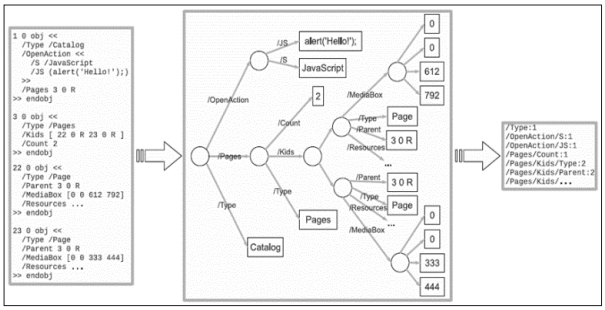
به طور مشــابه، یک مطالعه مرتبــط [۸] از الگوریتمی به نام PJscan برای ترجمه کد جاوا اســکریپت به توکن‌ها



شکل ۱: نمای کلی فایل پی‌دی‌اف



شکل ۲: فایل پی‌دی‌اف به شکل گراف جهت‌دار [۵]



شکل ۳: ساختار فیزیکی در مقابل ساختار سلسله مراتبی فایل پی‌دی‌اف

استفاده کرد که دقت ۸۵ درصد را به دست آورد. علاوه بر این، قدرت تشــخیص با ترکیب ویژگی‌های عمومی اضافی مرتبط با جاوا اسکریپت افزایش یافته است، و با استفاده از تکنیک یادگیری پشته‌ای با رگرسیون لجستیک فرایادگیرنده (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] بــر روی مجموعه داده CIC-Evasive-PDFMal2022، بــا اســتفاده از الگوریتــم درخت تصمیم

بهینه سازی (O_DT) انجام شده است که به دقت تشخیص ۹۸/۸٤ درصد دست یافته است. علاوه بر این، مجموعه‌های مختلفی از ویژگی‌ها برای فایل‌های پی‌دی‌اف پیشنهاد شده‌اند [٦]، 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

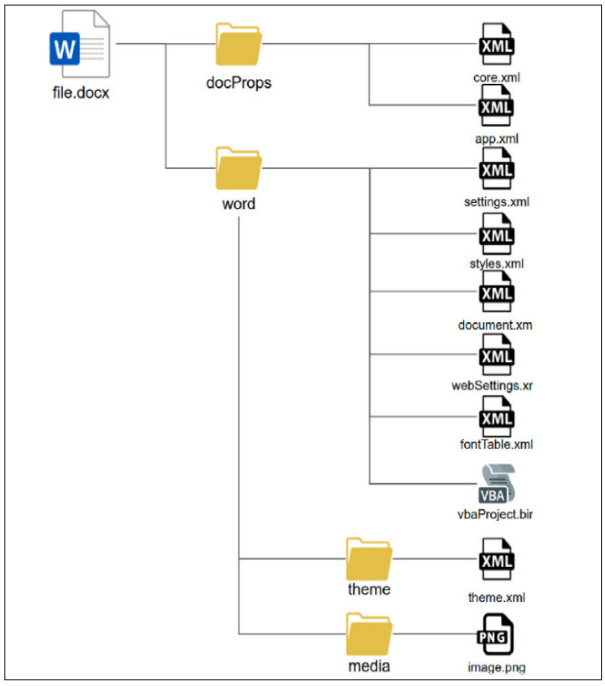
درک عمیق از ترکیب ساختاری اسناد اداری در تجزیه و تحلیل اسناد ضروری است. این ضرورت توسط تحقیقات علمی در این حوزه، مانند تحقیق‌های انجام شده توسط کوتسوکوستاس و همکاران [۲]، و سینگ و همکاران. [٥] که توضیحات مفصلی از سازماندهی سلسله مراتبی این انواع فایل ارائه کرده است، تأکید می‌شود. همان‌طور که در شکل ٤ نشان داده شده است، واضح است که اسناد آفیس ساختاری درخت‌مانند شبیه فایل‌های پی‌دی‌اف را نشان می‌دهند. این شباهت ساختاری بر پیچیدگی ذاتی در هر دو نوع سند تأکید می کند.

مطالعات و فعالیت‌های زیادی برای شناسایی و تحلیل اسناد اداری با تاکید بر بررسی کدهای ماکرو انجام شده است. به عنوان مثال، در مطالعه‌ای [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 93 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 فراخوانی‌های API و متغیرهای مبهم را استخراج کردند و به نرخ دقت ۸۲/٦٥ درصد دست یافتند. کوتسوکوستاس و همکاران [۲] مطالعه خود را در مورد تجزیه‌وتحلیل ماکروهای VBA گسترش دادند و از تبادل پویای داده DDE استفاده کردند، ویژگی‌ای که تبادل داده بین اسناد اداری مختلف را تسهیل می‌کند. این مطالعه به دقت تشخیص ۹۷/٥ درصد با مدل جنگل تصادفی RF دست یافت.

علاوه بر این، روشی توسعه یافته است [١٤] که تجزیه‌وتحلیل تصویر را با تجزیه‌وتحلیل کد VBA برای تشخیص اسناد ترکیب می‌کند. این رویکرد در حالی که بسیار موثر است، محدود به نمونه‌های حاوی تصاویر است که ممکن است در همه موارد وجود نداشته باشد. مطالعه انجام شده توسط Nissim و همکاران [۱٥] با تمرکز بر مسیرهای فایل به عنوان ویژگی‌های فایل، و معرفی ویژگی‌هایی که نشان‌دهنده مسیر از ریشه تا هر برگ در ساختار فایل هستند، از مطالعات قبلی فاصله گرفت. با استفاده از مدل SVM، این روش با موفقیت فایل‌های docx مخرب را با قدرت تشخیص بالا شناسایی می کند. با این حال، حدود ۹۸ درصد از فایل‌ها سالم بودند، و تنها حدود ۲ درصد مخرب بودند، که می‌تواند بر یادگیری تاثیر بگذارد

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 طبقه‌بندی فایل‌های مخرب از انواع پیچیده غیرقابل اجرا، توسعه مدل‌های یادگیری یک فرآیند حیاتی است. این فرآیند در شکل ٥ و ٦ نشان داده شده است که روش شناسی فایل‌های پی‌دی‌اف و آفیس را مشخص می‌کند. این رویکرد شامل چهار مرحله مجزا است: جمع‌آوری داده‌ها، برچسب‌گذاری قابل اعتماد،

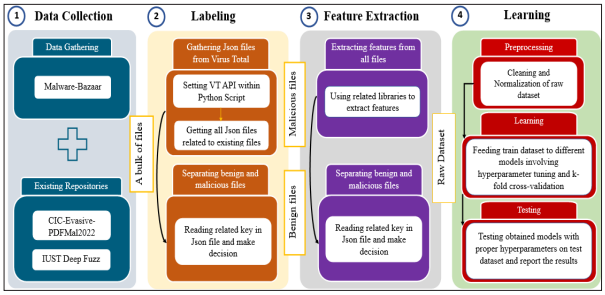


شکل ۴: ساختار اسناد آفیس (docx).

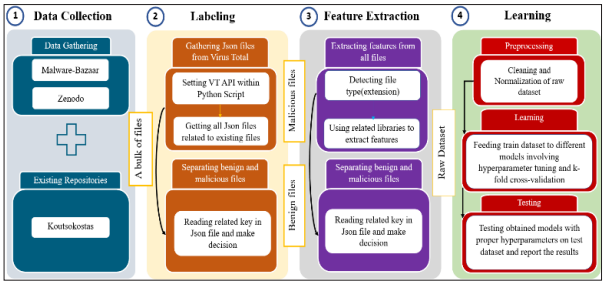
اســـتخراج ویژگی، و پیش‌پردازش و بـــه دنبال آن ایجاد مدل‌های طبقه‌بندی‌کننده. نوآوری‌های این مطالعه شامل ایجاد مجموعه داده جدید و تقویت‌شده، ارائه ویژگی‌های جدید جهت تشـــخیص بدافزارها و همچنین تنظیم هدفمند فراپارامترها جهت رسیدن به نتایج قابل اعتماد اســـت که به ترتیب در بخش‌های جمع‌آوری داده، استخراج ویژگی و یادگیری در فرآیندهای نشـــان داده شده توضیح داده خواهند شد.

۱-۳- جمع‌آوری داده

این بخـــش مجموعه قابل‌توجـــهی از فایل های مرتبط را بـــرای کمک بـــه یادگیری بهتـــر و دقیق‌تـــر در مراحل بعـــدی جمـــع‌آوری می‌کنـــد. در ابتدا، ما به منبـــعی به نام MalwareBazaar [۲۳]، دســـترسی داشـــتیم که فایل‌های بدافـــزار را از ســـال ۲۰۲۰ تا کنون در بســـته‌های روزانه جمع‌آوری می‌کند. به طور خاص، ما بســـته‌های روزانه را برای ســـال‌های ۲۰۲۱، ۲۰۲۲ و ۲۰۲۳ به‌دست آوردیم. با توجه به تمرکز پروژه بر فایل‌های آفیس و پی‌دی‌اف، ما از یک اسکریپت پایتون برای شناسایی و جداسازی فایل‌های



شکل ۵: فرآیند تشخیص فایل‌های مخرب پی‌دی‌اف



شکل۶: فرآیند تشخیص فایل‌های مخرب آفیس

مربوطه از بقیه انواع موجود در بســـته‌های دریافت شـــده استفاده کردیم.

برای تقویت مجموعه داده برای این پروژه، نمونه‌هایی از فایل‌های مرتبط از تحقیقات قبلی را ترکیب کردیم. برای فایل‌هـــای پی‌دی‌اف، از مجموعه داده‌هـــای CIC-Evasive-PDFMal2022 که شـــامل فایل‌های مخرب و سالم است و مجموعـــه داده‌های IUST Deep Fuzz [۱۶] که تنها شـــامل فایل‌های سالم اســـت، همان‌طور که در شکل ۵ نشان داده شده است، استفاده کردیم. برای تقویت مجموعه داده اسناد آفیس، ما یک crawler برای جمع‌آوری داده‌های مناســـب، از جمله فرمت‌های ppt، xlsx، xls، docx، doc و pptx ایجاد کردیم. علاوه بر این، ما مجموعـــه داده‌هایی را که در کار مرتبط توضیح داده شده است [۲] را به مجموعه جمع‌آوری شـــده اضافه کردیـــم. در نهایت یک مجموعه داده شـــامل ۵۲,۳۰۴ فایل پی‌دی‌اف و ۳۸,۵۷۶ 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 دســـت آمد. در بخش چهارم به تفصیل آمار مربوط به فایل‌های جمع‌آوری شده توضیح داده خواهد شد.

۳-۲- برچسب‌گذاری

در مرحلـــه دوم، وظیفـــه حیاتی تعییـــن مخرب بودن فایل‌هـــایی کـــه در ابتدا جمـــع‌آوری شـــده‌اند، بسیار مهم است. برای این منظور، یک فایل JSON، مربوط به هر فایل جمع‌آوری شـــده، از ویروس توتال [۱۷] گرفته می‌شـــود. برای اختصاص دادن یک برچسب ۰ یا ۱ به هر نمونه برای آموزش مدل‌های طبقه‌بندی دودویی، فایل JSON به‌دســـت آمده از ویروس توتال تجزیه‌وتحلیل شده و سپس برچسب مناسب به فایل متناظر تعلق می‌گیرد.

۳-۳- استخراج ویژگی

در مرحله سوم تجزیه‌وتحلیل، مجموعه‌ای از ویژگی‌ها را با دقت اســـتخراج کردیم. با توجه بـــه تفاوت‌های ذاتی فایل‌های پی‌دی‌اف و اســـناد آفیس در ســـاختار و محتوا، همان‌طور که در شـــکل‌های ۵ و ٦ نشـــان داده شده است، اســـتفاده از روش‌های متفاوت جهت استخراج ویژگی‌های مربوطه ضروری است.

در ایـــن مطالعه ۳٤ ویـــژگی برای فایل هـــای آفیس و ٦٦ ویژگی برای فایل‌های پی‌دی‌اف اســـتخراج شده است. دســـته‌بندی کلی ویژگی‌های اســـتخراج شـــده برای فایل پی‌دی‌اف شامل ویژگی‌های محتوایی، ساختاری، فراداده‌ای و مبتنی بر اشیاء می‌باشد. این دسته‌بندی‌ها برای فایل‌های آفیس شامل ویژگی‌های محتوایی، ساختاری، فراداده‌ای و مبتنی بر کد VBA می‌باشد.

۳-۴- یادگیری

این مطالعه بر بهینه‌سازی مدل‌های طبقه‌بندی دودویی از طریـــق پیش‌پردازش داده‌ها، تنظیم فراپارامتر و ارزیابی مدل تأکیـــد دارد. این فرایند با تقسیـــم مجموعه داده‌ها به مجموعه‌هـــای آمـــوزشی و آزمایـــشی (۷۵ درصد و ۲۵ درصد) شروع می‌شود و مراحل پیش‌پردازش مانند حذف موارد تکراری و عادی‌ســـازی ویژگی‌های عددی را اعمال می‌کند. بهینه‌ســـازی فراپارامترها از رویکرد kfold cross validation با مقدار k=5 و GridSearchCV استفاده می‌کند و پارامترهـــای بهینه را برای مدل‌های مختلف شناســـایی می‌کند. این مطالعـــه از مدل‌های یادگیـــری دودویی مانند تقویـــت گرادیان [۱۸]، پرســـپترون چندلایـــه [۱۹]، جنگل تصـــادفی [۲۰]، و روش‌هـــای یادگیـــری گـــروهی ماننـــد Voting Classifiers-model3 و Voting Classifier-model5 اســـتفاده می‌کند. هدف این مطالعه به کارگیری همه موارد لازم جهت تعیین دقیق‌ترین مدل‌ها برای وظایف طبقه‌بندی دودویی فایل‌های موردنظر است.

۳-۵- مجموعه داده

همان‌طور که اشـــاره شـــد هـــدف این پـــروژه ایجاد مدل‌هایی برای تمایز بین فایل‌های مخرب و سالم، با تمرکز بر اســـتخراج ویژگی‌هـــای منحصر به فرد از نســـخه‌های فایل‌های فیزیکی فایل‌های مدنظر است.

فایل‌هـــا از منابع مختلف، از جمله Zenodo، پایگاه داده بدافـــزار Malwarebazaar و فایل‌های ســـالم که توســـط پروژه Zenodo و ذاکری نصرآبادی و همکاران تهیه شده بودند، جمع‌آوری شـــد. مجموعه نهایی با ترکیب این منابع و ارائـــه یک مجموعه داده جامع برای توســـعه و ارزیابی مدل‌های دودویی اشاره شده شکل گرفت. آمار مربوط به فایل‌های جمع‌آوری شده برای فایل‌های پی‌دی‌اف و آفیس در جداول ۱ و ۲ نشان داده شده است.

۴- ارزیابی

پس از توضیح روش پیشـــنهادی، اجزای تشکیل‌دهنده معمـــاری و ویژگی‌های مجموعه داده، این بخش به مرحلهٔ آزمایشی می‌پـــردازد. این بخش آزمون‌های انجام شـــده، نحوه انجام آنها، نتایج به‌دســـت آمده و معیارهای ارزیابی مورد استفاده را تشریح می‌کند.

۴-۱- سؤالات پژوهشی

به منظـــور درک بهتر زمینه مطالعه، تدوین ســـؤالات

جدول ۱: آمار مربوط به فایل‌های پی‌دی‌اف جمع‌آوری شده

Maliciousness	Resource			
	CIC-Evasive-PDFMal2022 [22]	Malware bazaar [23]	IUST Deep Fuzz [16]	Overall
Benign	۹۱۰۷	۳۸	۶۱۰۹	۱۵۲۵۴
Malicious	۲۱۸۹۸	۱۴۲۴	۰	۲۳۳۲۲
Total	۳۱۰۰۵	۱۴۶۲	۶۱۰۹	۳۸۵۷۶

جدول ۲: آمار مربوط به اسناد آفیس جمع‌آوری شده

Maliciousness	Resource			
	Malware Bazaar [23]	Koutsokostas [2]	Zenodo [21]	Overall
Benign	۸۷	۲۹۶۸	۳۹۹۸	۷۰۵۳
Malicious	۳۰۵۲۹	۱۴۹۶۵	۰	۴۵۵۰۴
Total	۳۰۱۶	۱۷۹۳۳	۳۹۹۸	۵۲۵۵۷

پژوهشی مهم امری ضروری است. بر این اساس، ما سه سؤال اصلی تحقیق را به عنوان دستورالعملی برای آزمایش‌های خود به شرح زیر طراحی کرده‌ایم:

پرسش ۱: کدام مدل طبقه‌بندی‌کننده، بدافزار پیچیده غیرقابل اجرا را به‌طور مؤثرتر شناسایی می‌کند؟

پرسش ۲: مهمترین ویژگی‌ها برای تشخیص فایل‌های مخرب پیچیده غیرقابل اجرا چیست؟

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

پرسش ٤: تاثیر مجموعه داده نامتوازن بر خروجی مدل طبقه‌بندی‌کننده چیست؟

۴-۲- 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 [۲٤]. بر اساس این تعاریف، معیارهای بررسی شده برای مدل‌های دودویی به شرح زیر است:

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

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

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

$$\text{Recall} = \frac{TP}{TP+FN} \quad (٤)$$

۵- نتایج

در این بخش، با ارائه شواهد، نتایج و خروجی‌هایی که از آزمایش‌های خود به‌دست آورده‌ایم، 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.

۵-۱- عملکرد مدل‌های 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 پی‌دی‌اف نشان می‌دهد. بهترین مقدار به دست آمده برای هر معیار ارزیابی در این

جدول پررنگ شده است. همچنین جدول ٤ نتایج آزمایش‌ها برای مجموعه داده‌های آفیس را نشان می دهد.

جدول ۳: نتایج مدل‌های دودویی تشخیص بدافزارهای پی‌دی‌اف

Model name	Evaluation criteria			
	Accuracy	F-score	Precision	Recall
Random Forest	۰/۹۹۱	۰/۹۸۹	۰/۹۹۲	۰/۹۸۷
Gradient Boosting	۰/۹۹۳	۰/۹۹۲	۰/۹۹۴	۰/۹۹۰
Decision Tree	۰/۹۶۰	۰/۹۵۲	۰/۹۹۳	۰/۹۱۳
MLP	۰/۹۸۹	۰/۹۸۷	۰/۹۹۱	۰/۹۸۴
KNN	۰/۹۸۰	۰/۹۷۶	۰/۹۹۱	۰/۹۶۳
AdaBoost	۰/۹۸۶	۰/۹۸۳	۰/۹۸۵	۰/۹۸۱
Logistic Regression	۰/۹۶۴	۰/۹۵۸	۰/۹۷۹	۰/۹۳۸
SVM	۰/۹۷۷	۰/۹۷۳	۰/۹۸۷	۰/۹۶۰
Voting Classifier-model3	۰/۹۹۳	۰/۹۹۱	۰/۹۹۴	۰/۹۸۹
Voting Classifier-model5	۰/۹۹۱	۰/۹۸۹	۰/۹۹۳	۰/۹۸۵

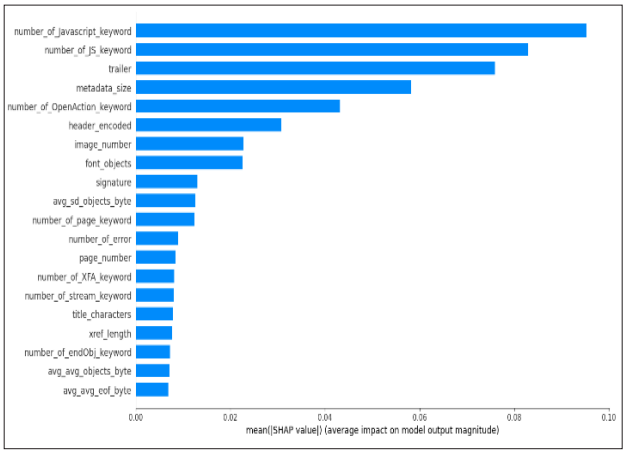
جدول ۴: نتایج مدل‌های دودویی تشخیص بدافزارهای آفیس

Model name	Evaluation criteria			
	Accuracy	F-score	Precision	Recall
Random Forest	۰/۹۹۴	۰/۹۹۶	۰/۹۹۷	۰/۹۹۵
Gradient Boosting	۰/۹۹۲	۰/۹۹۵	۰/۹۹۷	۰/۹۹۴
Decision Tree	۰/۹۸۲	۰/۹۸۹	۰/۹۹۳	۰/۹۸۶
MLP	۰/۹۸۹	۰/۹۹۴	۰/۹۹۲	۰/۹۹۵
KNN	۰/۹۸۶	۰/۹۹۲	۰/۹۹۲	۰/۹۹۱
AdaBoost	۰/۹۸۷	۰/۹۹۲	۰/۹۹۱	۰/۹۹۳
Logistic Regression	۰/۹۶۹	۰/۹۸۲	۰/۹۹۰	۰/۹۷۳
SVM	۰/۹۸۴	۰/۹۹۱	۰/۹۸۹	۰/۹۹۳
Voting Classifier-model3	۰/۹۹۴	۰/۹۹۶	۰/۹۹۷	۰/۹۹۵
Voting Classifier-model5	۰/۹۹۰	۰/۹۹۴	۰/۹۹۳	۰/۹۹۵

پرسش ۱: کدام مدل طبقه‌بندی‌کننده، بدافزار پیچیده غیرقابل 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 پی‌دی‌اف در چندین معیار از جمله دقت، امتیاز F، صحت و فراخوانی دارد. برعکس، جنگل تصادفی و طبقه‌بند رأی ۳ بهترین عملکرد را در مجموعه داده آفیسر تمام معیارهای تعریف شده دارند.

۵-۲- مهمترین ویژگی‌ها

با توجه به این که 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 نظریه بازی‌های مشارکتی دارد، به ویژه در افزایش شفافیت و تفسیرپذیری مدل‌های یادگیری ماشین عملکرد خوبی دارد.

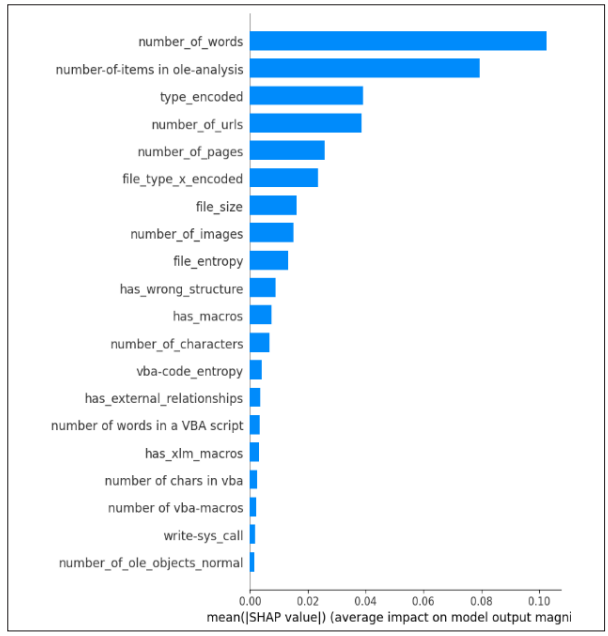


شکل ۷: مهمترین ویژگی‌ها در فایل‌های پی‌دی‌اف

با توجه به توضیحات ارائه شده، لازم است در این تحقیق به ویژگی‌هایی اشاره شود که بیشترین تاثیر را در یادگیری بهترین مدل داشته‌اند. همان‌طور که اشاره شد گرادیان تقویتی برای مجموعه داده پی‌دی‌اف 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 های ۷ و ۸ ارائه شده است.

۵-۳- مقایسه عملکرد روش‌های مختلف تشخیص بدافزار

مقایسه روش پیشنهادی ما با کارهای مختلف مرتبط، چالش بزرگی را به دلیل عدم وجود اندازه‌گیری‌های کامل و



شکل ۸: مهمترین ویژگی‌ها در فایل‌های آفیس

دقیق و مشکل در دسترسی به مجموعه داده‌های دقیق مورد استفاده برای آموزش و آزمایش مدل‌ها، در بر دارد. با این وجـــود، ما برخی از مقالاتی را که نتایـــج قابل اندازه‌گیری گـــزارش کرده‌اند، خلاصه کرده‌ایم و یافته‌های آن‌ها را در جدول ۵ مقایســـه کرده‌ایم. رویکرد ما دقت آزمون برتر را در مقایسه با آزمایش‌های دیگر نشان داده است.

اگر چه تحقیق انجام شده توسط عیسی‌خانی و همکاران [۱] نتایج بهتری را نشـــان داد، در ایـــن مطالعه آزمایشی بـــرای اثبات برتری نتایج این تحقیق با توجه به تفاوت‌های مجموعه داده انجام شد. جدول ٦ نشان می‌دهد که مدل ما نتایج بهتری را در هر دو مجموعه داده به دســـت می‌آورد و اســـتحکام و مقاومت در برابر تغییـــرات اندازه مجموعه دارد. برعکس، مدل ارائه شـــده در مطالعه مذکور [۱] نتایج کاهش یافته را بر روی یک مجموعه داده بزرگتر با بهترین فرا آموز[1] خود که رگرسیون خطی اســـت نشـــان داد، که حساسیت آن را به اندازه مجموعه داده برجسته می کند.

علاوه بـــر ایـــن، تحقیقات ما به دلیل شـــفافیت آن، از جمله در دســـترس بودن منبع مجموعه داده و مشـــخص بودن فراپارامترهای انتخاب شده برای مدل‌های یادگیری، در کنار جزئیات پیاده‌ســـازی، متمایز است. این عناصر بر وضـــوح آزمایش‌های ما تأکید می‌کننـــد. همان‌طور که در جدول ۷ نشان داده شده است، این مطالعه مزایای متمایز را نســـبت به کارهای مرتبط در شناسایی بدافزار آفیس ارائه می‌کند. تحقیق انجام شـــده در این مطالعه نه تنها مجموعه داده‌هـــای جامع‌تر و متنوع‌تری را ارائـــه می‌کند که هر دو فایـــل ppt و pptx را در بر می‌گیرد، بلکه به نتایج برتر در هر چهار معیار معرفی شـــده یعنی دقت، صحت، امتیاز F و یادآوری، دســـت می‌یابد. علاوه بر این، این مطالعه تلاش می‌کند تا پارامترهای بهینه هر مدل یادگیرنده را شناسایی کند و تنظیم دقیق را پیاده‌ســـازی کنـــد، ویژگی که معمولاً در مطالعـــات دیگر مشـــاهده نمی‌شـــود. علاوه بر این، در دســـترس بودن پیاده‌ســـازی‌های اجرا شـــده و مجموعه داده‌های جمع‌آوری شـــده، این تحقیـــق را از همتایان خود متمایز می‌کند.

پرســـش ۲: مهمترین ویژگی‌ها برای تشخیص فایل‌های مخرب پیچیده غیرقابل اجرا چیست؟
پاسخ به پرسش دوم: همانطور که در شکل ۷ نشان داده شده اســـت، موثرترین و مهمترین ویژگی‌ها در تشخیص بدافزار پی‌دی‌اف به ترتیب عبارتند از: number_of_javascript، num-ber_of_js_keyword، trailer و metadata_size که نشان‌دهنده اهمیت کدهای جاوا اسکریپت برای تجزیه‌وتحلیل در فایل‌های پی‌دی‌اف است. از سوی دیگر، بر اساس آنچه در شکل ۸ نشان داده شده است، ضروری‌ترین ویژگی‌ها در تشخیص بدافزار آفیس به ترتیب تعداد واژه‌ها، تعداد موارد در تحلیل OLE، نوع فایل و تعداد urls است که اهمیت تحلیل پیوندهای خطرناک و تعداد کلمات مشکوک در کد را نشان می دهد.

پرســـش ۳: اثربخشی رویکرد ما در مدل‌های طبقه‌بندی در مقایسه با سایر کارهای مرتبط چیست؟
پاســـخ به پرسش سوم: همان‌طور که در چندین جدول نشان داده شده است، مدل‌های یادگیری معرفی شده در این مطالعه به طور موثر بدافزار پیچیده غیرقابل اجرا را شناسایی کردند. برای تشـــخیص بدافزار پی‌دی‌اف، 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 مـــا به دقت ۹۹/۴ درصد دســـت یافت که بـــا مدل جنگل تصادفی در این زمینه بهتر از سایرین عمل کرد.

1-Meta Learner

جدول ۵: مقایسه کارهای مرتبط با تشخیص بدافزارهای پی‌دی‌اف

Reference	Dataset	Features	Outcomes
[۸]	Collection of PDF files obtained from Virus Total for a total of 65942	A set of JavaScript-related features	Best Accuracy 85%
[۷]	Collection of 22196 malware and 40441 benign files	A set of JavaScript-related features	Best Accuracy 96.93%
[۱]	A refined version of the Contagio dataset [26] integrated with Virus Total is called the Evasive-PDFMal2022 data-.set, which has 5557 malicious and 4468 benign entries	File structure, File content, and JavaScript for a total of 28 features	The Best F-score with LR is 99.86%, recall is 99.88%, precision is 99.84%, and accuracy is 99.89%.

جدول ۶: مقایسه بین استحکام و پایداری نتایج دو مطالعه انجام شده

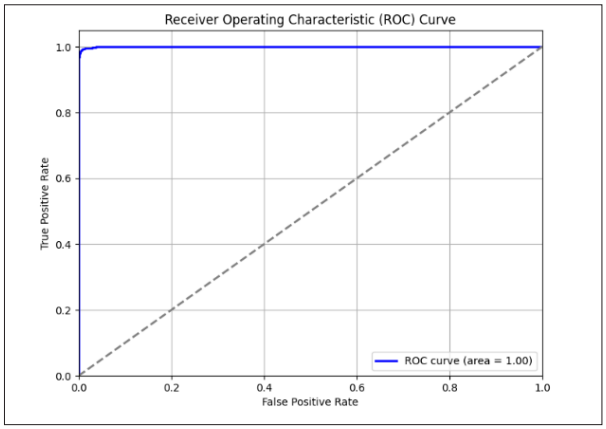
Method	Dataset	
	Evasive-PDFMal2022	Our dataset
The best model of Issakhani et al. study [1](Linear Regression)	Accuracy of 99.89%, F-score of 99.86%, Precision of 99.84%, and Recall of 99.88%	Accuracy of 97.26%, F-score of 94.11%, precision of 95.68%, and recall of 92.58%
Best model of our study (Gradient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

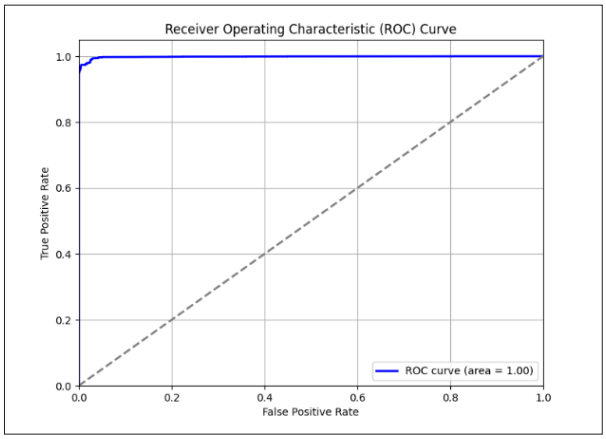
Reference	Dataset	Features	Outcomes
[12]	A random collection of office files, collected according to VirusTotal analysis, included 2,537 files and 4,212 macros (sometimes more than one per file), of which 877 were obfuscated.	A set of 15 lexicographical and function call features	The different machine-learning approaches report accuracies of around 90% in the task of identifying obfuscated macros. MLP was the most prominent with a 92% F2 score.
[11]	7,145 samples, including macros retrieved from VirusTotal.	Different language processing-related features, including SCDV, LSI, Doc2vec, Bag-of-words	The best F1 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.

۴-۵- تاثیر مجموعه داده نامتوازن در خروجی مدل

همان‌طور که در بخش چهارم که مربوط به مجموعه داده می‌باشد بحث شد، عدم توازن در مجموعه داده‌ها می‌تواند چالشی را ایجاد کند که در روند یادگیری و در نتیجه بر نتایج مدل‌ها تأثیر بگذارد. با وجود این که، مقادیر امتیاز F نتایج قابل‌قبول و امیدوارکننده‌ای را نشان می‌دهند که نشان‌دهنده حداقل تأثیر داده‌های نامتوازن بر مدل‌ها است با این حال، برای اثبات بیشتر این ادعا، لازم است شواهد بیشتری ارئه شود. بنابراین، این پژوهش اقدام به ایجاد نمودار منحنی ROC که دورنمای گرافیکی از عملکرد و قابلیت اطمینان مدل‌ها ارائه می‌دهد، کرده است. این نمودارها در شکل‌های ۹ و ۱۰ که به ترتیب مربوط به مجموعه داده پی‌دی‌اف و آفیس هستند ارائه شده‌اند.



شکل ۹: نمودار منحنی ROC برای مجموعه داده پی‌دی‌اف



شکل ۱۰: نمودار منحنی ROC برای مجموعه داده آفیس

پرسش ۴: تاثیر مجموعه داده نامتوازن بر خروجی مدل طبقه‌بندی کننده چیست؟
پاسخ به پرسش چهارم: با استناد بر نمودارهای منحنی ROC برای مجموعه داده پی دی اف و آفیس که در این بخش طی شکل‌های ۹ و ۱۰ نشان داده شد و همچنین امتیاز معیارهای مرتبط نظیر امتیاز F که در بخش ۱-۳-۵ به آن اشاره شد و عدد قابل قبولی را نشان می‌دهند، عدم توازن در مجموعه داده‌های آفیس و پی‌دی‌اف تاثیری نداشته و نتایج حاصل از مدل‌های طبقه‌بندی کننده قابل اتکا و معتبر هستند.

۶- نتیجه‌گیری

در این مطالعه به اهمیت تشخیص بدافزارهای پیچیده غیراجرایی اشاره شد. با توجه به ساختار متمایز فایل‌های پی‌دی‌اف 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 پی‌دی‌اف و جنگل تصادفی برای آفیس به ترتیب به دقت ۹۹/۳ و ۹۹/٤درصد دست پیدا کند و نسبت به مطالعات مشابه عملکرد بهتری را به ثبت برساند.

مراجع

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

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

[3] Kaspersky, 2023. "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. 2020 May. Malware detection in pdf and office documents: A survey. Information Security Journal: A Global Perspective. 3; 29(3):134-53.

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

[7]Gu, J., Kong, R., Sun, H., Zhuang, H., Pan, F., & Lin, Z. 2022 May 24. 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) (Vol. 12260, pp. 599-607). SPIE.

[8]Laskov, P. & Šrndić, N. 2011. 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. 2022 Sep 30. PDF malware detection based on optimizable decision trees. Electronics.; 11(19):3142.

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

[11]Liu, JK., Huang, X. 2019 Dec 10editors. Network and System Security: 13th International Conference, NSS 2019, Sap-

poro, Japan, December 15–18, 2019, Proceedings. Springer Nature.
[12]Kim, S., Hong, S., Oh, J., Lee, H. 2018 Jun 25. Obfuscated VBA macro detection using machine learning. In2018 48th annual ieee/ifip international conference on dependable systems and networks (dsn) (pp. 490-501). IEEE.
[13]Ravi, V., Gururaj, SP., Vedamurthy, HK., Nirmala, MB. 2022 Jun 1Analysing corpus of office documents for macro-based attacks using machine learning. Global Transitions Proceedings. 3(1):20-4.
[14]Casino, F., Totosis, N., Apostolopoulos, T., Lykousas, N, Patsakis, C. 2023 Aug 10. Analysis and correlation of visual evidence in campaigns of malicious office documents. Digital Threats: Research and Practice.;4(2):1-9.
[15]Nissim, N., Cohen, A., Elovici, Y. 2016 Dec 1. 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. 12(3):631-46.
[16]Zakeri-Nasrabadi, M., Parsa, S., Kalaee, A. 2021 Mar 33. Format-aware learn&fuzz: deep test data generation for efficient fuzzing. Neural Computing and Applications. (5):1497-513.
[17]Virus Total, https://www.virustotal.com/gui/home/upload., 2023.
[18]Natekin, A., Knoll, A. 2009 Jul 1. 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.;8(7):579-88.
[20]Breiman L. 2001 Oct. Random forests. Machine learning.45:5-32.
[21]Zenodo, https://zenodo.org/., 2023.
[22]Issakhani, “CIC- Evasive- PDFMal 2022 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. 2017 Oct 5. Baykal N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In2017 International Conference on Computer Science and Engineering (UBMK) (pp. 821-826). IEEE.
[25]Rodríguez-Pérez, R., & Bajorath, J. 2020 Oct. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of computer-aided molecular design.;34(10):1013-26.
[26]Contagio,“https://contagiodump.blogspot.com/2013/03/16800-clean-and-11960-malicious-files.html.”