نشریه علمی "پدافند الکترونیکی و سایبری"

سال یازدهم، شماره ۲، تابستان ۱۴۰۲، ص ۵۵-۴۹

شناسایی بدافزار اندرویدی روز صفر با استفاده از شبکه‌های عصبی

بهزاد لک* [۱]، وحید یادگاری [۲]، احمدرضا متین فر [۳]

۱- استادیار، دانشگاه علوم انتظامی امین،۲- دانشجوی دکتری، دانشگاه علامه طباطبایی،۳- استادیار دانشگاه جامع امام حسین(ع)، تهران، ایران

(دریافت: ۱۴۰۱/۱۲/۰۶، بازنگری: ۱۴۰۲/۰۲/۲۲، پذیرش: ۱۴۰۲/۰۴/۱۱، انتشار: ۱۴۰۲/۰۷/۰۶)

DOR: https://dorl.net/dor/20.1001.1.23224347.1402.11.3.5.0

چکیده

با افزایش ضریب نفوذ اینترنت در زندگی و استفاده آحاد مردم از این فناوری در همه ابعاد، به‌کارگیری از دستگاه‌های گوشی تلفن همراه نیز به همین نسبت افزایش داشته است. این موضوع در کنار خلق مزایای فراوان، موجب گسترش و تسریع انتشار برخی برنامه‌های مخرب به نام بدافزار گردیده است. در این پژوهش سعی بر آن است که با استفاده از شبکه عصبی چندلایه و یادگیری ماشین تشخیص بدافزارهای روز صفر در تلفن‌های هوشمند صورت گیرد. برای این منظور از دیتاست استاندارد با بیش از ۱۵ هزار نمونه از انواع بدافزار و خوب افزار به‌صورت برچسب‌گذاری شده بهره‌گیری شده است. در مرحله پیش‌پردازش ابتدا با استفاده از نرمال‌سازی و یکسان‌سازی داده‌ها انجام می‌شود و با تجزیه‌وتحلیل مؤلفه‌های اصلی عمل انتخاب ویژگی صورت گرفته و از تعداد ۱۱۸۳ ویژگی تعداد ۲۱۵ ویژگی که واریانس بالاتری دارند انتخاب می‌شود و پس‌ازآن مدل پیشنهادی معرفی‌شده است که از طبقه بند شبکه عصبی چندلایه و الگوریتم بهینه‌سازی مبتنی بر آموزش و یادگیری است که با اعمال آن بر روی پایگاه داده‌های ذکرشده و مقایسه نتایج طبقه‌بندی آن با الگوریتم‌های ماشین بردار، الگوریتم ژنتیک ، نزدیک‌ترین همسایه و ... می‌توان دریافت که آموزش شبکه عصبی چندلایه یادگیری دقت و صحت را بالا می‌برد. نتایج استفاده از شبکه عصبی چندلایه مبتنی بر آموزش و یادگیری حاکی از دقت ۹۹٪ و صحت ۹۸٪ است.

واژه‌های کلیدی: بدافزار، اندرزید، تجزیه‌وتحلیل ، انتخاب ویژگی، یادگیری ماشین.

۱. مقدمه

با افزایش ضریب نفوذ اینترنت در زندگی و استفاده آحاد مردم از این فناوری در همه ابعاد و زمینه‌های کاری، استفاده از دستگاه‌های گوشی تلفن همراه نیز به همین نسبت افزایش داشته است. این موضوع در کنار خلق مزایای فراوان، موجب گسترش و تسریع انتشار برخی برنامه‌های مخرب به نام بدافزار گردیده است. بدافزار در اصطلاح به نرم‌افزاری گفته می‌شود که باهدف هک کردن و آسیب رساندن به دستگاه‌ها، سرقت داده‌ها و به‌طورکلی ایجاد خرابکاری نوشته‌شده است. اولین نکته‌ای که باید توجه داشته باشید بدافزارهای اندرویدی عامل اصلی مشکلات امنیتی مختلف در اینترنت است. همه‌روزه به‌طور فزاینده‌ای صنعت بدافزار ندروید در حال رشد هست که تقریباً ۱۲۰۰۰ نمونه‌های جدید بدافزار ندروید هرروز تولید و منتشر می‌شود. تشخیص بدافزارهای اندرویدی درگوشی‌های هوشمند یک هدف ضروری برای جامعه سایبری برای رهایی از نمونه‌های بدافزار تهدیدکننده است [۱]. در این پژوهش قصد داریم با استفاده از یادگیری ماشین و همچنین الگوریتم شبکه عصبی به شناسایی بدافزارها درگوشی‌های هوشمند بپردازیم

۲. مبانی و مفاهیم

۲-۱. ساختار فایل‌های اندرویدی

اندروید یک چارچوب مبتنی بر لینوکس است که توسط شرکت گوگل و اتحادیه ابزارهای منبع باز توسعه داده‌شده است. برنامه‌نویسی برنامه‌های اندرویدی منحصراً در جاوا انجام می‌شود. پَروَنجاهایی با پسوند آی پی کی که مخفف عبارت یا بسته برنامه اندروید می‌باشد. این برنامه‌ها برای استفاده در سیستم‌عامل اندروید تهیه و توسعه داده می‌شوند [۲].

فایل‌های اندرویدی به راحتی می‌توانند توسط بدافزارها آلوده شوند و فروشگاه‌های ندرویدی مثل گوگل پلی بااین‌حال که بر روی برنامه‌های فروشگاهی خود و فایل‌های اندرویدی نظارت و

* رایانامه نویسنده مسئول: behzad_lak@yahoo.com

کنترل دقیقی دارند و معمولاً به‌سرعت برنامه‌های مخرب و تروجان را شناسایی و حذف می‌کنند ولی در برخی موارد این‌گونه برنامه‌ها به دلیل استفاده از الگوریتم‌های پیچیده‌تر و همچنین استفاده از آسیب‌پذیرهای جدید قابل شناسایی نبوده و سیستم را آلوده می‌کنند.[۳]

۲-۲. بدافزارها

به زبان ساده بدافزار به یک برنامه مخرب گفته می‌شود که برای کاربران کامپیوتری مضر است. بدافزارها عملکردهای متفاوتی را مثل سرقت اطلاعات – تغییر یا حذف داده‌ها – کدگذاری ناخواسته روی‌داده‌ها – و یا مانیتور کردن فعالیت‌های کاربر بدون مجوز او را انجام می‌دهند [۴].

برای مثال اطلاعات مربوط به کارت شناسایی ، شماره‌حساب‌های بانکی ، رمزهای عبور و نام‌های کاربری و امثال این اطلاعات را جمع‌آوری و برای نویسنده آن بدافزار ارسال می‌کنند.[۵]

۲-۳. روز صفر

آسیب‌پذیری روز صفر یک آسیب‌پذیری در یک محصول نرم‌افزاری است که یا هنوز کشف نشده و اگر کشف شده هنوز وصله‌ای برای آن ارائه نشده است. هکرها از این آسیب‌پذیری‌ها استفاده می‌کنند و حملات هدفمندی را علیه سیستم‌ها و شبکه‌های سازمان‌هایی که دارای چنین آسیب‌پذیری‌هایی هستند طراحی می‌کنند [۵].

۲-۴. یادگیری ماشین

یادگیری ماشین مطالعه الگوریتم‌های کامپیوتری است که به‌طور خودکار از طریق تجربه بهبود می‌یابند. به‌طورکلی می‌توان الگوریتم‌های یادگیری ماشین را در سه دسته تقسیم‌بندی کرد: با نظارت، بدون نظارت و نیمه نظارتی[۵].

۲-۵. یادگیری بانظارت

در حالت یادگیری با نظارت برای هر داده ورودی، یک خروجی که از قبل برچسب خورده موجود می‌باشد. الگوریتم بر اساس این ورودی‌ها و خروجی‌ها آموزش می‌بینند تا بتواند در مواجهه با یک ورودی جدید که تاکنون آن را مشاهده نکرده خروجی درستی را تولید کند. در یادگیری با نظارت اگر داده‌های مساله برای یادگیری به صورت گسسته باشند این مساله طبقه‌بندی نامیده می‌شود و اگر مقادیر داده‌ها به‌صورت پیوسته باشند به آن رگرسیون گویند [۵].

۲-۶. یادگیری بدون‌نظارت

در حالت بدون نظارت، الگوریتم یادگیرنده برای ورودی‌های مشخص هیچ خروجی را در دسترس ندارد و به بیانی دیگر داده‌ها برچسب نخورده‌اند. هدف در این یادگیری تخمین این خروجی‌ها توسط خود الگوریتم از طریق تحلیل داده‌ها و شناسایی الگوهای خاص موجود در ساختار داده‌ها می‌باشد[۶].

۲-۷. یادگیری نیمه‌نظارتی

یادگیری نیمه‌نظارتی ترکیبی از یادگیری بانظارت و بدون‌نظارت می‌باشد، یعنی یادگیری بر اساس داده‌های برچسب‌دار و داده‌های بدون برچسب انجام می‌گیرد و بیشتر هنگامی استفاده می‌شود که تعداد داده‌های بر چسب‌دار کم باشد[۶].

۳. مروری بر پیشینه

رحمان و همکاران یک چارچوب ترکیبی کارآمد برای تشخیص بدافزار در برنامه‌های کاربردی اندروید ارائه کردند که از دو روش تحلیل اکتشافی و امضا بهره می‌گرفت. به این منظور دسته‌بندی، از ماشین بردار پشتیبان ، درخت تصمیم، نزدیک‌ترین همسایه و جی ۴۸ استفاده کردند. نتایج به‌دست‌آمده نشان داد که ماشین بردار پشتیبان در حالت دودویی‌ها و نزدیک‌ترین همسایه در حالت فایل‌های منیفست بهترین گزینه برای تشخیص بدافزار در دستگاه‌های اندروید هستند.[۶].صیادی و همکاران از روش‌های یادگیری تجمعی جهت بهبود عملکرد سامانه‌های تشخیص بدافزار مبتنی بر عملکرد شمارنده‌های سخت‌افزاری استفاده کردند. برای این منظور آن‌ها از هشت مدل یادگیری ماشین مقاوم و دو دسته‌بندی تجمعی مشهور استفاده کردند.[۷]

پاتل و همکاران به تحلیل کاملی از روش‌های یادگیری ماشین جهت دسته‌بندی برنامه‌های سالم و مخرب پرداختند. آن‌ها به‌منظور پیاده‌سازی نرم‌افزاری، تمامی دسته‌بندها را در هسته سامانه عامل به‌طور کامل پیاده‌سازی کردند تا تأثیرات نرم‌افزاری مختلف مشخص شود. علاوه بر این به‌منظور اجرای سریع، این الگوریتم‌ها را بر روی اف پی جی آی پیاده‌سازی نمودند. [۸]

ساینگ و همکاران یک تحلیل جامع از به‌کارگیری رویکردهای یادگیری ماشین و شمارنده‌های عملکرد سخت‌افزار را برای یک زیرمجموعه خاص از نرم‌افزارهای مخر روتکیت‌های هسته ارائه دادند. علاوه براین آن‌ها از روش‌های انتخاب ویژگی جهت تعیین ویژگی‌های کارآمد استفاده نمودند. [۹]

لئو و همکاران یک روش تشخیص بدافزار جدید بر اساس یادگیری ماشین ارائه کرده‌اند. در این روش با استفاده از الگوی ان گرام از او پی کدهای بدافزارها تصاویر سیاه‌وسفید ایجاد کرده که این تصاویر برای استخراج ویژگی‌ها به‌منظور خوشه‌بندی بدافزارهای ناشناس بر اساس الگوریتم نزدیک‌ترین همسایه مشترک استفاده می‌شود. نویسندگان از یک مجموعه داده شامل ۲۰۰۰۰ نمونه بدافزار و فایل بی‌خطر برای ارزیابی مدلشان

استفاده کرده‌اند. نتایج ارزیابی آن‌ها نشان می‌دهد که بهترین دقت برای دسته‌بندی بدافزارهای ناشناس ۹۸,۹ درصد بوده است همچنین دقت میانگین روش آن‌ها برای تشخیص بدافزارهای مدرن ۸۶,۷ درصد است[۲]

در مقاله آرورا و همکاران یک الگوریتم برای اولویت‌بندی ویژگی‌های ترافیک شبکه با حداقل تعداد ویژگی‌ها برای تجزیه‌وتحلیل دقت بیشتر و زمان پردازش ارائه کرد. نتایج نشان داد که ۹ ویژگی از ۲۲ مورد برای ارائه حداکثر دقت تشخیص (۸۵ بیش از تا ۱۰۰) کافی است. علاوه بر این ، نویسندگان ادعا کردند که این ۹ ویژگی همچنین زمان قابل‌توجهی را برای آموزش و آزمایش مجموعه داده صرفه‌جویی می‌کند. زمان آموزش ۳۰۰ برنامه کاربردی از ۱۱,۷ ثانیه به ۵,۸ ثانیه و زمان تست ۲۳۰ برنامه از ۲۵,۱ ثانیه به ۱۷,۳ ثانیه کاهش می‌یابد.[۱۰]

هَنسِن و همکاران با سفارشی‌سازی سند باکس کوکو ابزاری برای تحلیل حجم انبوه بدافزارهایی که برای تحلیل ایستا مبهم بودند ارائه کرده‌اند. نویسندگان از الگوریتم آر اف سی برای تشخیص و دسته‌بندی استفاده کرده‌اند. مجموعه داده نویسندگان از ۲۷۰۰۰۰ بدافزار و ۸۳۷ خوب افزار تشکیل شده است.[۱۱]

• ایمران و همکاران از روش تحلیل نمادین برای تشخیص بدافزارها استفاده کرده‌اند. روش آن‌ها یک روش ترکیبی بوده که از فراخوانی‌های سیستمی استخراج شده از تحلیل ایستا و پویا به‌عنوان نماد استفاده کرده است. نویسندگان از مدل مارکوف مخفی برای دسته‌بندی بهره برده‌اند.[۱۲]

اکثر روش‌های موجود که برای شناسایی بدافزارها به کار می‌روند مبتنی بر امضا می‌باشند ، بنابراین در جلوگیری و شناسایی از بدافزارهای جدید توسعه‌یافته شکست می‌خورند. هدف از این تحقیق طراحی یک شبکه عصبی است که می‌تواند احتمال وجود یک بدافزار از نوع روز صفر را درون برنامه کاربردی اندروید تعیین کند

۴. روش تحقیق

از آنجایی که شبکه عصبی بر روی هزاران برنامه کاربردی اندرویدی موجود آموزش می‌بینند ، می‌تواند احتمال وجود هرگونه برنامه جدید در صورت داشتن بدافزار را با موفقیت پیش‌بینی کند ، حتی اگر این بدافزار به تازگی توسعه‌یافته باشد. در صورت عدم توجه به توسعه الگوریتمیک بدافزارهایی از نوع روز صفر، توسعه این دسته از بدافزار موجب تخریب و اختلال در سیستم عامل اندورید گوشی‌های هوشمند خواهد شد.

ابتدا بایستی فایل مورد نظر به یکی از دو روش ایستا و پویا و یا با ترکیب این دو روش مورد آنالیز قرار گرفته و کلیه ویژگی‌های رفتاری و دسترسی‌ها و ورودی و خروجی‌های برنامه بصورت یک خروجی با هم ترکیب شده و تشکیل یک مجموعه را می‌دهد. پس از تجمیع خروجی‌ها و تشکیل به یک بانک اطلاعاتی، داده‌ها مورد بررسی قرار گرفته و ویژگی‌های مؤثر و پرکاربرد شناسایی و استخراج می‌گردد. سپس ویژگی‌های مفید به‌عنوان به دیتاست در اختیار سامانه‌های آموزش ماشین قرار می‌گیرد تا در صورت تشخیص الگوی مشابه با بدافزارهای شناسایی‌شده قبلی به‌عنوان بدافزار شناسایی گردد.

استخراج ویژگی‌ها یکی از مهمترین مراحل تشخیص است چرا که ویژگی‌های استخراج شده از برنامه‌ها می‌توانند رفتار و عملکردهای برنامه‌ها را مشخص کنند و بایستی ویژگی‌هایی بررسی شوند که این رفتار را به‌طور کامل و صریح نمایان می‌کنند [۱۹].

علاوه بر مطالب ذکرشده ویژگی‌ها بایستی این توانایی را داشته باشند که رفتار ایستا و پویا برنامه را تشریح کنند. در همین راستا در روش پیشنهادی از دودسته ویژگی‌های ایستا و پویا استفاده می‌شود. این ویژگی‌ها سطوح دسترسی، کلیه تعاملات برنامه و همچنین رفتارهای محیطی برنامه‌های اندروید را مشخص می‌کنند[۱۸].

۴-۱. تحلیل ایستا

فایل اندروید مانیفست دات ایکس ام ال شامل ویژگی‌های زیادی است که می‌تواند برای تجزیه‌وتحلیل استاتیک مورد استفاده قرار گیرد[۱۷].

• فعالیت‌ها : فعالیت اندروید یک صفحه از رابط کاربری اندروید است.

• گیرنده‌ها و ارائه‌دهندگان پخش

• متادیتا : اساساً یک گزینه اضافی برای ذخیره اطلاعاتی است که می‌تواند از طریق کل پروژه قابل دسترسی باشد.

• مجوزهای درخواست شده توسط برنامه: این حریم خصوصی کاربر را محافظت می‌کند و برای دسترسی به داده‌های حساس کاربر مورد نیاز است (مانند مخاطبین و پیامک)

• ویژگی‌های سیستم (مانند دوربین و اینترنت)

۴-۲. تحلیل پویا

برای درک تغییرات رفتاری این دسته‌بندی‌های بدافزار و خانواده‌ها، پس از اجرای نرم‌افزارهای مخرب در یک محیط خالی، شش دسته از ویژگی‌ها شامل حافظه، ای پی آی ، شبکه، باتری، اکسس ، لاگ چت و فرایند استخراج و در سمت گوشی و یا سرور ، مراحل پیش‌پردازش و بررسی و تحلیل داده‌ها

را در گوشی‌های هوشمند کاربران و یا سرور انجام می‌دهند [۱۶] [۲۰].

۴-۳. تحلیل مؤلفه‌های اساسی

حجم بالای داده‌ها باعث پیچیدگی داده‌ها و در نتیجه پیچیده‌تر شدن مدل‌های ما شده و افزایش حجم محاسبات را در پی دارد، از این رو نیاز به روش‌هایی جهت کاهش ابعاد داریم .

هنگامی که یک مجموعه داده دارای متغیرها یا ویژگی‌های زیادی باشند، می‌توان با روش‌هایی مثل همبستگی بعضی متغیرها و... و الگوریتم‌های کاهش بعد، امکان بهبود تحلیل مؤلفه‌ها را انجام داد[۲۰].

۴-۴. الگوریتم مورد استفاده شبکه عصبی چندلایه

شبکه‌های عصبی قابلیت خود تطبیقی، خودسازمان‌دهی و عملیات بلادرنگ و ... را دارند. ساختار این شبکه شامل یک‌لایه ورودی، یک‌لایه میانی و یک‌لایه خروجی است. در هر لایه، یک یا چند عنصر پردازشگر(نورون)، وجود دارد که با تمامی نورون‌های لایه بعدی با اتصالات وزن‌دار، به هم مربوط می‌شوند. تعداد نورون‌های لایه‌های ورودی و خروجی، بستگی به تعداد متغیرهای ورودی و خروجی مدل دارد؛ ولی انتخاب تعداد نورون‌های لایه میانی به‌صورت سعی و خطا تعیین می‌شود که در مسأله ما معمولاً بهترین جواب تشخیص بدافزار با تعداد نورون‌های ۱۰ در لایه پنهان به دست می‌آید.[۲۱]

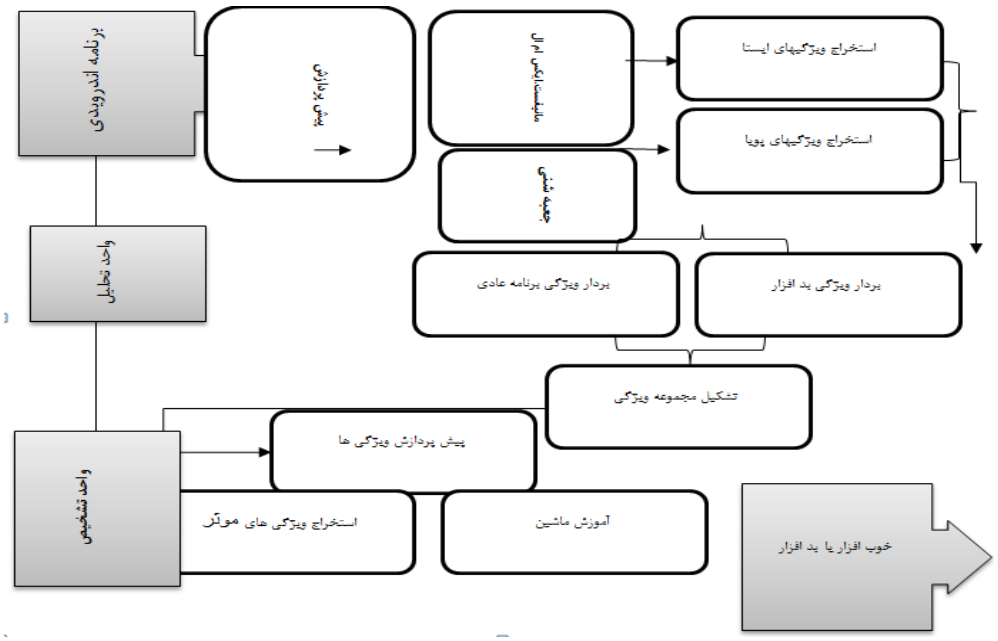
۴-۵. دیتاست

دیتاست استاندارد مورد استفاده در این تحقیق محصول پروژه موسسه امنیت سایبری کانادا می‌باشد.این کار تحقیقاتی یک مجموعه داده بدافزار جامع و عظیم اندرویدی جدید می‌باشد. مجموعه داده‌ها شامل نمونه‌های بدافزار و خوب افزار با حجم K۲۰۰ و K۲۰۰ است که مجموعاً K۴۰۰ می‌باشد و شامل ۱۴ دسته بدافزار و ۱۹۱ خانواده بدافزار می‌باشد. (البته در این تحقیق فقط از دسته بدافزارهای روز صفر استفاده شده است) [۲۲].

۵. مدل پیشنهادی

روش پیشنهادی یک روش تشخیص ترکیبی مبتنی بر اجزای اصلی برنامه‌های اندروید برای تشخیص بدافزارهای اندروید است. این روش شامل "مرحله‌ی آموزش"، "مرحله‌ی تشخیص" و پنج زیرسیستم "پیش‌پردازش ایستا"، "پیش‌پردازش پویا"، "استخراج ویژگی‌های ایستا"، "استخراج ویژگی‌های پویا"و "مدل‌سازی" است. که در شکل ۱ معماری سیستم پیشنهادی مشاهده می‌شود.

در ادامه زیرسیستم‌های معماری پیشنهادی ارائه می‌شود.



شکل (۱). مدل پیشنهادی

۵-۳. انتخاب ویژگی

همان‌طور که گفته شد در این پروژه از ۱۳۸۱ ویژگی استفاده‌شده است. از مزایایی که تعداد ویژگی بالا می‌تواند داشته باشد، تشخیص دقیق‌تر و کامل‌تر است؛ اما از طرفی معایبی نیز می‌تواند داشته باشد. یکی از معایب مهم آن رخداد بیش‌ازحد مناسب است. بر اثر این رخداد باعث می‌شود که مدل یادگیری شده دقیقاً مطابق داده‌های آموزش شود و در تشخیص داده‌های جدید ناتوان باشد. از معایب دیگر آن یادگیری و تشخیص بسیار زمان‌بر و با سرعت بسیار پایین است. به‌طورکلی هر چه ابعاد یا همان ویژگی‌های مساله‌ی مورد کاوش بالاتر باشد، باعث خواهد شد که رکوردها در فضای جستجو پراکنده‌تر شوند.

روش پیشنهادی انتخاب ویژگی، ویژگی‌های مؤثر

از تعداد کل ویژگی‌های مورداستفاده در این پژوهش که ۱۳۸۱ ویژگی است، ۷۲ ویژگی از گروه ویژگی‌های اجزای سخت‌افزاری ۸۱۲، ویژگی از گروه ویژگی‌های مجوزهای درخواست شده ۳۷۹، ویژگی از گروه ویژگی‌های قصدهای فیلترشده ۴۷۰، ویژگی از گروه ویژگی‌های فراخوانی رابطه‌های برنامه‌نویسی محدود شده ۷۰، ویژگی از گروه ویژگی‌های مجوزهای استفاده شده و ۲۱۵ ویژگی از گروه ویژگی‌های فراخوانی رابطه‌های برنامه‌نویسی حساس است. در این روش، زیرمجموعه‌ای از این تعداد ویژگی انتخاب می‌شود.

به‌طور مثال بدافزارها نسبت به خوب افزارها تمایل بیشتری به دسترسی به بخش پیامک‌های یک سیستم اندرویدی را دارند و ویژگی‌های مرتبط با این بخش را بیشتر از خوب افزارها دارند. تفاوت در نوع ویژگی‌ها می‌تواند نقطه قوتی برای تشخیص بهتر بدافزارهای اندرویدی از خوب افزارها باشد و توجه به همین تفاوت‌ها می‌تواند رفتار مخرب یا غیر مخرب بودن یک نرم‌افزار را به‌خوبی نشان دهد. به همین دلیل در این روش ما از این تفاوت‌ها برای تشخیص بدافزارها استفاده کردیم.

انتخاب ویژگی را می‌توان به‌عنوان فرآیند شناسایی ویژگی‌های مرتبط و حذف ویژگی‌های غیر مرتبط و تکراری با هدف مشاهده زیرمجموعه‌ای از ویژگی‌ها که مساله را به‌خوبی و با حداقل کاهش درجه کارایی تشریح می‌کنند تعریف کرد. در سناریوهای «تحلیل کلان داده»، انتخاب ویژگی نقشی اساسی ایفا می‌کند. تحلیل مؤلفه‌های اصلی یکی از انواع روش‌های تحلیل داده‌های چندمتغیره است که هدف اصلی آن تقلیل بُعد مسأله موردمطالعه است. یکی از کاربردهای مهم تحلیل مؤلفه‌های اصلی، در طبقه‌بندی است. الگوریتم تحلیل مؤلفه‌های اصلی داده‌ها را از فضای ورودی به فضایی جدید نگاشت می‌کند به‌طوری ویژگی‌ها بر اساس پراکندگی(واریانس) مرتب می‌شوند و وابستگی بین ویژگی‌ها در فضای جدید وجود ندارد. در این تحقیق با استفاده از الگوریتم تحلیل مؤلفه‌های اصلی از میان ۱۳۸۱ ویژگی موجود در پایگاه داده تعداد ۴۷۰ ویژگی برای طبقه‌بندی دقیق‌تر و در انتها با یک تحلیل دقیق‌تر ۲۱۵ ویژگی به‌عنوان تأثیرگذارترین ویژگی‌ها انتخاب‌شده‌اند. در این روش ابتدا برای هر ویژگی تعداد بدافزار و خوب افزار دارای آن ویژگی را مشخص کردیم. به دلیل اختلاف تعداد کل بدافزارها و خوب افزارها، نسبت آن‌ها به‌کل بدافزارها و خوب افزارها محاسبه شد. به‌عبارت‌دیگر تعداد بدافزارهای هر ویژگی تقسیم‌بر تعداد کل بدافزارها یعنی ۶۸۲۰ و تعداد خوب افزارها تقسیم‌بر تعداد کل خوب افزارها یعنی ۱۲۰۱۵ شد. این عمل نشان می‌دهد که یک ویژگی به چه اندازه توسط بدافزارها و خوب افزارها مورداستفاده قرارگرفته است، یا به‌عبارت‌دیگر چند درصد از بدافزارها و خوب افزارها از یک ویژگی استفاده کرده‌اند. اختلاف بین نسبت تعداد بدافزارها و نسبت تعداد خوب افزارها نشانگر این است که یک ویژگی بیشتر توسط کدام گروه از نرم‌افزارها مورداستفاده قرارگرفته است. در این مرحله فهرستی از تمام ویژگی‌ها به همراه اختلاف بین نسبت تعداد بدافزارها و نسبت تعداد خوب افزارها به دست آمد که میانگین آن‌ها ۰,۰۰۰۱۱۱۹۷ و انحراف معیار آن ۰,۰۰۳۲۰۴۳۲ است .انحراف معیار می‌تواند در این تحقیق تفاوت بین بدافزار و خوب افزار را بهتر می‌تواند نشان دهد. درنتیجه انحراف معیار آن‌ها محاسبه شد و ویژگی‌هایی که مقدار اختلاف بین تعداد بدافزارها و خوب افزارهای آن‌ها بیشتر از یک انحراف معیار [که منظور مجموع میانگین و انحراف معیار است] بود به‌عنوان ویژگی‌های انتخابی در این روش انتخاب شدند. در این روش از ۱۳۸۱ ویژگی که در ابتدای کار داشتیم،۲۱۵ ویژگی به‌عنوان ویژگی‌هایی که مؤثر در تشخیص بدافزار و خوب افزار اندروید هستند، انتخاب شدند.

۶. ارزیابی و مقایسه روش

در این بخش به ارزیابی مدل‌های مختلف پرداخته و خروجی نهایی حاصل از اجرای الگوریتم‌ها را با روش پیشنهادی مبتنی بر شبکه‌های عصبی چندلایه مقایسه می‌کنیم

۶-۱. مقایسه دقت

در جدول ۱ دقت محاسبه‌شده برای الگوریتم شبکه عصبی عمیق۹۹٪ هست که در مقایسه با دیگر الگوریتم‌ها از دقت بیشتری برخوردار هست.

جدول (۱). مقایسه دقت

روش (مدل)	دقت
ماشین بردار پشتیبان	۹۷,۵۲
شبکه عصبی چندلایه	۹۸,۴۲
نزدیک‌ترین همسایه	۹۷,۳۷
گرادیان کاهشی تصادفی	۹۷,۶۸
تقویت‌کننده(Ada)	۹۶,۳۲
بیز ساده	۷۴,۸۹

۶-۲. مقایسه صحت

در جدول ۲ برای محاسبه و ارزیابی صحت مدل پیشنهادی در مقایسه با دیگر مدل‌ها مشاهده می‌شود که دارای بیشترین صحت در بین آن‌ها هست.

جدول (۲). مقایسه صحت

روش (مدل)	صحت
ماشین بردار پشتیبان	۹۷،۳۲
شبکه عصبی چندلایه	۹۸،۲۹
نزدیک‌ترین همسایه	۹۷،۳۱
گرادیان کاهشی تصادفی	۹۷،۰۸
تقویت‌کننده(Ada)	۹۵،۵۶
بیز ساده	۶۰،۰۱

۶-۳. ارزیابی فراخوانی

در اینجا با محاسبه مقدار فراخوانی الگوریتم‌ها و مقایسه آن‌ها با الگوریتم پیشنهادی می‌توان به بهتر بودن آن در نسبت با دیگر مدل‌ها پی برد جدول ۵ مقدار محاسبه‌شده برای هر یک از الگوریتم‌ها را نشان می‌دهد.

جدول (۳). مقایسه فراخوانی مدل پیشنهادی با دیگر مدل‌ها

روش (مدل)	فراخوانی	
ماشین بردار پشتیبان	۹۵،۹۶	
شبکه عصبی چندلایه	۹۷،۶۶	
نزدیک‌ترین همسایه	۹۵،۵۵	
گرادیان کاهشی تصادفی	۹۶،۵۴	
تقویت‌کننده(Ada)	۹۴،۴۷	
بیز ساده	۹۷،۲۱	

۶-۴. ارزیابی F1

در جدول ۶ مقایسه نتایج حاصله از محاسبه ارزیابی F1 الگوریتم‌ها قابل‌مشاهده هست.

جدول (۴).مقایسه ارزیابی F1 الگوریتم‌ها

روش (مدل)	F1
ماشین بردار پشتیبان	۹۶،۶۲
شبکه عصبی چندلایه	۹۸،۰۸
نزدیک‌ترین همسایه	۹۶،۴۲
گرادیان کاهشی تصادفی	۹۶،۹۹
تقویت‌کننده(Ada)	۹۵
بیز ساده	۷۴،۱۹

ما در این پژوهش برای پیاده‌سازی الگوریتم پیشنهادی از زبان برنامه‌نویسی پایتون ۳ استفاده کرده‌ایم و به‌منظور استفاده از شرایط مساوی برای افرادی که در آینده می‌خواهند این روش را بسط و گسترش دهند از محیط ابری کولب ارائه‌شده توسط گوگل استفاده‌شده است.

۷. نتیجه‌گیری

در این پژوهش سعی شد تا به روشی بهبودیافته برای تشخیص و شناسایی بدافزارهای سیستم‌عامل اندروید، با بررسی ویژگی‌های مختلف یک نرم‌افزار اندروید دست‌یافت. نرم‌افزارها در این پژوهش بر اساس ویژگی‌ها و قابلیت‌هایی که از دو فایل اندرویدمانیفست دات ایکس ام ال و کلسس دات دکس استخراج‌شده بود مورد ارزیابی قرار گرفتند. ویژگی‌های استخراج‌شده از این دو فایل شامل ۶ گروه مختلف بود که عبارت‌اند از : گروه ویژگی‌های اجزای سخت‌افزاری، گروه ویژگی‌های مجوزهای درخواست شده، گروه ویژگی‌های اینتنت (قصد)های فیلترشده، گروه ویژگی‌های فراخوانی رابطه‌های برنامه‌نویسی محدودشده، گروه ویژگی‌های مجوزهای استفاده‌شده و گروه ویژگی‌های فراخوانی رابطه‌های برنامه‌نویسی حساس. برای بهینه‌تر شدن روش تشخیص بدافزار، در این پژوهش، ابتدا با روش انتخاب ویژگی‌های مؤثر بود که با بررسی کاربرد ویژگی‌ها در خوب افزارها و بدافزارها از بین ۱۳۸۱ ویژگی که در ابتدا وجود داشت، ۴۷۰ ویژگی به‌عنوان ویژگی‌های مؤثر انتخاب شدند. و سپس در گام دوم با انتخاب ویژگی‌های وزن بالا و با بررسی تأثیر و میزان وزنی که هر ویژگی در تشخیص بدافزارها داشت، ۲۱۵ ویژگی برای ارزیابی نرم‌افزارها انتخاب شدند.

در ادامه به جهت تشخیص بدافزارها با استفاده از ویژگی‌های عملکردی هرکدام از مجوزهای مورداستفاده در برنامه‌ها، روش بهینه‌شده شبکه عصبی چندلایه ارائه شد. با توجه به قابلیت تحلیل بسیار بالای ویژگی‌های رفتاری و همچنین تعداد مجوزهای مورداستفاده و نیز کارایی مناسب دسته‌بندی به کار گرفته‌شده، روش پیشنهادی توانست با دقت بیش از ۹۹ درصد فایل‌های آلوده به بدافزار را از فایل‌های سالم تشخیص دهد. نتایج حاصل از مقایسه‌ها نشان می‌دهد روش پیشنهادی در مقایسه با بهترین روش‌های پیشین بیش از ۰/۰۱ درصد بهبود دارد.

۸. مراجع

[1] Bartos, Karel, Michal Sofka, and Vojtech Franc. "Optimized Invariant Representation of Network Traffic for Detecting Unseen Malware Variants." USENIX Security Symposium. 2016.

[2] Liu, L.; Wang, B. Sh.; Yu, B.; Zhong, Q. X. “Automatic Malware Classification and New Malware Detection Using Machine Learning”; Front. Inf. Technol. Electron. Eng. 2017, 18, 1336–1347. Doi: 10.1631/FITEE.1601325.

[3] Seo, S. H.; Gupta, A.; Mohamed Sallam, A.; Bertino, E.; Yim,K. “Detecting Mobile Malware Threats to Homeland Security through Static Analysis”; J. Netw. Comput. Appl. 2014, 38, 43-53. Doi:10.1016/j.jnca.2013.05.008.

[4] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. 13, (2018). Doi: 10.1109/MCI.2018.2840738.

[5] Nayeem, Kh.; Johari, A.; Adnan, Sh. “Defending Malicious Script Attacks Using Machine Learning Classifiers”; Wirel.Commun. Mob. Com. 2017.

[6] Z.-U. Rehman et al., “Machine learning-assisted signature and heuristic-based detection of malwares in Android devices,” Computers & Electrical Engineering, vol. 69, pp.828-841, 2018. Doi:10.1016/j.compeleceng.2017.11.028.

[7] H. Sayadi, N. Patel, S. M. PD, A. Sasan, S. Rafatirad, and H.Homayoun, “Ensemble learning for effective run-time hardware-based malware detection: A comprehensive analysis and classification,” in 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC), IEEE, pp. 1-6, 2018. Doi:10.1145/3195970.3196047.

[8] N. Patel, A. Sasan, and H. Homayoun, “Analyzing hardware based malware detectors,” in 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC), IEEE, pp. 1-6, 2017. Doi:10.1145/3061639.3062202.

[9] B. Singh, D. Evtyushkin, J. Elwell, R. Riley, and I.Cervesato, “On the detection of kernel-level rootkits using hardware performance counters,” in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 483-493, 2017. Doi:10.1145/3052973.3052999.

[10] Arora, Anshul, and Sateesh K. Peddoju. ”Minimizing Network Traffic Features for Android Mobile Malware Detection.”Proceedings of the 18th International Conference on Distributed Computing and Networking. ACM, 2017. Doi:10.1145/3007748.3007763.

[11] Hansen, S.; Larson, M. L.; Stevanovic, M.; Pedersen, J. M. “An Approach for Detection and Family Classification of Malware Based on Behavioral Analysis”; Int. Conf. on Computing, Networking and Communications, 2016. Doi: 10.1109/ICCNC.2016.7440587.

[12] Imran, M.; Afzal, M. T.; Qadir, M. A.; Xiao, Zh.; Li, K. “Malware Classification using Dynamic Features and Hidden Markov Model”; J. Intell. Fuzzy Syst. 2016, 31, 837. Doi: 10.3233/JIFS-169015.

[13] S. Dash, Suarez-Tangil, K. G, T. S, A. K, K. J. M and L. Cavallaro, "DroidScribe: Classifying android malware based on runtime behavior," in Mobile Security Technologies (MoST 2016), 2016 . Doi:10.1109/SPW.2016.25.

[14] Mohaisen, A.; Alrawi, O.; Mohaisen, M. “AMAL: High-Fidelity, Behavior-Based Automated Malware Analysis and Classification”; Comput. Secur. 2015, 52, 251–266. Doi:10.1016/j.cose.2015.04.001.

[15] S. Dai and A. Tongaonkar and X. Wang and A. Nucci and D.Song, Network Profiler: Towards automatic fingerprinting of Android apps, Proceedings IEEE INFOCOM,p809-817, 2013. Doi:10.1109/INFCOM.2013.6566868.

[16] J. Sahs and L. Khan, "A Machine Learning Approach to Android Malware Detection," in European Intelligence and Security Informatics Conference - IEEE, 2012 . Doi:10.1109/EISIC.2012.34.

[17] G. Dini, F. Martinelli, A. Saracino and D. Sgandurra, "MADAM: a MultiLevel Anomaly Detector for Android Malware," Computer Network Security. MMM-ACNS 2012. Springer, vol. 7531, pp. 240-253,2021. https://link.springer.com/chapter/10.1007/978-3-642-33704-8_21.

[18] B. Sanz, I. Santos, C. Laorden, X. Ugarte-Pedrero, P. G. Bringas and G. Alvarez, "PUMA: Permission Usage to detect Malware in Android," Advances in Intelligent Systems and Computing, vol. 189, no. AISC, pp. 289-298,2020۰. https://link.springer.com/chapter/10.1007/978-3-642-33018-6_30 .

[19] Javaheri, D. “A Solution for Recognition and Confronting of Obfuscation and Stealth Techniques of Behavior in Spywares”;Ph.D. Thesis, Islamic Azad University, Science and Research Branch, Tehran, Iran, 2018 (In Persian).

[20] M. Damshenas, A. Dehghantanha, K.-K. R. Choo and R. Mahmud, "M0Droid: An Android Behavioral-Based Malware Detection Model," Journal of Information Privacy and Security, vol. 11, no. 3, pp. 141-157 , 2015 . Doi:10.1080/15536548.2015.1073510.

[21] G. Ciaburro and B. Venkateswaran, Neural Networks with R. Packt Publishing, 2017.

Scientific Journal of Electronic & Cyber Defense
Vol. 11, No. 3, Autumn 2023, Serial No. 43
ISSN: 2322-4347 , E-ISSN: 2980-8979

dentifying Zero Day Android Daily through Neural Networks

Behzad lak[1]*, Vahid yadeghari[2], Ahmadreza matinfar

Assistant Professor, Amin University of Management Sciences, Tehran, Iran

((Received: 2023/02/25, Revised: 2023/05/12, Accepted: 2023/07/02, Published: 1402/09/28))

DOR: https://dorl.net/dor/20.1001.1.23224347.1402.11.3.5.0

Abstract

With the increase in the Internet's penetration rate in life and the use of this technology in all aspects, the use of mobile phones has increased as well. This, in addition to creating many benefits, has expanded and accelerated the release of some malicious programs called malware. In this study, it is attempted to use a multilayer neural network and learning machine diagnosis of zero daytime malware on smartphones. For this purpose, the standard database has been labeled with more than 15,000 samples of malware and goodware. In the pre -processing phase, the data is first performed using normalization and alignment of the data and by analyzing the main components of the feature of the selection of the feature and selected from 1183 features 215 features that have higher variances, followed by the model. A suggestion is introduced from the multilayer neural network class and the optimization algorithm based on the training and learning that apply it to the databases and compare its classification results with vector algorithms, genetic algorithm, nearest neighbor. And ... it can be seen that the neural network training increases accuracy and accuracy. The results of the use of multilayer neural network based on education and learning indicate 99% accuracy and 98% accuracy.

Keywords: Malware, Android, analysis, feature selection, machine learnin

*Corresponding Author Email: *behzad_lak@yahoo.com*