دانشگاه تربیت مدرس

دانشکده مهندسی برق و کامپیوتر

گزارش سمینار کارشناسی ارشد

مهندسی کامپیوتر (گرایش مهندسی نرم افزار)

گزارش جامع از مقالات در زمینه استفاده از یادگیری ماشین برای شناسایی بدافزار در سیستم‌های مختلف

دانشجو

علیرضا حیدرآبادی زاده

استاد راهنما

دکتر مریم لطفی

مقدمه

با گسترش روزافزون استفاده از اینترنت و فناوری‌های دیجیتال، تهدیدات امنیتی نظیر بدافزارها به یکی از معضلات جدی تبدیل شده‌اند. این بدافزارها با استفاده از روش‌های پیشرفته و گوناگون به سیستم‌ها حمله کرده و می‌توانند خسارات جبران‌ناپذیری به امنیت داده‌ها، اطلاعات شخصی و زیرساخت‌های حیاتی وارد کنند. یکی از چالش‌های بزرگ در مقابله با بدافزارها، افزایش پیچیدگی و تطابق‌پذیری آن‌ها با سیستم‌های مختلف است. در این زمینه، یادگیری ماشین (ML) به عنوان ابزاری برای شناسایی و پیش‌بینی حملات به طور گسترده مورد استفاده قرار گرفته است.

این دو مقاله به بررسی استفاده از تکنیک‌های یادگیری ماشین برای شناسایی بدافزار در سیستم‌های مختلف مانند کامپیوترهای شخصی، دستگاه‌های موبایل، اینترنت اشیاء (IoT) و محیط‌های ابری می‌پردازند. مقاله اول به بررسی روش‌های یادگیری ماشین در تشخیص بدافزارهای چندپلتفرمی اختصاص دارد و مقاله دوم رویکردی برای استفاده از سیستم‌های توزیع‌شده همراه با یادگیری ماشین در شناسایی بدافزارها را بررسی می‌کند.

اهداف مقاله‌ها

در این بخش، اهداف دقیق و فنی هر دو مقاله در زمینه استفاده از یادگیری ماشین برای تشخیص بدافزارها در سیستم‌های مختلف، به طور جامع و تکنیکی توضیح داده خواهد شد. هر مقاله به جنبه‌های خاصی از تشخیص بدافزار پرداخته است و اهداف مشخصی را برای حل مشکلات موجود در زمینه امنیت سایبری دنبال کرده‌اند.

۱.مقاله اول «A Survey on ML Techniques for Multi-Platform Malware Detection» :

این مقاله بیشتر به **بررسی و تحلیل** استفاده از تکنیک‌های یادگیری ماشین در تشخیص بدافزارها در پلتفرم‌های مختلف اختصاص دارد. اهداف اصلی مقاله به‌طور دقیق شامل موارد زیر است:

۱.۱توسعه روش‌های تطبیقی برای تشخیص بدافزار در چند پلتفرم مختلف

- **چالش پلتفرم‌های مختلف** :در دنیای امروز، بدافزارها دیگر محدود به یک نوع سیستم‌عامل یا پلتفرم خاص نیستند. بدافزارها می‌توانند به راحتی از ویندوز به اندروید، iOS، لینوکس، یا حتی سیستم‌های IoT منتقل شوند و برای مقابله با این تهدیدات، الگوریتم‌های یادگیری ماشین باید قادر به شناسایی این بدافزارها در چند پلتفرم مختلف باشند.
- **هدف** :هدف این بخش، توسعه و بهبود روش‌های یادگیری ماشین برای شناسایی و تحلیل ویژگی‌های بدافزار در پلتفرم‌های مختلف) ویندوز، موبایل، ابری و (IoT است. این روش‌ها باید توانایی شناسایی ویژگی‌های عمومی و خاص هر پلتفرم را داشته باشند تا بتوانند بدافزارها را به‌طور مؤثر شناسایی کنند.

۱.۲شناسایی تهدیدات جدید و ناشناخته

- **چالش بدافزارهای جدید** :بدافزارها به طور مداوم در حال تکامل هستند و از تکنیک‌هایی مانند رمزنگاری، اختلاط (Obfuscation) و تغییرات کد برای پنهان شدن از سیستم‌های تشخیص استفاده می‌کنند.
- **هدف** :ایجاد مدل‌هایی که بتوانند با استفاده از داده‌های شبیه‌سازی شده و ویژگی‌های داینامیک و استاتیک، حتی بدافزارهای جدید و ناشناخته را شناسایی کنند. این به معنی شناسایی رفتارهای غیرمعمول و ناهنجاری‌ها در سیستم‌ها و تطبیق مدل‌های یادگیری ماشین برای تشخیص بدافزارهای جدید است.

۱.۳استفاده از یادگیری عمیق برای شناسایی ویژگی‌های پیچیده بدافزار

- **چالش ویژگی‌های پیچیده** :بدافزارهای پیچیده ممکن است ویژگی‌هایی داشته باشند که به راحتی با روش‌های سنتی شناسایی نشوند. استفاده از الگوریتم‌های یادگیری عمیق می‌تواند این مشکل را حل کند.
- **هدف** :استفاده از مدل‌های پیچیده‌ای مانند **شبکه‌های عصبی کانولوشنی (CNN)** و **شبکه‌های عصبی بازگشتی (RNN)** برای استخراج ویژگی‌های پیچیده بدافزار و شناسایی تهدیدات که با روش‌های سنتی قابل شناسایی نیستند.

۱.٤بهبود عملکرد سیستم‌های تشخیص بدافزار با استفاده از ویژگی‌های چندمنظوره

- **چالش عدم تطابق ویژگی‌ها** :ویژگی‌های یک بدافزار در یک پلتفرم ممکن است در پلتفرم دیگری به‌طور متفاوت عمل کنند. برای مثال، ویژگی‌های یک بدافزار در ویندوز ممکن است با ویژگی‌های آن در اندروید تفاوت داشته باشد.
- **هدف** :طراحی سیستم‌هایی که از ویژگی‌های مختلف (مانند ویژگی‌های ایستا و دینامیک) برای تشخیص بدافزار استفاده کنند. این سیستم‌ها باید بتوانند ویژگی‌ها را در پلتفرم‌های مختلف ترکیب کنند تا دقت سیستم‌های تشخیص بدافزار را افزایش دهند.

۲.مقاله دوم «A Study on the Application of Distributed System Technology-Guided Machine Learning in Malware Detection» :

این مقاله بیشتر بر **استفاده از سیستم‌های توزیع‌شده** و **یادگیری ماشین** برای شناسایی بدافزارها در شبکه‌های بزرگ و پیچیده متمرکز است. اهداف این مقاله به شرح زیر هستند:

۲.۱توسعه سیستم‌های توزیع‌شده برای مقیاس‌پذیری و افزایش عملکرد

- **چالش مقیاس‌پذیری** :در محیط‌های بزرگ، مانند شبکه‌های توزیع‌شده و ابری، تشخیص بدافزار نیازمند پردازش و ذخیره‌سازی داده‌های زیادی است. این امر نیازمند سیستم‌های مقیاس‌پذیر است که قادر به پردازش داده‌ها به‌طور همزمان در چندین گره باشند.
- **هدف** :توسعه یک **سیستم توزیع‌شده** برای شناسایی بدافزارها که بتواند به‌طور مؤثر در مقیاس بزرگ داده‌ها را پردازش کند. این سیستم باید به‌طور خودکار داده‌ها را از گره‌های مختلف جمع‌آوری کرده و از آن‌ها برای تشخیص تهدیدات استفاده کند.

۲.۲استفاده از الگوریتم‌های یادگیری ماشین برای شناسایی بدافزارهای ناشناخته

- **چالش بدافزارهای جدید** :یکی از مشکلات اصلی در تشخیص بدافزار، شناسایی انواع جدید آن‌ها است که ویژگی‌های قبلی آن‌ها تغییر کرده باشد.

- **هدف** :استفاده از **الگوریتم‌های یادگیری ماشین** مانند **جنگل تصادفی (Random Forest)** و **PCA**(کاهش ابعاد ویژگی‌ها) برای شناسایی بدافزارهای ناشناخته که تاکنون در پایگاه داده‌ها وجود نداشته‌اند.

۲.۳کاهش پیچیدگی پردازش داده‌ها با استفاده از سیستم‌های توزیع‌شده

- **چالش پردازش حجم بالای داده‌ها** :در سیستم‌های توزیع‌شده، حجم بالای داده‌ها می‌تواند پردازش و تجزیه‌وتحلیل را کند کند.
- **هدف** :طراحی سیستمی که از پردازش موازی استفاده کند و قادر باشد حجم بالای داده‌ها را به‌طور سریع و کارآمد پردازش کند. این سیستم باید قادر به شناسایی سریع بدافزارها و تهدیدات بالقوه باشد.

۲.٤ارزیابی عملکرد سیستم‌های توزیع‌شده در محیط‌های واقعی

- **چالش ارزیابی عملکرد در محیط‌های واقعی** :بسیاری از الگوریتم‌ها ممکن است در محیط‌های آزمایشی عملکرد خوبی داشته باشند، اما در محیط‌های واقعی کارایی پایین‌تری نشان دهند.
- **هدف** :انجام آزمایش‌های عملی با استفاده از دیتاست‌های عمومی مانند Ember ۲۰۱۷ و Ember ۲۰۱۸برای ارزیابی دقت و کارایی سیستم‌های توزیع‌شده. این ارزیابی‌ها باید نشان دهند که سیستم‌های پیشنهادی در محیط‌های واقعی چگونه عمل می‌کنند و آیا قادر به شناسایی بدافزارها با دقت بالا هستند.

۲.۵ارتقاء مدل‌های یادگیری ماشین برای تشخیص سریع‌تر و دقیق‌تر بدافزارها

- **چالش‌های مربوط به سرعت و دقت** :برای مقابله با تهدیدات سریع، سیستم‌ها باید قادر باشند بدافزارها را با سرعت بالا شناسایی کنند.
- **هدف** :استفاده از مدل‌های یادگیری ماشین مانند **جنگل تصادفی** و **مدل‌های مبتنی بر PCA** برای کاهش ابعاد داده‌ها و شناسایی بدافزارها در زمان واقعی با دقت بالا.

جمع‌بندی اهداف

هر دو مقاله با اهداف خاصی به مسئله شناسایی بدافزار پرداخته‌اند. مقاله اول تمرکز بیشتری بر شناسایی بدافزارها در پلتفرم‌های مختلف با استفاده از یادگیری ماشین و الگوریتم‌های پیشرفته دارد. مقاله دوم، هدف خود را بهبود عملکرد سیستم‌های توزیع‌شده برای شناسایی سریع و دقیق بدافزار در مقیاس‌های بزرگ و پیچیده قرار داده است. در نهایت، هر دو مقاله به دنبال توسعه سیستم‌های مقاوم‌تر، مقیاس‌پذیرتر و هوشمندتر برای مقابله با تهدیدات روزافزون بدافزارها هستند.

مفاهیم و موضوعات کلیدی

در این بخش به توضیح دقیق‌تر مفاهیم و موضوعات کلیدی موجود در هر دو مقاله پرداخته می‌شود. این مفاهیم شامل یادگیری ماشین، سیستم‌های توزیع‌شده، تحلیل ویژگی‌ها، کشف ناهنجاری، و استفاده از الگوریتم‌های مختلف یادگیری ماشین برای شناسایی بدافزارها هستند. در اینجا توضیحات فنی و عمیق‌تری برای هر کدام از این مفاهیم آورده می‌شود:

۱. یادگیری ماشین(Machine Learning)

یادگیری ماشین یکی از مهم‌ترین و کلیدی‌ترین مفاهیم در تشخیص بدافزار است. این تکنیک به‌طور خاص برای شناسایی الگوهای پیچیده و شبیه‌سازی رفتارهای بدافزار به‌کار می‌رود. به‌طور کلی، یادگیری ماشین شامل دو دسته الگوریتم است:

۱. **یادگیری نظارت‌شده:(Supervised Learning)**

- در این نوع یادگیری، مدل‌ها با استفاده از داده‌های برچسب‌خورده (مجموعه‌ای از نمونه‌های آلوده و سالم) آموزش داده می‌شوند. سپس، مدل آموزش‌دیده می‌تواند برای شناسایی بدافزارهای ناشناخته استفاده شود.
- الگوریتم‌های معروف در این حوزه شامل **ماشین بردار پشتیبان(SVM)** ، **درخت تصمیم (DT)** و **کدگذاری گراف نایب (KNN)** هستند.

۲. یادگیری بدون نظارت:(Unsupervised Learning)

- این نوع یادگیری برای شناسایی الگوها و گروه‌بندی داده‌های بدون برچسب (یعنی بدون دانستن اینکه داده‌ها آلودگی دارند یا خیر) استفاده می‌شود. این روش برای کشف بدافزارهای جدید و ناشناخته که هیچ برچسبی ندارند، بسیار مفید است.
- **کشف ناهنجاری (Anomaly Detection)** یکی از تکنیک‌های رایج در این زمینه است که به تشخیص رفتارهای غیرمعمول در داده‌ها پرداخته و آن‌ها را به عنوان بدافزار شناسایی می‌کند.

۲.سیستم‌های توزیع‌شده(Distributed Systems)

سیستم‌های توزیع‌شده به‌ویژه در تشخیص بدافزار در محیط‌های مقیاس بزرگ بسیار مهم هستند. این سیستم‌ها با استفاده از معماری‌های توزیع‌شده قادر به انجام پردازش‌های موازی و همکاری گره‌های مختلف در شبکه برای شناسایی تهدیدات هستند. در این روش:

- **معماری توزیع‌شده** :در این معماری، گره‌ها به‌طور مستقل داده‌ها را پردازش کرده و اطلاعات را با یکدیگر به اشتراک می‌گذارند. به‌طور خاص، هر گره می‌تواند بخشی از اطلاعات بدافزار را شناسایی کرده و آن را به گره‌های دیگر ارسال کند. این همکاری بین گره‌ها به سیستم کمک می‌کند تا اطلاعات جدید از تهدیدات را سریع‌تر دریافت کرده و قادر به شناسایی تهدیدات جدید در کل شبکه باشد.
- **پردازش موازی** :پردازش داده‌های بزرگ و پیچیده نیاز به زمان زیاد دارد. استفاده از پردازش موازی در سیستم‌های توزیع‌شده به این معناست که عملیات مختلف به‌طور همزمان در چندین سرور انجام می‌شود، که سرعت پردازش را به‌طور چشمگیری افزایش می‌دهد.

۳.تحلیل ویژگی‌ها(Feature Analysis)

یکی از بخش‌های اساسی در یادگیری ماشین و شناسایی بدافزار، استخراج ویژگی‌ها از داده‌ها است. این ویژگی‌ها به مدل کمک می‌کنند تا تفاوت‌ها و شباهت‌های میان بدافزار و نرم‌افزار سالم را تشخیص دهد. ویژگی‌های مورد استفاده در هر پلتفرم می‌تواند متفاوت باشد و به‌طور کلی به دو دسته تقسیم می‌شود:

۱. ویژگی‌های ایستا:(Static Features)

- ویژگی‌هایی که بدون نیاز به اجرای برنامه‌ها و تنها از طریق تجزیه و تحلیل فایل‌های برنامه استخراج می‌شوند. مانند تجزیه و تحلیل هدرهای فایل‌ها، امضاهای بدافزار، و متا دیتاهای موجود در فایل‌های اجرایی.
- به‌عنوان مثال، در سیستم‌های ویندوز، ویژگی‌هایی مانند اطلاعات هدر فایل‌های **PE** (Portable Executable)استخراج می‌شود.

۲. ویژگی‌های دینامیک:(Dynamic Features)

- ویژگی‌هایی که در حین اجرای برنامه‌ها و در هنگام تعامل برنامه با سیستم به‌طور داینامیک استخراج می‌شوند. این ویژگی‌ها شامل فراخوانی‌هایAPI ، تغییرات در فایل‌ها، یا رفتارهای شبکه‌ای هستند که در هنگام اجرای بدافزار اتفاق می‌افتد.
- به‌طور مثال، در پلتفرم‌های اندروید، ویژگی‌هایی مثل **دسترسی‌ها به مجوزها** و **فراخوانی‌های API** از جمله ویژگی‌های دینامیک به‌شمار می‌روند.

٤.کشف ناهنجاری(Anomaly Detection)

کشف ناهنجاری یک تکنیک کلیدی در یادگیری بدون نظارت است که برای شناسایی رفتارهای غیرمعمول یا غیرقابل پیش‌بینی در داده‌ها استفاده می‌شود. این تکنیک در برابر بدافزارهایی که تا به‌حال شناسایی نشده‌اند، به‌ویژه بدافزارهای نوظهور، بسیار مؤثر است.

- **روش‌های آماری** :در این روش‌ها، ناهنجاری‌ها با استفاده از تجزیه و تحلیل آماری از داده‌ها شناسایی می‌شوند. برای مثال، اگر رفتار سیستم در مقایسه با رفتارهای نرمال شبکه یا سیستم به‌شدت متفاوت باشد، به‌عنوان یک ناهنجاری شناسایی می‌شود.
- **روش‌های مبتنی بر مدل** :در این رویکرد، مدل‌هایی برای پیش‌بینی رفتارهای نرمال ایجاد می‌شود و هر گونه انحراف از این پیش‌بینی‌ها به‌عنوان ناهنجاری در نظر گرفته می‌شود.

٥.الگوریتم‌های یادگیری ماشین برای شناسایی بدافزار

در این مقاله‌ها، الگوریتم‌های مختلفی برای تشخیص بدافزار مورد استفاده قرار گرفته‌اند. این الگوریتم‌ها شامل مدل‌های کلاسیک یادگیری ماشین و مدل‌های یادگیری عمیق هستند.

- **الگوریتم‌های کلاسیک:**
  - **ماشین بردار پشتیبان (SVM):** یکی از الگوریتم‌های محبوب برای طبقه‌بندی داده‌ها است که با استفاده از ابرصفحه‌ها (Hyperplanes) داده‌ها را به دو دسته تقسیم می‌کند.
  - **درخت تصمیم (DT):** الگوریتمی است که با استفاده از ساختار درختی، داده‌ها را به‌طور بازگشتی تقسیم کرده و بهترین تصمیم را می‌گیرد.
- **الگوریتم‌های یادگیری عمیق:**
  - **شبکه‌های عصبی کانولوشنی (CNN):** برای تجزیه و تحلیل داده‌های تصویری و باینری بسیار مؤثر هستند. این شبکه‌ها می‌توانند ویژگی‌های پیچیده را از داده‌ها استخراج کنند و برای شناسایی بدافزارهای پیچیده مناسب هستند.
  - **شبکه‌های عصبی بازگشتی (RNN):** این مدل‌ها برای تحلیل داده‌های دنباله‌ای مانند فراخوانی‌های API و کدهای اجرایی به‌کار می‌روند.

٦.الگوریتم جنگل تصادفی(Random Forest)

جنگل تصادفی (Random Forest) یک الگوریتم یادگیری ماشین است که از ترکیب چندین درخت تصمیم برای انجام پیش‌بینی‌ها استفاده می‌کند. این الگوریتم به‌ویژه در محیط‌های توزیع‌شده و در مواقعی که داده‌های پیچیده و متنوع در دسترس هستند، بسیار مفید است. از ویژگی‌های اصلی آن می‌توان به موارد زیر اشاره کرد:

- **استحکام بالا در برابر داده‌های ناهنجار** :جنگل تصادفی از طریق ایجاد چندین درخت تصمیم مختلف از داده‌های مختلف، می‌تواند خطاهای فردی را کاهش دهد و دقت بالاتری ارائه دهد.
- **مقیاس‌پذیری بالا** :جنگل تصادفی قابلیت پردازش داده‌های بزرگ را در مقیاس‌های مختلف دارد و می‌تواند در محیط‌های توزیع‌شده اجرا شود.

کارهای انجام‌شده

در این بخش، به تفصیل به کارهایی که در هر دو مقاله انجام شده است پرداخته می‌شود و تکنیک‌ها و رویکردهای فنی استفاده‌شده به‌طور دقیق‌تر شرح داده می‌شود.

۱.مقاله اول **A Survey on ML Techniques for Multi-Platform Malware Detection**:

مقاله اول، به طور جامع به بررسی استفاده از تکنیک‌های یادگیری ماشین در تشخیص بدافزار در پلتفرم‌های مختلفPC) ، موبایل، IoTو ابری (می‌پردازد. این مقاله از روش‌های یادگیری ماشین برای شناسایی بدافزارهایی استفاده می‌کند که قادر به عبور از روش‌های سنتی و امضا-محور تشخیص هستند. در این مقاله:

۱.۱**تحلیل ویژگی‌ها**:

- برای هر پلتفرم) ویندوز، اندروید، iOS، لینوکس، IoTو محیط‌های ابری(، ویژگی‌های خاصی از بدافزار استخراج می‌شود. این ویژگی‌ها شامل داده‌های ایستا (مانند امضای فایل‌ها) و دینامیک) مانند الگوهای فراخوانی API یا تغییرات در شبکه (است.
- ویژگی‌های مختلف برای تشخیص بدافزار در این سیستم‌ها استخراج و تحلیل می‌شوند. برای مثال، در پلتفرم‌های ویندوز و لینوکس، ویژگی‌های فایل‌های اجرایی) مثل بخش‌های PE در ویندوز و ELFدر لینوکس (به دقت بررسی می‌شود. در پلتفرم‌های موبایل، ویژگی‌های فایل‌های APK و IPA، مجوزها، فراخوانی‌های API و رفتارهای شبکه‌ای مورد توجه قرار می‌گیرند.

۱.۲**استفاده از الگوریتم‌های یادگیری ماشین**:

- در این مقاله، انواع الگوریتم‌های یادگیری ماشین بررسی شده است. این الگوریتم‌ها به دو دسته اصلی تقسیم می‌شوند:
  - **الگوریتم‌های کلاسیک یادگیری ماشین** :مانند **ماشین بردار پشتیبان(SVM)** ، **K-نزدیک‌ترین همسایه‌ها(KNN)** ، **درخت تصمیم (DT)** و **ناوی بیز.(NB)**

- **روش‌های یادگیری عمیق** :از مدل‌هایی مانند **شبکه‌های عصبی کانولوشنی (CNN)**و **شبکه‌های عصبی بازگشتی (RNN)** برای شناسایی ویژگی‌های پیچیده‌تر و دنباله‌ای استفاده می‌شود. به‌ویژه CNN در پردازش ویژگی‌های تصویری از فایل‌های اجرایی یا داده‌های باینری بسیار مؤثر است.

۱.۳چالش‌ها و مسائل پلتفرم‌ها:

- در این مقاله همچنین چالش‌های هر پلتفرم نیز بررسی شده است. مثلاً، برای سیستم‌های IoT که معمولاً از منابع محدود برخوردارند، تشخیص بدافزار باید به گونه‌ای طراحی شود که مصرف منابع کم باشد و بتواند به سرعت رفتارهای مشکوک را شناسایی کند.
- در محیط‌های ابری که معمولاً دارای مقیاس بزرگ و چندگانه هستند، تشخیص بدافزار باید بتواند در میان تعداد زیادی از ماشین‌های مجازی و داده‌های متنوع به‌طور کارآمد عمل کند.

۲.مقاله دوم «A Study on the Application of Distributed System Technology-Guided Machine Learning in Malware Detection» :

این مقاله رویکردی جدید برای تشخیص بدافزار در سیستم‌های توزیع‌شده با استفاده از یادگیری ماشین پیشنهاد می‌کند. رویکرد اصلی این مقاله استفاده از معماری توزیع‌شده برای گسترش قابلیت شناسایی بدافزار در شبکه‌های بزرگ است.

۲.۱استفاده از سیستم‌های توزیع‌شده:

- در این مقاله، یک چارچوب توزیع‌شده طراحی شده است که در آن هر **گره (Node)** در سیستم توزیع‌شده برای تجزیه و تحلیل داده‌ها و شناسایی بدافزار فعالیت می‌کند. این گره‌ها در نقاط مختلف شبکه قرار دارند و به‌طور مستقل آنالیزهایی را روی داده‌ها انجام می‌دهند.
- **سیستم کنترل تحلیلی** :این بخش از سیستم مسئول هماهنگی و جمع‌آوری نتایج از گره‌های مختلف است. داده‌های شناسایی شده از هر گره به سیستم کنترل مرکزی ارسال می‌شود تا بررسی‌های بیشتر و تجزیه و تحلیل دقیق‌تری انجام شود.

- **عملکرد توزیع‌شده** :این مقاله به ویژه از پروتکل‌هایی مانند **Hadoop**و **Spark**برای پردازش داده‌های بزرگ و انجام تحلیل‌های موازی در سیستم توزیع‌شده استفاده می‌کند.

۲.۲**الگوریتم‌های یادگیری ماشین**:

- برای تشخیص بدافزار، این مقاله از الگوریتم‌های مختلف یادگیری ماشین، از جمله **جنگل تصادفی (Random Forest)**، استفاده می‌کند. جنگل تصادفی به‌ویژه در این مقاله به دلیل توانایی در مدیریت ویژگی‌های پیچیده و داشتن مقاومت بالا در برابر داده‌های متنوع و ناهنجاری‌ها انتخاب شده است.
- در این روش، داده‌ها ابتدا از طریق الگوریتم **PCA (Principal Component Analysis)** کاهش ابعاد داده‌ها برای استخراج ویژگی‌های اصلی و سپس برای طبقه‌بندی از الگوریتم **جنگل تصادفی** استفاده می‌شود.

۲.۳**تحلیل ویژگی‌ها**:

- در این مقاله، از ویژگی‌های مختلف مانند **فایل‌های PE** برای شناسایی بدافزار استفاده می‌شود. فایل‌های PE به دلیل ساختار خاص خود در ویندوز، می‌توانند اطلاعات مفیدی در مورد ویژگی‌های اجرایی بدافزارها ارائه دهند.
- همچنین، برای شناسایی بدافزارهای ناشناخته از تکنیک‌های کشف ناهنجاری (Anomaly Detection)استفاده می‌شود. در این بخش، رفتارهای غیرمعمول در شبکه یا تغییرات در الگوهای اجرایی به عنوان سیگنال‌هایی برای شناسایی بدافزار در نظر گرفته می‌شود.

۲.۴**ارزیابی عملکرد**:

- برای ارزیابی عملکرد روش‌های پیشنهادی، از **دیتاست‌های عمومی** مانند Ember ۲۰۱۷ و Ember ۲۰۱۸استفاده شده است. این دیتاست‌ها شامل داده‌های واقعی از بدافزارها و نرم‌افزارهای سالم است که می‌توانند به‌عنوان معیار برای ارزیابی دقت مدل‌های یادگیری ماشین به کار روند.

- نتایج نشان می‌دهند که روش‌های پیشنهادی نسبت به الگوریتم‌های سنتی مانند **SVM**و **K-means**عملکرد بهتری دارند. در برخی آزمایش‌ها، دقت مدل‌های توزیع‌شده بیش از ۹۹٪ بوده است.

۲.۵پیشرفت‌های سیستم توزیع‌شده:

- این مقاله همچنین به نحوه بهبود عملکرد سیستم توزیع‌شده با استفاده از پردازش موازی و ذخیره‌سازی توزیع‌شده می‌پردازد. استفاده از **HDFS**و **Hive**برای ذخیره‌سازی داده‌ها و **Spark**برای پردازش آن‌ها در این سیستم موجب افزایش کارایی و کاهش زمان پاسخ‌دهی شده است.

نتیجه‌گیری

در مجموع، هر دو مقاله رویکردهای پیشرفته‌ای برای استفاده از یادگیری ماشین در تشخیص بدافزار در سیستم‌های مختلف و مقیاس‌های بزرگ ارائه می‌دهند. مقاله اول تمرکز بیشتری بر شناسایی بدافزارها در پلتفرم‌های مختلف با استفاده از یادگیری ماشین و الگوریتم‌های پیشرفته دارد. مقاله دوم، هدف خود را بهبود عملکرد سیستم‌های توزیع‌شده برای شناسایی سریع و دقیق بدافزار در مقیاس‌های بزرگ و پیچیده قرار داده است. در نهایت، هر دو مقاله به دنبال توسعه سیستم‌های مقاوم‌تر، مقیاس‌پذیرتر و هوشمندتر برای مقابله با تهدیدات روزافزون بدافزارها هستند.