نشریه علمی «علوم و فناوری‌های پدافند نوین»
سال سیزدهم، شماره ۲، تابستان ۱۴۰۱؛ ص ۹۹- ۸۹

علمی- پژوهشی

تشخیص نفوذ در شبکه‌های رایانه‌ای با استفاده از انتخاب ویژگی ترکیبی مؤثر مبتنی بر روش اشتراک‌گیری اطلاعات متقابل، آزمون F تحلیل واریانس و الگوریتم ژنتیک

حمید بیگدلی[۱]*، جلیل مظلوم[۲]

۱- استادیار، دانشگاه فرماندهی و ستاد آجا، ۲- دانشیار، دانشگاه علوم و فنون هوایی شهید ستاری

(دریافت: ۰۲/ ۱۴۰۰/۱۲، پذیرش: ۱۴۰۱/۰۵/۱۶)

چکیده

سامانه تشخیص نفوذ (IDS) حجم عظیمی از داده‌ها را مدیریت می‌کند که شامل ویژگی‌های نامرتبط و زائد است که منجر به مصرف منابع قابل توجه، روندهای آموزش و آزمایش طولانی مدت و نرخ تشخیص پایین می‌شود. از این رو، انتخاب ویژگی یک گام مهم در تشخیص نفوذ در نظر گرفته شده است. هدف این پژوهش، معرفی یک راهبرد مبتنی بر اشتراک است که به‌طور بهینه ویژگی‌ها را برای طبقه‌بندی انتخاب می‌کند. این انتخاب ویژگی شامل اشتراک‌گیری از روش‌های اطلاعات متقابل بر اساس مدل انتقال (MIT-MIT)، آزمون F تحلیل واریانس و الگوریتم ژنتیک (GA) است. یک مجموعه داده معیار، به نام NSL-KDD، برای ارزیابی اثربخشی رویکرد پیشنهادی استفاده می‌شود. این مطالعه شامل صحت، دقت، یادآوری و امتیاز F1 به‌عنوان معیارهای ارزیابی برای IDS است که روش پیشنهادی را با طبقه‌بندی کننده‌های پیشرفته تحلیل می‌کند. نتایج ارزیابی تأیید کرده است که الگوریتم انتخاب ویژگی ما ویژگی‌های ضروری‌تری را برای IDS جهت دستیابی به دقت بالا فراهم می‌نماید و از سایر الگوریتم‌های مقایسه‌ای برتری می‌جوید.

کلیدواژه‌ها: سامانه تشخیص نفوذ، انتخاب ویژگی، اطلاعات متقابل، آزمون F تحلیل واریانس، الگوریتم ژنتیک

Network Intrusion Detection in Computer Networks Using an Efficacious Combined Feature Selection Technique Based on the Intersection Method of Mutual Information, Anova F-Test and Genetic Algorithm

H. Bigdeli*, J. Mazloum,
Shahid Sattari Aeronautical University of Science and Technology
(Received: 21/ 02/2022; Accepted: 07/08/2022)

Abstract

The intrusion detection system (IDS) manages a massive volume of data that comprises irrelevant and redundant features, leading to more significant resource consumption, long-time training and testing procedures, and low detection rate. Hence, feature selection is a crucial phase in intrusion detection. The aim of this paper is to introduce an intersection-based strategy that optimally selects the features for classification. This feature selection involves an intersection of simultaneous mutual information based on the transductive model (MIT-MIT), Anova F-test, and genetic algorithm (GA) methods. A benchmark dataset, named NSL-KDD, is applied to evaluate the effectiveness of the proposed approach. This study includes accuracy, precision, recall, and F1 score as the evaluation metrics for IDS, which analyzes the proposed method with state-of-the-art classifiers. The evaluation results confirm that our feature selection algorithm provides more essential features for IDS to achieve high accuracy, outperforming other comparative algorithms.

Keywords: Intrusion Detection System; Feature Selection; Mutual Information; Anova F-Test; Genetic Algorithm

*Corresponding Author E-mail: H.Bigdeli@casu.ac.ir

۱. مقدمه

اطلاعات مهم همیشه برای نفوذگران از جذابیت بالایی برخوردار بوده است و از این رو نسبت به نفوذهای شبکه حساس است. نفوذ، به فرآیندی گفته می‌شود که در آن نفوذگر به سرور یا سامانه جهت دانلود، تغییر یا آسیب رساندن به هر داده ضروری یا مخفی دسترسی پیدا می‌کند. آسیب‌پذیری‌های موجود سامانه، مانند استفاده نادرست کاربر، نقص برنامه، یا پیکربندی اشتباه، می‌تواند باعث نفوذ شود. در نتیجه، طرح‌های تشخیص نفوذ هوشمند با هدف محافظت از سامانه شبکه مورد نیاز است.

سامانه‌های تشخیص نفوذ (IDS) کارآمد، معمولاً از طریق استفاده از روش‌های داده‌کاوی به‌طور ماهرانه‌ای پیشرفت می‌کنند. با این وجود، اجرای چنین سامانه‌هایی به‌طور کلی پیچیده به نظر می‌رسد. علاوه بر این، روند یادگیری سامانه به مقادیر زیادی از داده‌های استاندارد و مطالعه شده نیاز دارد که در عین معتبر بودن حاوی انواع حملات جهت تشخیص و مقایسه با سایر پژوهش‌ها باشد. در این راستا، مجموعه داده‌های مختلفی در سال‌های اخیر معرفی شده‌اند که یکی از معتبرترین و گسترده‌ترین آن‌ها، KDD CUP 99[1] نام دارد. این مجموعه داده گسترده شامل بیش از پنج میلیون ثبت و ۴۱ ویژگی است. آخرین و قابل اعتمادترین نسخه سنتی KDD CUP 99، مجموعه‌ای به نام NSL-KDD[2] است که در این کار مورد استفاده قرار گرفته است.

از طرف دیگر، پردازش "داده‌های بزرگ"، کل فرآیند تشخیص از جمله ساخت و بررسی یک طبقه‌بندی کننده را به تأخیر می‌اندازد؛ یا باعث می‌شود طبقه‌بندی کننده به دلیل حافظه ناکافی قادر به عملکرد درست نباشد که منجر به دقت پایین طبقه‌بندی به دلیل مشکلات محاسباتی در مدیریت حجم انبوه داده می‌گردد.

علاوه بر این، مجموعه‌های داده در مقیاس بزرگ دارای ویژگی‌های اضافی، غیر مفید یا نویز هستند که چالش‌های مهم تجزیه و تحلیل اطلاعات و مدل‌سازی داده‌ها را فراهم می‌نمایند. یک گام پیش‌پردازش به‌عنوان کاهش ابعاد برای رسیدگی به مشکلات ذکر شده معرفی شده است که انتخاب ویژگی نامیده می‌شود، به‌طوری که یک مرحله اساسی در IDS، انتخاب ویژگی‌های مربوطه و حذف ویژگی‌های نامربوط و زائد است. انتخاب ویژگی منجر به اصلاح داده‌ها، درک داده‌ها و فضای ذخیره‌سازی مورد نیاز و محدودیت هزینه می‌گردد.

با توجه به مزایای متعدد الگوریتم ژنتیک (GA) از قبیل درک آسان، کار با جمعیت به جای یک نقطه، سهولت استفاده در مسائل موازی‌سازی و کار با قوانین احتمال، به‌عنوان یکی از مؤثرترین و جامع‌ترین الگوریتم‌های بهینه‌سازی جهت انتخاب ویژگی مطرح شده است. عملکرد الگوریتم‌های طبقه‌بندی در کنار الگوریتم‌های تکاملی مانند GA در سال‌های اخیر امیدوار کننده بوده است [۱ و ۲]. به همین ترتیب، یکی دیگر از معیارهای کاربردی برای انتخاب ویژگی‌ها، روش اطلاعات متقابل[3] (MI) است که بر اساس وابستگی بین دو ویژگی عمل می‌کند [۳ و ۴]. همچنین، راهبرد آزمون F تحلیل واریانس در مرحله پیش‌پردازش تحقیقات اخیر به‌طور گسترده‌ای یافت شده است [۵ و ۶].

الگوریتم‌های یادگیری ماشین[4] (ML) و یادگیری عمیق[5] (DL) اخیراً در تحقیقات مربوط به توسعه IDS به‌کار گرفته شده‌اند. با توجه به ماهیت داده‌محور این روش‌ها، آن‌ها در شناسایی حملات ناشناخته بسیار موفق عمل کرده‌اند. در میان روش‌های طبقه‌بندی، درخت تصمیم[6] (DT) [۷ و ۸]، درخت مازاد[7] [۹ و ۱۰]، تقویت گرادیان[8] [۱۱ و ۱۲]، جنگل تصادفی[9] (RF) [۷ و ۱۲]، بیز ساده[10] (NB) [۷ و ۱۳]، ماشین بردار پشتیبان[11] (SVM) [۷ و ۱۴]، و پرسپترون چندلایه[12] (MLP) [۱۲ و ۱۵]، نتایج رضایت‌بخشی را برای داده‌های شبکه ارائه کرده‌اند. این روش‌ها جهت ارزیابی طبقه‌بندی برگزیده شده‌اند. با اضافه نمودن گام انتخاب ویژگی به الگوریتم‌های طبقه‌بندی در حوزه IDS، افزایش دقت و کاهش زمان یادگیری با حجم داده کمتری مشاهده شده است [۱۶ و ۱۷].

مشارکت‌های کلیدی این پژوهش به شرح زیر ارائه می‌گردد:

- یک بلوک انتخاب ویژگی ادغام شده جدید که دربردارنده الگوریتم‌های اطلاعات متقابل مبنی بر مدل انتقال (MIT-MIT)، آزمون F تحلیل واریانس و GA است به‌طوری که ویژگی‌های ورودی را به‌صورت موازی بهینه‌سازی کرده و از طریق یک واحد اشتراک‌گیری فیلتر می‌نمایند.
- یک مرحله ارزیابی که انتخاب ویژگی پیشنهادی را با طبقه‌بندی کننده‌های DT، درخت مازاد، تقویت گرادیان، RF، NB، SVM و MLP ترکیب می‌کند تا عملکرد و امکان‌سنجی روش پیشنهادی را از نظر معیارهای ارزیابی تأیید نماید.
- مجموعه داده NSL-KDD برای مطالعه رویکرد پیشنهادی مورد استفاده قرار می‌گیرد.

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99
[2] http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDDdataset

[3] Mutual Information
[4] Machine Learning
[5] Deep Learning
[6] Decision Tree
[7] Extra Tree
[8] Gradient Boosting
[9] Random Forest
[10] Naïve Bayes
[11] Support Vector Machine
[12] Multi-layer Perceptron

۲. مبانی نظری و پیشینه پژوهش

۲-۱. مبانی نظری

با توجه به اهمیت روش انتخاب ویژگی در IDS، راهبردهای MIT-MIT، آزمون F تحلیل واریانس و GA در این بخش به‌عنوان مبنای روش انتخاب ویژگی پیشنهادی معرفی می‌گردند.

- الگوریتم MIT-MIT

روش پیشنهادی MIT-MIT به‌عنوان یک نسخه اصلاح شده از روش اطلاعات متقابل (MI) فرض شده است که ارتباط بین دو متغیر تصادفی را محاسبه می‌کند، به‌صورتی که [۱۸]:

$$I(X,Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (۱)$$

که در آن، $p(x)$ و $p(y)$ توابع احتمال حاشیه‌ای را نشان می‌دهند و $p(x,y)$ بیانگر احتمال مشترک است. در اینجا، معیار ارتباط (بیشینه سازی) و معیار افزونگی (کمینه سازی) مستقل فرض می‌شوند. در روند ارتباط، MI بین کاندیدها و مقدار کلاس ارزیابی می‌شود. با این حال، نمی‌تواند داده‌های آزمایشی بدون برچسب اخیراً اضافه شده را استخراج کند. از این رو نمی‌تواند نسخه انتقالی مدل را توسعه دهد. علاوه بر این، روند افزونگی، MI همه کاندیدهای انتخاب شده را بدون توجه به مقدار کلاس درنظر می‌گیرد. بنابراین، داده‌های آزمایش بدون برچسب بر اساس مدل انتقال تعیین می‌گردند. در کل، تابع هدف در این روش را می‌توان به‌صورت زیر توصیف کرد:

$$\underset{x_j\in\left|X-S_{m-1}\right.}{Sup}\left[I(x_j^t;c^t)-\frac{1}{m-1}\sum_{x_j\in S_{m-1}}I(x_j^{tt};x_i^{tt})\right] \qquad (۲)$$

در این رابطه، x_j^t نشان دهنده ستون ویژگی j ام مربوط به داده‌های آموزشی است. به همین ترتیب، x_j^{tt} به داده‌های آموزش و آزمایش اختصاص داده می‌شود. همچنین، مقدار کلاس داده‌های آموزشی با c^t تعریف می‌گردد.

برای جلوگیری از به‌دست آوردن خطای قابل توجه ناشی از گرد کردن مقادیر بسیار نزدیک فنوتیپ‌ها، محاسبه z-score در هر نمونه از طریق $\frac{x-\mu}{\sigma^2}$ درنظر گرفته می‌شود که منجر به مقدار گسسته زیر خواهد شد:

$$descretized-value=\begin{cases}-1, & if \quad z-score<-1\\ 1, & if \quad z-score>1\\ 0, & otherwise\end{cases} \qquad (۳)$$

علاوه بر این، خطای سامانمند برای دو بردار به‌صورت زیر قابل تعریف است:

$$\Delta I(X,Y)\approx\frac{M_{xy}-M_x-M_y+1}{2N} \qquad (۴)$$

که در آن، M_x، M_y و M_{xy} ستون های هیستوگرام اعمال شده را برای بردارهای X، Y و هر دو بردار به ترتیب ارائه می کنند. همچنین، N تعداد نمونه را مشخص می‌کند. به‌منظور محاسبه کارآمد افزونگی جدید برای ارزیابی MI در کاندید انتخابی، نیازی به تخمین x_j در $\sum_{x_i\in S_{m-1}}I(x_j;x_i)$ نیست. علاوه بر این، $\sum_{x_i\in S_{m-2}}I(x_j;x_i)$ و $\sum_{x_i\in S_{m-1}}I(x_j;x_i)$ فقط در $I(x_j;x_{m-1})$ تفاوت دارند. در نتیجه، $\sum_{x_i\in S_{m-1}}I(x_j;x_i)$ می‌تواند در هر مرحله حفظ شده و در مرحله بعد دوباره استفاده گردد.

- آزمون F تحلیل واریانس

آزمون تحلیل واریانس برای مقایسه متغیرهای "میانگین چندگانه" مجموعه داده و تجزیه و تحلیل اینکه آیا تفاوت اساسی بین میانگین متغیرهای کلاس‌های متعدد وجود دارد یا خیر، استفاده می‌گردد. آمار تحلیل واریانس، یعنی آزمون F، از طریق مراحل زیر قابل محاسبه است [۵]:

انحراف بین کلاس‌ها به‌صورت زیر تعریف می‌گردد:

$$BSS=\sum_{i=1}^{N}n_i\left(\bar{x}_i-\bar{x}\right)^2 \qquad (۵)$$

$$BMS=BSS/df \qquad (۶)$$

به‌طوری که BSS و BMS به ترتیب بیانگر مجموع مربعات بین کلاس‌ها و میانگین مربعات آن‌ها است.

انحراف در کلاس‌ها به‌صورت زیر تعیین می‌شود:

$$WSS=\sum_{i=1}^{k}\sum_{j=1}^{N}(n_j-1)\sigma_i^2 \qquad (۷)$$

$$WMS=WSS/df_\omega \qquad (۸)$$

از طرف دیگر، WSS و WMS بیانگر مجموع مربعات درون کلاس‌ها و میانگین مربعات آن‌ها هستند. همچنین، N و k به ترتیب تعداد نمونه‌ها و کلاس‌ها را مشخص می‌کنند و df درجه آزادی، σ انحراف معیار و $df_\omega=(N-k)$ است.

آمار مقدار F خواهد بود:

$$F-value=BMS/WMS \qquad (۹)$$

• الگوریتم ژنتیک (GA)

GA یک الگوریتم تصادفی برای بهینه‌سازی مسئله با توجه به ژنتیک بیولوژیکی و تکامل طبیعی است. به‌طور طبیعی، ژن‌های موجودات زنده در طول نسل‌های متوالی توسعه می‌یابند تا به‌طور رضایت‌بخشی با محیط سازگار شوند. GA روی گروهی از نمونه‌ها کار می‌کند تا تخمین بهتر و بهتری حاصل نماید. با توجه به مقدار برازش در هر نسل از الگوریتم، جمعیت جدیدی ایجاد می‌گردد. در ادامه روند این الگوریتم به‌صورت مرحله به مرحله توضیح داده شده است [۱۹].

مقداردهی اولیه- در مرحله اول، افراد در جمعیت تولید و مقداردهی اولیه می‌شوند. با توجه به روش تصادفی GA، مقداردهی اولیه ژن‌های افراد به‌صورت تصادفی صورت می‌گیرد.

تخصیص برازش- در ادامه، ضروری است که یک تابع برازش جهت ارزیابی هر فرد در جمعیت تعریف گردد. خطای قابل توجه انتخاب نشان دهنده برازش پایین است. افرادی که برازش بالاتری دارند، احتمال بیشتری برای انتخاب ترکیب مجدد را دارا می‌باشند.

انتخاب- پس از آن، گام انتخاب، نیمی از جمعیت را برای ترکیب مجدد نسل بعدی برمی‌گزینند که بر اساس سطح برازش آن‌ها تمایل به زنده ماندن دارند.

عملگر آمیزش- در این مرحله، جمعیت جدیدی از طریق ترکیب مجدد افراد منتخب تولید می‌شود که در آن چهار فرزند به‌طور تصادفی با ترکیب ویژگی‌های دو والد ایجاد می‌گردند. راهبرد آمیزش تعیین می‌کند که آیا هر یک از ویژگی‌های فرزندان از یک والد نشئت می‌گیرد یا دیگری.

عملگر جهش- در گام بعدی، به دلیل تنوع کم عملگر آمیزش که منجر به شباهت فرزندان به والدین شده است، عملگر جهش سعی می‌کند مقادیر برخی از ویژگی‌ها را به‌طور تصادفی تغییر دهد. از این رو، یک عدد تصادفی بین ۰ و ۱ ایجاد می‌شود به‌طوری که اگر کمتر از نرخ جهش باشد، یک ویژگی جهش یافته و متغیر برگردانده می‌شود. نرخ جهش درنظر گرفته می‌شود که نشان دهنده تعداد ویژگی‌ها است.

فرآیند و نتایج- تمام فرآیند فوق تا زمانی که یک معیار توقف برآورده گردد، تکرار خواهد شد. درنهایت، مناسب‌ترین افراد از جمعیت بازگردانده می‌شوند.

۲-۲. پیشینه پژوهش

در طول دهه‌های اخیر، مطالعات متعددی بر افزایش دقت طبقه‌بندی در حوزه IDS متمرکز شده‌اند. در مراحل اولیه، برخی از مدل‌های ML معرفی شدند که در طول زمان به‌عنوان شبکه‌های عصبی عمیق[1] (DNN) توسعه یافتند. با این وجود، به دلیل حجم زیاد مجموعه داده ورودی، از جمله داده‌های تکراری و نویز، اصلاح ورودی مدل به‌عنوان یک مرحله ضروری با استفاده از روش انتخاب ویژگی پیاده‌سازی شده است. در ادامه، پیشینه مربوط به مباحث انتخاب ویژگی و طبقه‌بندی بررسی شده است.

به‌منظور تسریع یادگیری و افزایش کیفیت مفهوم، فرآیند انتخاب ویژگی، مرتبط‌ترین ویژگی‌ها به هدف را در مسائل یادگیری استخراج می‌کند. برای مثال، رویکردهای موازی مبتنی بر GA برای انتخاب ویژگی، با اتخاذ یک کتابخانه MapReduce منبع‌باز (Hadoop) پیاده‌سازی شدند که در آن، روش‌های ML مانند SVM، ANN[2]، RT[3]، رگرسیون لجستیک[4] و NB به همراه مجموعه داده NSL-KDD، در اجرا به‌کار گرفته شدند [۱]. یک ساختار ترکیبی دو مرحله‌ای پیشنهاد شد، به‌طوری که GA ویژگی‌های مناسب را با هدف بهبود دقت انتخاب می‌کند. در مرحله بعد، الگوریتم‌های ML، متشکل از SVM، طبقه‌بندی کننده ensemble و DT، بر روی یک پایگاه داده چند کلاسه NSL-KDD پیاده‌سازی شدند [۲۰]. یک طرح تشخیص نفوذ ارائه گردید که زیرمجموعه ویژگی‌ها را با MI، کای دو[5]، آزمون F تحلیل واریانس و RF سریع‌تر به‌عنوان انتخاب‌گر نهایی و بهینه‌ساز پارامترها برگزید [۲۱]. علاوه بر این، SAE[6]، CNN[7] و RF سه طبقه‌بندی کننده هستند که به‌طور موازی به‌عنوان طبقه‌بندی ensemble اتخاذ شده‌اند. برای تحلیل‌های مقایسه‌ای، آن‌ها از NSL-KDD کلاسیک و جدیدترین مجموعه داده CICIDS2018 استفاده کردند [۲۲]. به یک روش تشخیص ناهنجاری مبتنی بر جریان در کنترل کننده OpenFlow با استفاده از روش ترکیبی GRU-LSTM[8] دست یافت که در آن یک راهبرد انتخاب ویژگی مبنی بر آزمون F تحلیل واریانس و FE[9]R کاربردی به نام ANOVA F-RFE، با آزمایش بر روی معیار NSL-KDD مورد مطالعه قرار گرفته بود. از سه روش مختلف انتخاب ویژگی از قبیل MI، الگوریتم اطلاعات متقابل کرم شب‌تاب[10] (MIFA) با C4.5 به‌عنوان ارزیاب، و MIFA با یک شبکه

[1] Deep Neural Networks
[2] Artificial Neural Network
[3] Random Tree
[4] Logistic Regression
[5] Chi-squared
[6] Simple Autoencoder Ensemble
[7] Convolutional Neural Network
[8] Gated Recurrent Unit Long Short-term Memory
[9] Recursive Feature Elimination
[10] Mutual Information Firefly Algorithm

بیز[1] (BN)، به‌عنوان ارزیاب استفاده گردید و سپس رأی‌گیری (انتخاب یک ویژگی از سه زیرمجموعه، اگر حداقل در دو مجموعه وجود داشته باشد) صورت گرفت. ویژگی‌های به‌دست آمده به طبقه‌بندی‌کننده‌های C4.5 و شبکه‌های بیز، با درنظر گرفتن KDD CUP 99 به‌عنوان یک مجموعه‌داده، وارد شدند [۲۳]. یک انتخاب ویژگی مبتنی بر اطلاعات متقابل چند متغیره[2] (MVMIFS) به همراه ماشین بردار پشتیبان حداقل مربعات[3] (LSSVM) برای طبقه‌بندی داده‌های ترافیک را پیشنهاد شده است که در سه مجموعه داده برجسته KDD Cup 99، NSL-KDD، و Kyoto 2006 + مورد تأیید قرار گرفت [۲۴].

انواع متعددی از تحقیقات پیشین نشان داده‌اند که فرآیند IDS با راهبردهای ML و DNN می‌تواند به نتایج با دقت بالایی دست یابد. به‌عنوان مثال، مدل‌های طبقه‌بندی نظارت شده متفاوتی با استفاده از طبقه‌بندی کننده‌های RF، SVM، DT، LG[4]، BM درخت مازاد، تقویت گرادیان، AdaBoost، NN[5]K، MLP، بیز ساده گوسی و رگرسیون لجستیک ساخته شدند که روی مجموعه داده NSL-KDD مورد بررسی قرار گرفتند [۲۵]. DNN و دو روش ensemble به نام‌های RF و GBT[6]، جهت طبقه‌بندی مجموعه داده‌های ترافیک شبکه استفاده کرد. یک معیار همگنی برای ارزیابی ویژگی‌ها اتخاذ شد. همچنین، UNSW NB15 و CICIDS2017، به‌عنوان مجموعه داده‌های اخیراً منتشر شده، برای تأیید روش پیشنهادی مورد استفاده قرار گرفتند [۲۶]. روش‌های RF، SVM، رگرسیون لجستیک، NB و GBT پیاده‌سازی شده‌اند. همچنین، یک مدل MLP عمیق برای مقایسه روش‌های کلاسیک ML با رویکرد DL مورد تجزیه و تحلیل قرار گرفت. پس از آن، نتایج نشان داد که مدل DL به صحت، دقت و یادآوری امیدوار کننده‌ای دست یافته است، در حالی که بررسی داده‌ها زمان‌برتر خواهد بود [۲۷]. چندین مدل برای طبقه‌بندی و شناسایی ترافیک شبکه‌های خصوصی مجازی[7] (VPN) درنظر گرفته شد. به‌طور دقیق‌تر، ترکیب DT و تقویت گرادیان به روش bagging به‌عنوان طبقه‌بندی کننده ensemble مورد آزمایش قرار گرفت که نتایج تأیید شده‌ای را به‌دست آورد زیرا عملکرد بهتری از طبقه‌بندی کننده‌های منفرد مانند KNN، MLP و DT را ارائه نمود [۲۸]. هفت مدل ML، از جمله AdaBoost، MLP، DT، NB، KNN، QDA[8] و RF، از طریق مجموعه داده تشخیص نفوذ CICIDS2017 مورد بررسی قرار گرفتند. علاوه بر این، نتایج تأیید کرد که طبقه‌بندی کننده KNN از نظر دقت، یادآوری، صحت و امتیاز F1 از مدل‌های مقابل پیشی گرفته است. با این حال، همه طبقه‌بندی‌کننده‌های پیاده‌سازی شده به جز KNN، زمان آموزش منطقی داشتند [۲۹].

۳. روش‌شناسی پژوهش

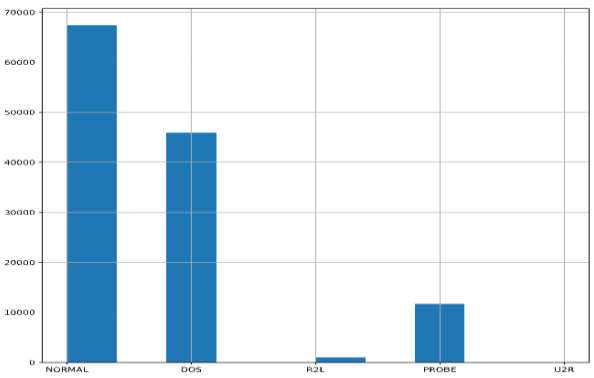
این بخش به شرح مجموعه داده می‌پردازد و با توضیح روش انتخاب ویژگی پیشنهادی و ساختار کلی ادامه می‌یابد.

۳-۱. شرح مجموعه داده

با توجه به معایب ذاتی مجموعه داده KDD Cup 99، یک نسخه اصلاح شده از آن، به نام NSL-KDD، ایجاد شده است که بر کارایی تشخیص مدل‌های IDS بسیار تأثیرگذار است. این مجموعه داده جدید شامل اطلاعات لازم از کل مجموعه اده KDD است. ۴۱ ویژگی در آن وجود دارد که مشخصات هر نمونه را توصیف می‌کند،و هر نمونه با یک عنوان حمله یا نرمال برچسب‌گذاری می‌شود. ویژگی ۴۲ام شامل داده‌های مربوط به ۵ کلاس مختلف شبکه است که به‌عنوان یک کلاس نرمال و چهار کلاس حمله گروه‌بندی شده‌اند. ۴ کلاس حمله به عناوین انکار سرویس (DoS)، حمله کاوشگر (Probe)، از راه دور به محلی (R2L) و کاربر به ریشه (U2R) اختصاص داده شده‌اند [۳۰]. جدول (۱) کلاس حمله را با نوع حمله نگاشت کرده است. برای درک بهتر، شکل‌های (۱) و (۲) توزیع نمونه‌های عادی و حمله موجود در زیرمجموعه‌های آموزش شامل ۱۲۵۹۷۳ رکورد و آزمایش شامل ۲۲۵۴۴ رکورد از مجموعه داده NSL-KDD را نمایش می‌دهد. در مجموعه داده آموزش، کلاس‌های نرمال، DOS، R2L، U2R و Probing به ترتیب ۵۳/۴۷٪، ۳۶/۴۷٪، ۰/۷۶٪، ۰/۰۴٪ و ۹/۲۶٪ از کل را به خود اختصاص داده‌اند. در مجوعه داده آزمایش، این نسبت‌ها ۵۱/۳۵٪، ۳۰/۶۷٪، ۱۱/۷۹٪، ۰/۲۴٪ و ۵/۹۵٪ برای به ترتیب کلاس‌های نرمال، DOS، R2L، U2R و Probing است.

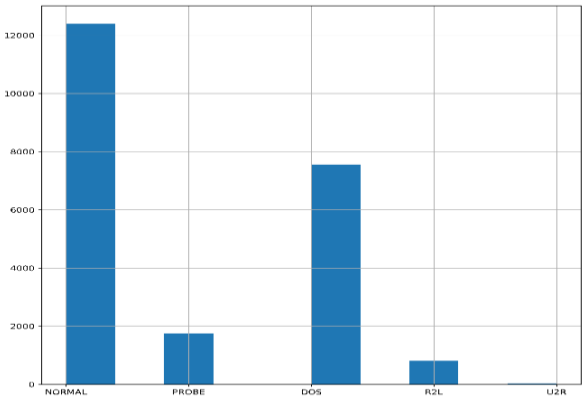
جدول ۱. کلاس‌بندی حملات مجموعه داده NSL-KDD

کلاس حمله	۲۲ نوع حمله
DOS	back, land, neptune, pod, smurf, teardrop
R2L	ftp_write, guess_passwd, imap, multhop, phf, spy, warezclient, warezmaster
U2R	buffer_overflow, perl, loadmodule, rootkit
Probing	ipsweep, nmap, portsweep, satan



شکل ۱. کلاس‌بندی مجموعه داده آموزش NSL-KDD

[1] Bayesian Network
[2] Multivariate Mutual Information-based Feature Selection
[3] Least Square Support Vector Machine
[4] Light Gradient Boosting Machine
[5] K Nearest Neighbor
[6] Gradient Boosting Tree
[7] Virtual Private Networks
[8] Quadratic Discriminant Analysis



شکل ۲. کلاس‌بندی مجموعه داده آزمایش NSL-KDD

۳-۲. انتخاب ویژگی

به‌منظور فراهم نمودن بهترین بردار ویژگی بهینه، ابتدا روش‌های انتخاب ویژگی MIT-MIT، آزمون F تحلیل واریانس و GA برای انتخاب زیرلیست‌های ویژگی به‌طور جداگانه و هم‌زمان پیاده‌سازی می‌شوند. پس از این، آن‌ها از یک واحد فیلتر اشتراک‌گیری عبور خواهند کرد تا زیرمجموعه ویژگی‌های با امتیاز بالا را ایجاد کنند. شایان ذکر است که روش پیشنهادی جهت دستیابی به تمامی ویژگی‌های با امتیاز برتر که همان ویژگی‌های مشترک هستند، مورد تحلیل قرار گرفته است.

اگر $F=\{f_1,...,f_n\}$ به‌عنوان مجموعه ویژگی‌های اولیه تعیین شود، که از مرحله پیش‌پردازش در مجموعه داده به‌دست می‌آید، $F_{s1}=\{f_{m1},...,f_{mn}\}$، $F_{s2}=\{f_{a1},...,f_{an}\}$ و $F_{s3}=\{f_{g1},...,f_{gn}\}$ به‌عنوان زیرمجموعه ویژگی‌ها تعریف می‌گردند که به ترتیب از طریق الگوریتم‌های MIT-MIT، آزمون F تحلیل واریانس و GA برگزیده شده‌اند. به‌منظور تهیه یک زیرفهرست ویژگی‌ها با فیلتر اشتراک‌گیری، تنها ویژگی‌های مشترک در هر سه زیرمجموعه قبلی به‌صورت زیر انتخاب خواهند شد:

$$F_{s4}=F_{s1}\cap F_{s2}\cap F_{s3} \qquad (۱۰)$$

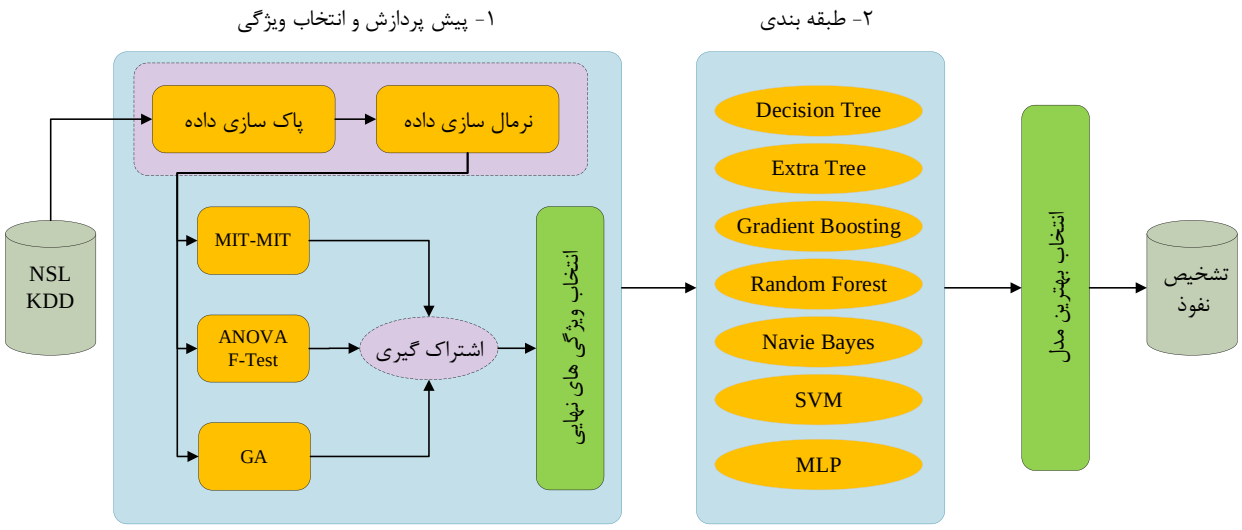
شبه‌کد این روش انتخاب ویژگی ترکیبی در شکل (۳) نشان داده شده است. در مرحله بعدی، این زیرلیست ویژگی‌های ادغام شده برای آموزش به طبقه‌بندی کننده‌ها تحویل می‌گردد.

```
Input: Historical dataset
Output: A subset of features
        //Preprocessing dataset
        Prep_DS = Replace missing and defective values
                  with the average value
        //Normalizing dataset
        Prep_DS = MinMaxScaler( Prep_DS )
        SelectedFeatures _GA =
                  Genetic_Algorithm(Prep_DS)
        SelectedFeatures _Anova = Anova(Prep_DS)
        SelectedFeatures _MIT-MIT = MIT-MIT(Prep_DS)
        SelectedFeatures = INTERSECTION(
                           SelectedFeatures_GA,
                           SelectedFeatures _Anova,
                           SelectedFeatures _MIT-MIT)
        Return  SelectedFeatures
end
```

شکل ۳. شبه‌کد روش انتخاب ویژگی پیشنهادی

۳-۳. ساختار کلی پیشنهادی

در شکل (۴)، ساختار کلی روش پیشنهادی آورده شده است که بر اساس آن مجموعه داده اولیه به مرحله پیش‌پردازش وارد می‌شود که شامل تمیز کردن و نرمال‌سازی داده‌ها است. به دنبال آن، بلوک انتخاب ویژگی، لیستی از مرتبط‌ترین ویژگی‌ها و نه تکراری را انتخاب می‌کند و به موتور طبقه‌بندی ارائه می‌دهد. طبقه‌بندی کننده، تشخیص نفوذ را بر اساس زیرمجموعه‌های آموزش و آزمایش انجام می‌دهد و نتایج را از طریق معیارهای شناخته شده ارزیابی می‌کند.



شکل ۴. ساختار کلی مدل پیشنهادی

۴. نتایج و بحث

در این بخش، معیارهای ارزیابی، تنظیمات پیکربندی، به همراه نتایج عملکرد انتخاب ویژگی و طبقه‌بندی ارائه شده و مورد بحث قرار می‌گیرند.

۴-۱. معیارهای ارزیابی عملکرد

به‌منظور برآورد اثربخشی مدل ارائه شده، فاکتورهای امتیاز F1، دقت، یادآوری و صحت معرفی می‌شوند. این معیارها از ماتریس درهم‌ریختگی (جدول ۲)، حاوی مؤلفه‌های مثبت صحیح (TP)، منفی صحیح (TN)، مثبت کاذب (FP) و منفی کاذب (FN) حاصل می‌گردند. در واقع، فرآیند ارزیابی بیشتر بر صحت متمرکز است، زیرا که در اکثر مطالعات در این زمینه وارد شده است [۲۰، ۲۷ و ۲۹].

جدول ۲. ماتریس درهم‌ریختگی

		پیش‌بینی شده	
		نرمال	حمله
واقعی	نرمال	TP	FN
	حمله	FP	TN

- دقت

به‌عنوان نسبت حملات شناسایی شده درست، نسبت به تمام داده‌های طبقه‌بندی شده به‌عنوان حمله تعریف می‌شود که به‌صورت زیر محاسبه خواهد شد:

$$\text{Pr}ecision = \frac{TP}{TP + FP} \qquad (۱۱)$$

- یادآوری

یادآوری در این زمینه، نسبت حملات به‌طور صحیح شناسایی شده، به تمام داده‌های حمله است. یادآوری به‌صورت زیر تعیین می‌شود:

$$\text{Re}call = \frac{TP}{TP + FN} \qquad (۱۲)$$

- امتیاز F1

این معیار ترکیبی را نشان می‌دهد که میانگین هارمونیک دقت و یادآوری را به‌صورت زیر تخمین می‌زند:

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \qquad (۱۳)$$

- صحت

به‌عنوان نسبت رکوردهایی که به‌طور صحیح به‌عنوان حمله یا نرمال تصمیم‌گیری شده‌اند به همه داده‌ها درنظر گرفته می‌شود. تخمین صحت به‌صورت زیر بیان می‌گردد:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۱۴)$$

۴-۲. تنظیمات پیکربندی

عملیات مبتنی بر تصادفی بودن شبکه‌های DL باعث نتایج متنوعی در هر اجرا شده است. از این رو، در این پژوهش هر الگوریتم ده بار اجرا می‌شود و نتیجه نهایی به‌صورت مقدار میانگین آن‌ها محاسبه خواهد شد. لازم به ذکر است که تمامی شبیه‌سازی‌ها در پایتون ۳/۸ پیاده‌سازی شده‌اند و کتابخانه Keras برای دسترسی به DNNها مورد استفاده قرار گرفته است. همه شبیه‌سازی‌ها روی یک لپ‌تاپ با مشخصات سخت‌افزاری Intel Core i7، ۳/۱ GHz CPU و ۱۶ گیگابایت رم صورت گرفته است.

۴-۳. ارزیابی عملکرد انتخاب ویژگی

پس از تکمیل اولین گام انتخاب ویژگی، روش‌های MIT-MIT، آزمون F تحلیل واریانس و GA به ترتیب ۲۲، ۱۶ و ۱۶ ویژگی را انتخاب نمودند. پس از آن، ویژگی‌های برگزیده از سه گروه به‌طور هم‌زمان از فیلتر اشتراک‌گیری عبور کرده و در نتیجه ۱۴ ویژگی فهرست شده در جدول (۳) استخراج گردید.

جدول ۳. ویژگی‌های برگزیده نهایی

ویژگی‌های انتخاب شده
src_bytes, dst_bytes, count,same_srv_rate,dst_host_count, dst_host_srv_count,dst_host_same_srv_rate, dst_host_diff_srv_rate,dst_host_same_src_port_rate, dst_host_srv_diff_host_rate,dst_host_srv_serror_rate, rerror_rate, dst_host_srv_rerror_rate, Flag

در مرحله بعد، ویژگی‌های انتخابی MIT-MIT، آزمون F تحلیل واریانس و GA، به همراه راهبرد پیشنهادی، در معرض طبقه‌بندی کننده‌های معروف قرار گرفتند تا آن‌ها را از نظر معیارهای ارزیابی مقایسه نمایند. نتایج مقایسه در جدول‌های (۷-۴) به تفصیل بیان شده است. همان‌طور که مشخص است، روش انتخاب ویژگی ترکیبی پیشنهادی مبتنی بر اشتراک‌گیری در مورد تمام معیارهای ارزیابی به بالاترین عملکرد رسیده است. از نظر امتیاز F1 و دقت، روش پیشنهادی نسبت به الگوریتم‌های انتخاب ویژگی منفرد به‌طور میانگین به ترتیب ۸٪ و ۱۰٪ بهبودی را نشان می‌دهد. به همین ترتیب، روش مبتنی بر اشتراک‌گیری دارای نرخ میانگین افزایشی یادآوری و صحت برای IDS است که هر دو ۷٪ پیشرفت داشته‌اند؛ این مسئله برتری مدل پیشنهادی را نسبت به سایر روش‌های رقیب تأیید می‌کند.

از دیدگاه دیگر، در رابطه با مدل‌های طبقه‌بندی، مشخص شده است که الگوریتم‌های طبقه‌بندی ensemble (تقویت گرادیان و RF) به ترتیب نسبت به سایر طبقه‌بندی کننده‌ها از لحاظ امتیاز F1 و یادآوری کارآمدتر هستند. علاوه بر این، SVM به‌عنوان یک طبقه‌بندی کننده ML دقت ۰/۹ را به‌دست آورد. همچنین، MLP به‌عنوان یک مدل DL به فراخوانی ۰/۶۵۸۶۷ دست یافت که به وضوح نسبت به سایر روش‌های رقیب برتری دارند. قابل ذکر است که شکل‌های (۸-۵)، نتایج آزمایش‌های جدول‌های (۴-۷) را در

قالب نمودار میله‌ای تصویر کرده است تا با نگاهی کوتاه بتوان به تحلیل بیان شده دست یافت.

جدول ۴. مقایسه عملکرد مدل‌های کلاس‌بندی مبتنی بر انتخاب ویژگی MIT-MIT

	امتیاز F1	دقت	یادآوری	صحت
Decision Tree	۰/۵۲۶۹۳	۰/۵۱۶۹۷	۰/۵۷۵۵۰	۰/۸۱۵۸۹
Extra Tree	۰/۴۸۰۵۶	۰/۵۶۳۶۰	۰/۵۳۲۲	۰/۸۲۰۱۱
Gradient Boosting	۰/۵۶۹۷۰	۰/۵۷۳۲۵	۰/۶۱۸۲۸	۰/۸۴۳۷۱
Random Forest	۰/۵۲۳۳۹	۰/۶۳۰۷۲	۰/۵۸۲۴۷	۰/۸۵۲۰۵
Navie Bayes	۰/۳۸۸۰۹	۰/۴۵۴۳۷	۰/۵۳۵۹۰	۰/۴۷۲۰۰
SVM	۰/۵۲۱۶۳	۰/۶۶۵۲۴	۰/۵۸۳۳۹	۰/۸۵۳۷۴
MLP	۰/۵۲۹۸۸	۰/۶۶۴۸۷	۰/۵۹۱۱۸	۰/۸۲۵۴۰

جدول ۵. مقایسه عملکرد مدل‌های کلاس‌بندی مبتنی بر انتخاب ویژگی آزمون F تحلیل واریانس

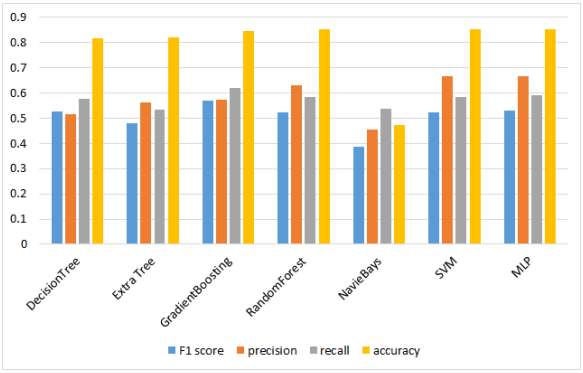
	امتیاز F1	دقت	یادآوری	صحت
Decision Tree	۰/۵۲۳۷۳	۰/۴۹۸۹۹	۰/۵۷۸۷۳	۰/۸۲۰۸۶
Extra Tree	۰/۵۰۸۶۳	۰/۶۱۲۵۴	۰/۵۷۱۹۲	۰/۸۲۸۷۲
Gradient Boosting	۰/۵۱۴۲۰	۰/۴۸۱۹۰	۰/۵۶۵۵۲	۰/۸۱۸۵۶
Random Forest	۰/۵۱۸۳۶	۰/۵۵۶	۰/۵۷۸۸۱	۰/۸۴۶۷۷
Navie Bayes	۰/۲۹۹۵۰	۰/۴۶۰۳۷	۰/۴۵۶۱۶	۰/۲۶۰۰۹
SVM	۰/۵۱۶۷۷	۰/۶۷۴۰۷	۰/۵۷۸۹۳	۰/۸۵۳۲۵
MLP	۰/۵۲۸۵۳	۰/۶۵۴۸۶	۰/۵۸۷۰۶	۰/۸۳۳۰۲

جدول ۶. مقایسه عملکرد مدل‌های کلاس‌بندی مبتنی بر انتخاب ویژگی GA

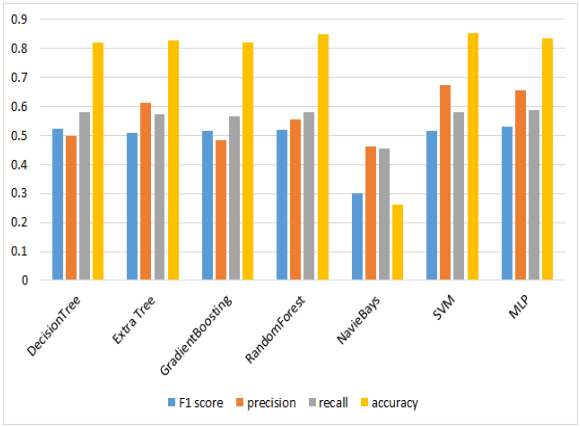
	امتیاز F1	دقت	یادآوری	صحت
Decision Tree	۰/۵۳۵۹۸	۰/۵۴۱۷۷	۰/۵۸۲۶۱	۰/۸۲۸۳۶
Extra Tree	۰/۵۳۸۸۴	۰/۶۷۸۴۷	۰/۵۶۸۲۰	۰/۸۳۵۲۸
Gradient Boosting	۰/۵۳۷۰۲	۰/۶۵۸۷۳	۰/۵۸۴۳۰	۰/۸۲۸۴۰
Random Forest	۰/۵۵۸۷۹	۰/۷۳۶۹۷	۰/۵۹۶۸۸	۰/۸۴۸۶۳
Navie Bayes	۰/۲۸۵۱۲	۰/۴۲۰۲۸	۰/۴۸۲۴۳	۰/۳۰۵۶۵
SVM	۰/۵۱۳۵۲	۰/۶۶۴۳۵	۰/۵۷۳۸۹	۰/۸۳۵۵۹
MLP	۰/۵۲۶۱۵	۰/۶۲۷۷۴	۰/۵۸۴۱۷	۰/۸۳۰۶۷

جدول ۷. مقایسه عملکرد مدل‌های کلاس‌بندی مبتنی بر انتخاب ویژگی ترکیبی پیشنهادی

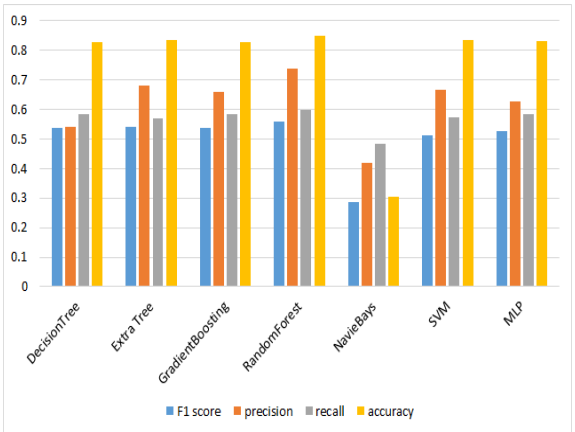
	امتیاز F1	دقت	یادآوری	صحت
Decision Tree	۰/۵۹۲۸۶	۰/۵۸۵۴۸	۰/۶۱۷۸۶	۰/۸۵۱۱۷
Extra Tree	۰/۵۸۰۴۲	۰/۷۵۸۸۷	۰/۵۹۲۶۴	۰/۸۶۳۲
Gradient Boosting	۰/۶۱۲۵۲	۰/۶۹۸۰۲	۰/۶۴۳۸۱	۰/۸۷۶۸۱
Random Forest	۰/۶۰۳۳۳	۰/۷۹۰۳۵	۰/۶۳۶۶۲	۰/۸۸۵۴۷
Navie Bayes	۰/۴۲۲۹۰	۰/۴۹۳۸۶	۰/۵۶۰۶۵	۰/۵۲۷۳۵
SVM	۰/۵۸۶۱۵	۰/۷۴۱۳۲	۰/۶۴۳۷۳	۰/۸۹۹۹۹
MLP	۰/۵۷۶۲۶	۰/۷۲۴۱۲	۰/۶۵۸۶۷	۰/۸۸۶۵۷



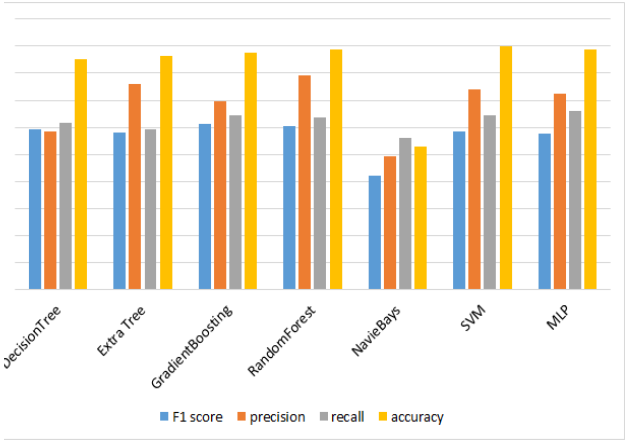
شکل ۵. نمودار میله‌ای ارزیابی روش انتخاب ویژگی MIT-MIT



شکل ۶. نمودار میله‌ای ارزیابی روش انتخاب ویژگی آزمون F تحلیل واریانس



شکل ۷. نمودار میله‌ای ارزیابی روش انتخاب ویژگی GA



شکل ۸. نمودار میله‌ای ارزیابی روش انتخاب ویژگی ترکیبی پیشنهادی مبتنی بر اشتراک‌گیری

علاوه بر این، جهت تکمیل ارزیابی روش پیشنهادی، مقایسه‌ای از کارایی این مدل با کارهای تحقیقاتی اخیر صورت گرفته است، به‌طوری که همگی روی مجموعه داده NSL-KDD پیاده‌سازی شده‌اند. این بررسی جوانبی همچون الگوریتم‌های انتخاب ویژگی و طبقه‌بندی را بر اساس معیار صحت مدنظر قرار داده است که در جدول (۸) ارائه شده‌اند. طبق جدول (۸)، مدل پیشنهادی با بهره‌گیری از روش ترکیبی انتخاب ویژگی مؤثر به پیشرفت ۲/۹۹٪ درصدی نسبت به بالاترین صحت دست یافته است.

جدول ۸. مقایسه مدل پیشنهادی با برخی از مقالات اخیر

صحت	انتخاب ویژگی/ طبقه بندی	مرجع، سال
٪ ۸۷/۰۰	GRU-LSTM /ANOVA F-RFE	[۲۲]، ۲۰۱۸
٪ ۷۶/۳۶	CART /MI-Firefly	[۳۱]، ۲۰۱۹
٪ ۸۴/۹۳	SVM /Local & EDA	[۳۲]، ۲۰۲۰
٪ ۸۱/۷۵	DT /PMSO-BAT	[۳۳]، ۲۰۲۱
٪ ۸۵/۵۰	Ensemble Classifiers /AFS	[۲۱]، ۲۰۲۱
٪ ۸۹/۹۹	SVM /MIT-GA-ANOVA	روش پیشنهادی

۴-۴. تحلیل زمانی

در حقیقت، یکی دیگر از مسائل مورد توجه در تحلیل مدل‌های ML و DL، اندازه‌گیری زمان آموزش و آزمایش در مقایسه روش‌ها است. بر این اساس، جدول (۹) زمان آموزش را برای الگوریتم‌های فوق‌الذکر مقایسه کرده است و به همین ترتیب، جدول (۱۰) مقایسه زمان‌های آزمایش را ارائه می‌دهد. از جدول‌های (۹) و (۱۰) می‌توان مشاهده کرد که رویکرد انتخاب ویژگی مبتنی بر اشتراک‌گیری پیشنهادی به‌طور متوسط در مقابل الگوریتم‌های انتخاب ویژگی فردی که در بازه‌های زمانی مختلف با توجه به نوع طبقه‌بندی کننده تحلیل شده‌اند، کوتاه‌ترین زمان آموزش و آزمایش را به خود اختصاص داده است.

جدول ۹. مقایسه عملکرد روش‌های انتخاب ویژگی بر حسب زمان آموزش

	آزمون F تحلیل واریانس	MI	GA	روش پیشنهادی
Decision Tree	۰/۶۶۸۹۱	۰/۸۳۵۰۷	۰/۶۰۶۳۵	۰/۶۴۹۸۰
Extra Tree	۱/۶۷۴۳۳	۱/۹۳۷۱۹	۱/۶۸۲۴۹	۱/۵۰۰۶۱
Gradient Boosting	۱۱۹/۵۰۳	۱۴۷/۳۲۱	۱۰۹/۱۷۳	۱۰۲/۸۳۱
Random Forest	۲/۶۸۰۴۴	۰/۰۹۶۹۳	۲/۸۳۸۲۶	۲/۷۲۱۳۳
Navie Bayes	۰/۰۸۴۸۷	۰/۱۰۳۱۱	۰/۰۹۲۶۰	۰/۰۷۱۷۸
SVM	۹۷/۶۶۳۴	۱۰۶/۷۳۲	۲۰۰/۳۳۱	۸۸/۲۳۴۳
MLP	۲۷۸/۹۳۱	۲۵۴/۱۹۶	۳۷۱/۵۳۹	۲۵۰/۳۲۰

جدول ۱۰. مقایسه عملکرد روش‌های انتخاب ویژگی بر حسب زمان آزمایش

	آزمون F تحلیل واریانس	MI	GA	روش پیشنهادی
Decision Tree	۰/۰۱۱۹۵	۰/۰۱۲۹۴	۰/۰۱۰۱۶	۰/۰۱۰۰۹
Extra Tree	۰/۰۹۸۱۴	۰/۰۹۴۶۷	۰/۰۹۳۷۵	۰/۰۸۸۱۸
Gradient Boosting	۰/۱۹۳۵۶	۰/۱۹۳۴۷	۰/۱۹۲۵۵	۰/۱۹۱۶۰
Random Forest	۰/۰۸۴۸۰	۰/۰۹۲۴۵	۰/۱۰۰۳۸	۰/۰۸۲۴۶
Navie Bayes	۰/۰۲۱۶۳	۰/۰۳۰۰۱	۰/۰۲۶۳۰	۰/۰۱۹۹۹
SVM	۱۷/۸۲۷۳	۱۳/۶۵۸۳	۳۱/۸۴۵۷	۱۳/۱۵۰۴
MLP	۲۸۹/۷۶۳	۲۵۵/۱۱۰	۳۷۲/۴۶۷	۲۲۳/۲۵۱

۵. نتیجه‌گیری

یکی از چالش‌های غالب IDS، افزایش دقت مدل از طریق راهبردهای پیشرفته و ترکیبی است. رویکردهای ترکیبی شامل انتخاب ویژگی و روش‌های ML یا DL، دقت طبقه‌بندی کننده‌ها را بهبود بخشیده‌اند، حجم داده‌های ورودی را کاهش داده‌اند و عملیات را سرعت بخشیده‌اند. بر این اساس، روش ارائه شده بر یک راهبرد انتخاب ویژگی ترکیبی به خوبی تعریف شده متمرکز شده است. تجزیه و تحلیل شبیه‌سازی بر روی مجموعه داده گسترده NSL-KDD اجرا شده است. با توجه به مقایسه رویکرد انتخاب ویژگی پیشنهادی با الگوریتم‌های فردی بر اساس معیارهای ارزیابی، تأیید شده است که روش پیشنهادی ادغام شده با طبقه‌بندی کننده‌های قدرتمند بالاترین عملکرد را به‌دست آورده است. علاوه بر این، از نقطه‌نظر زمان آموزش و آزمایش، انتخاب ویژگی پیشنهادی نتایج امیدوار کننده‌ای در مقایسه با سایر روش‌های فردی دارد. به‌عنوان پژوهش آینده، طبقه‌بندی کننده‌های ترکیبی را می‌توان برای IDS استفاده نمود، و همچنین اجزای خطی و غیر خطی مجموعه داده را برای اعمال یک روش ترکیبی آماری-DNN جداسازی کرد. علاوه بر

ایـن، روش‌هـای پیشـرفته بهینـه‌سـازی مـی‌تواننـد بـا محاسـبه فراپارامترهای بهینه مدل‌های DNN برای ارزیابی تـأثیر آن‌هـا بـر دقت مدل، نقش داشته باشند.

۵. مراجع

[1] Mehanović, D.; Kečo, D.; Kevrić, J.; Jukić, S.; Miljković, A.; Mašetić, Z. "Feature Selection Using Cloud-based Parallel Genetic Algorithm for Intrusion Detection Data Classification"; Neural Computing and Applications 2021, 33, 11861-11873.

[2] Najafi, M.; Rafeh, R. "A New Light Weight Intrusion Detection Algorithm for Computer Networks"; Adv. Defence Sci. & Technol. 2017, 10, 191-200 (In Persian).

[3] Dubey, G. P.; Bhujade, R. K. "Optimal Feature Selection for Machine Learning Based Intrusion Detection System by Exploiting Attribute Dependence"; Materials Today: Proc. 2021, 47, 6325-6331.

[4] Kamalov, F.; Moussa, S.; Zgheib, R.; Mashaal, O. "Feature Selection for Intrusion Detection Systems"; 13th Int. Symp. on Computational Intelligence and Design, IEEE, 2020.

[5] Shakeela, S.; Shankar, N. S.; Reddy, P. M.; Tulasi, T. K.; Sai, M. M. "Optimal Ensemble Learning Based on Distinctive Feature Selection by Univariate ANOVA-F Statistics for IDS"; Int. J. of Electronics and Telecommunications 2021, 67, 267-275.

[6] Ibrahim, Z. K.; Thanon, M. Y. "Performance Comparison of Intrusion Detection System Using Three Different Machine Learning Algorithms"; 6th Int. Conf. on Inventive Computation Tech., IEEE, 2021.

[7] Kalimuthan, C.; Renjit, J. A. "Review on Intrusion Detection Using Feature Selection with Machine Learning Techniques"; Materials Today: Proc. 2020, 33, 3794-3802.

[8] Tajari Siahmarzkooh, A. A. "Intrusion Detection in Computer Networks Using Decision Tree and Feature Reduction"; Electron. Cyber Defence 2017, 9, 99-108 (In Persian).

[9] Sarhan, M.; Layeghy, S.; Portmann, M. "Towards a Standard Feature Set for Network Intrusion Detection System Datasets"; Mobile Networks and Applications 2021, 1-14.

[10] Moualla, S.; Khorzom, K.; Jafar, A. "Improving the Performance of Machine Learning-based Network Intrusion Detection Systems on the UNSW-NB15 Dataset"; Computational Intelligence and Neuroscience 2021, 1-13.

[11] Khafajeh, H. A. Y. E. L. "An Efficient Intrusion Detection Approach Using Light Gradient Boosting"; J. Theor. Appl. Inform. Technol. 2020, 98, 825-835.

[12] Çalışır, S.; Atay, R.; Pehlivanoğlu, M. K.; Duru, N. "Intrusion Detection Using Machine Learning and Deep Learning Techniques"; 4th Int. Conf. Comput. Sci. Eng., IEEE, 2019.

[13] Taheri, R.; Parsaei, M. R.; Javidan, R. "Real-Time Intrusion Detection System Using a Combination of Discretization and Feature Selection"; Adv. Defence Sci. & Technol. 2017, 10, 251-263 (In Persian).

[14] Mohammadi, M.; Rashid, T. A.; Karim, S. H. T.; Aldalwie, A. H. M.; Tho, Q. T.; Bidaki, M.;... Hosseinzadeh, M. "A Comprehensive Survey and Taxonomy of the SVM-based Intrusion Detection Systems"; J. Network Comput. App. 2021, 178, 102983.

[15] Alizadeh, M.; Beheshti, M. T.; Ramezani, A.; Saadatinezhad, H. "Network Traffic Forecasting Based on Fixed Telecommunication Data Using Deep Learning"; 6th Iranian Conf. Signal Proc. Intell. Syst., IEEE, 2020.

[16] Thakkar, A.; Lohiya, R. "A Survey on Intrusion Detection System: Feature Selection, Model, Performance Measures, Application Perspective, Challenges, and Future Research Directions"; Artificial Intelligence Review 2021, 1-111.

[17] Sultana, N.; Chilamkurti, N.; Peng, W.; Alhadad, R. "Survey on SDN Based Network Intrusion Detection System Using Machine Learning Approaches"; Peer-to-Peer Networking App. 2019, 12, 493-501.

[18] Alizadeh, M.; Mousavi, S. E.; Beheshti, M. T.; Ostadi, A. "Combination of Feature Selection and Hybrid Classifier for Network Intrusion Detection System Based on GWO, BAT, and FA Algorithms"; 6th Iranian Conf. Signal Proc. Intell. Syst., IEEE, 2020.

[19] Desale, K. S.; Ade, R. "Genetic Algorithm Based Feature Selection Approach for Effective Intrusion Detection System"; Int. Conf. Comput. Commun. Inform., IEEE, 2015.

[20] Saba, T.; Sadad, T.; Rehman, A.; Mehmood, Z.; Javaid, Q. "Intrusion Detection System through Advance Machine Learning for the Internet of Things Networks"; IT Prof. 2021, 23, 58-64.

[21] Lin, C.; Li, A.; Jiang, R. "Automatic Feature Selection and Ensemble Classifier for Intrusion Detection"; J. Phys.: Conf. Series Vol. 1856. No. 1. IOP Publishing, 2021.

[22] Dey, S. K.; Rahman, M. M. "Flow Based Anomaly Detection in Software Defined Networking: a Deep Learning Approach with Feature Selection Method"; 4th Int. Conf. Electrical Eng. Inform. & Commun. Tech., IEEE, 2018.

[23] Selvakumar, B.; Muneeswaran, K. "Firefly Algorithm Based Feature Selection for Network Intrusion Detection"; Computers & Security 2019, 81, 148-155.

[24] Mohammadi, S.; Desai, V.; Karimipour, H. "Multivariate Mutual Information-based Feature Selection for Cyber Intrusion Detection"; IEEE Electrical Power and Energy Conf., IEEE, 2018.

[25] Abhale, A. B.; Manivannan, S. S. "Supervised Machine Learning Classification Algorithmic Approach for Finding Anomaly Type of Intrusion Detection in Wireless Sensor Network"; Optical Memory and Neural Networks 2020, 29, 244-256.

[26] Faker, O.; Dogdu, E. "Intrusion Detection Using Big Data and Deep Learning Techniques"; Proc. 2019 ACM Southeast Conf. 2019.

[27] Dobson, A.; Roy, K.; Yuan, X.; Xu, J. "Performance Evaluation of Machine Learning Algorithms in Apache Spark for Intrusion Detection"; 28th Int. Telecommun. Networks App. Conf., IEEE, 2018.

[28] Afuwape, A. A.; Xu, Y.; Anajemba, J. H.; Srivastava, G. "Performance Evaluation of Secured Network Traffic Classification Using a Machine Learning Approach"; Computer Standards & Interfaces 2021, 78, 103545.

[29] Alrowaily, M.; Alenezi, F.; Lu, Z. "Effectiveness of Machine Learning Based Intrusion Detection Systems"; Int. Conf. Security, Privacy and Anonymity in Computation, Communication and Storage, Springer, Cham, 2019.

[30] Dhanabal, L.; Shantharajah, S. P. "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification ALlgorithms"; Int. J. Adv. Res. Comput. Commun. Eng. 2015, 4, 446-452.

[31] Wang, Z.; Tang, M.; Deng, J.; Wang, Y.; Qian, J., Chen, X. "A New Feature Selection Method for Intrusion Detection"; Int. Conf. Ubiquitous Comput. & Commun. and Data Sci. and Computational Intelligence and Smart Computing, Networking and Services, IEEE, 2019.

[32] Sharifiasn, M.; Karshenas, H.; Sharifiasn, S. "Improving Network Intrusion Detection by Identifying Effective Features using Evolutionary Algorithms based on Support Vector Machine"; Comput. Intell. Electrical Eng. 2020, 11, 29-42.

[33] Alkafagi, S. S.; Almuttairi; R. M. "A Proactive Model for Optimizing Swarm Search Algorithms for Intrusion Detection System"; J. Phys.: Conf. Series, Vol. 1818, IOP Publishing, 2021.