



Smart Home Intrusion Detection Model based on Principal Component Analysis and Random Forest Classification

A. Tajari Siahmarzkooh 

Assistant Professor, Golestan University, Gorgan, Iran.

(Received: 2024/06/04, Revised: 2024/07/27, Accepted: 2024/08/13, Published: 2024/08/31)

DOR: <https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.2.2.2>

ABSTRACT

In recent years, the problem of maintaining the security of smart homes, where a large number of devices use Internet connections to communicate, has become one of the main concerns in the field of network security. Although a lot of research has been done to establish the security of smart homes, but considering the scope of the topic under discussion, most of these works do not have the necessary efficiency in terms of accuracy and speed of operation. In the proposed method, after performing some pre-processing operations on the dataset, with the help of Principal Component Analysis (PCA), a subset of the features of the dataset are selected to prepare the data for classification, which are the most effective features in intrusion detection. It is considered that this action leads to an increase in the accuracy and speed of the classification action. Also, in the classification stage, the random forest algorithm, which is a powerful algorithm based on machine learning, has been used on a very new dataset of the Internet of Things, called IoTID20. The proposed approach has shown high performance for intrusion detection with an accuracy of 99.73% and 98.46% for the classification of binary and multi-class attacks. Comparing the results of the proposed method with other works, it shows the superiority of the proposed method in detecting multi-class attacks.

Keywords: Smart home, Intrusion Detection, Principal Component Analysis (PCA), Random Forest Algorithm, IoTID20 dataset.

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Publisher: Imam Hussein University

 Authors



*Corresponding Author Email: a.tajari@gu.ac.ir

علمی - پژوهشی

مدل تشخیص نفوذ در خانه‌های هوشمند مبتنی بر تحلیل مؤلفه اصلی و دسته‌بندی جنگل تصادفی

علی اکبر تجری سیاه‌مرزکوه

استادیار، دانشگاه گلستان، گلستان، ایران.

(دریافت: ۱۴۰۳/۳/۱۵، بازنگری: ۱۴۰۳/۵/۰۶، پذیرش: ۱۴۰۳/۵/۲۳، انتشار: ۱۴۰۳/۶/۱۰)

DOR: <https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.2.2.2>



* این مقاله یک مقاله با دسترسی آزاد است که تحت شرایط و ضوابط مجوز (CC BY) Creative Commons توزیع شده است.
ناشر: دانشگاه جامع امام حسین (ع) نویسندگان

چکیده

در سال‌های اخیر، مسئله حفظ امنیت خانه‌های هوشمند که در آن، تعداد زیادی از وسایل برای برقراری ارتباط از اتصالات اینترنت استفاده می‌کنند به یکی از دغدغه‌های اصلی در حوزه امنیت شبکه تبدیل شده است. اگرچه تاکنون پژوهش‌های زیادی در جهت برقراری امنیت خانه‌های هوشمند انجام شده است، اما با توجه به گستردگی موضوع مورد بحث، اغلب این کارها در مواردی از جمله دقت و سرعت عمل، کارایی لازم را ندارند. در روش پیشنهادی پس از انجام برخی عملیات پیش‌پردازش روی مجموعه داده، به کمک تحلیل مؤلفه اصلی (PCA)، زیرمجموعه‌ای از ویژگی‌های مجموعه داده که به عنوان مؤثرترین ویژگی‌ها در تشخیص نفوذ به شمار می‌آیند برای آماده‌سازی داده‌ها جهت دسته‌بندی انتخاب شده‌اند که این عمل منجر به افزایش دقت و سرعت عمل دسته‌بندی می‌شود. همچنین در مرحله دسته‌بندی از الگوریتم جنگل تصادفی که یک الگوریتم قدرتمند مبتنی بر یادگیری ماشین است بر روی یک مجموعه داده بسیار جدید اینترنت اشیا، به نام IoTID20 استفاده شده است. رویکرد پیشنهادی عملکرد بالایی برای تشخیص نفوذ با دقت ۹۹٫۷۳٪ و ۹۸٫۴۶٪ برای دسته‌بندی حملات دودویی و چند کلاسه نشان داده است. مقایسه‌ی نتایج روش پیشنهادی با سایر کارهای انجام شده، نشان دهنده‌ی برتری روش پیشنهادی در تشخیص حملات چند کلاسه است.

کلیدواژه‌ها: خانه هوشمند، تشخیص نفوذ، تحلیل مؤلفه اصلی، الگوریتم جنگل تصادفی، مجموعه داده IoTID20.

۱- مقدمه

موضوع امنیت و حفاظت از حریم خصوصی در طراحی این دستگاه‌ها در اولویت قرار نگرفته است. به همین دلیل، بخش قابل توجهی از پژوهش‌ها به ارتقای امنیت برنامه‌های مبتنی بر IoT تمرکز دارند. با این وجود، برای حفاظت کامل از شبکه‌های IoT در مقابل خطرات نفوذ، نیاز به انجام مطالعات و تحقیقات بیشتری است.

نرم‌افزارهای مورد استفاده در خانه‌های هوشمند نقش مهمی در گسترش استفاده از اینترنت اشیا ایفا کرده‌اند. خانه هوشمند به افراد این امکان را می‌دهد تا لوازم خانگی خود را به هم متصل کرده و از راه دور آن‌ها را کنترل کنند، این امر می‌تواند سهولت در دسترسی به لوازم را نیز فراهم نماید. با این حال، اگر تدابیر امنیتی کافی رعایت نشود، این مزیت با یک خطر امنیتی همراه خواهد بود [۲]. یکی از بزرگ‌ترین چالش‌ها در یک خانه هوشمند، ناتوانی کاربران در درک و اجرای اقدامات احتیاطی امنیتی است. اغلب دستگاه‌های اینترنت اشیا در خانه‌های هوشمند با حسگرها و محرک‌های متعددی طراحی شده‌اند. به عنوان مثال، یک چراغ

اینترنت اشیا^۲ (IoT) به عنوان یک فن آوری نوین و نوآورانه با تأثیر بسیار زیاد بر زندگی روزمره ما شناخته می‌شود [۱]. برنامه‌های کاربردی مختلف و پرکاربردی بر مبنای IoT توسعه یافته‌اند، از جمله خانه‌های هوشمند^۴، سلامت هوشمند، شهر هوشمند، شبکه هوشمند، کشاورزی هوشمند، خودروهای هوشمند و ردیابی لجستیک. دستگاه‌های اینترنت اشیا از طریق اتصال به اینترنت، برای جمع‌آوری و تبادل داده‌ها در راستای هدف مشخصی با یکدیگر ارتباط برقرار می‌کنند. با توجه به تقاضای بالا برای برنامه‌های مبتنی بر IoT، دستگاه‌های اینترنت اشیا به طور پیوسته در جهت کاهش هزینه‌ها، سهولت حمل و نقل، اندازه کوچک‌تر و دقت حسگرها در حال پیشرفت هستند. اما متأسفانه،

* رایانامه نویسنده مسئول: a.tajari@gu.ac.ir

³ Internet of Things

⁴ Smart homes

حملات مناسب نیست. در این روش، بهترین ویژگی‌ها برای ارزیابی عملکرد مدل مورد بررسی قرار گرفته است. با آزمایش روش پیشنهادی بر روی مجموعه داده جدید اینترنت اشیا به نام IoTID20 و انتخاب ویژگی‌های مؤثرتر، مقادیر پارامترهای ارزیابی بهبود پیدا خواهند کرد.

روش انجام شده در این مقاله شامل بخش‌های زیر است:

۱. پیش‌پردازش داده‌های مجموعه داده جهت آماده‌سازی اولیه
۲. استفاده از یک روش مناسب (PCA) برای انتخاب بهترین ویژگی‌ها در ارزیابی مدل پیشنهادی با کاهش سربار.
۳. استفاده از روش جنگل تصادفی برای دسته‌بندی / طبقه‌بندی داده‌ها به داده عادی یا حمله.
۴. ارزیابی و مقایسه روش پیشنهادی با روش‌های ارائه شده در قبل بر مبنای پارامترهای دقت (با اجرا بر روی مجموعه داده جدید IoTID20).

مقاله به شرح زیر سازمان‌دهی شده است: بخش ۲ پیشنهادی پژوهشی مرتبط با سیستم‌های تشخیص نفوذ خانه‌های هوشمند و همچنین تحقیقات انجام شده با استفاده از مجموعه داده IOTID20 و UNSWNB15 را مورد بحث قرار می‌دهد. در بخش ۳، جزئیات مربوط به راه کار پیشنهادی شامل پیش‌پردازش داده‌ها، انتخاب بهترین (مؤثرترین) ویژگی‌ها و توضیحات در مورد استفاده از روش جنگل تصادفی برای تشخیص نفوذ ارائه شده است. در بخش ۴، معیارهای ارزیابی مربوط به دسته‌بندی دودویی (عادی یا حمله) و چند کلاسه (نوع حمله) مطرح و بررسی شده‌اند. نتایج دسته‌بندی و تجزیه و تحلیل در بخش ۵ آورده شده‌اند. نتیجه‌گیری انجام شده و گزینه‌های کاری آینده در بخش ۶ مشخص شده است.

۲- پیشنهادی پژوهشی

تحقیقات زیادی در زمینه امنیت سایبری که در آن الگوریتم‌های یادگیری ماشین برای تشخیص ناهنجاری و حمله استفاده می‌شود، انجام شده است. مروری کوتاه بر برخی از مقالات که مربوط به این کار است در بخش زیر ارائه شده است.

۲-۱. برخی از روش‌های موجود در تشخیص نفوذ خانه‌های هوشمند

همان‌طور که در بالا ذکر شد، تشخیص نفوذ و ناهنجاری مبتنی بر یادگیری ماشین، یک حوزه تحقیقاتی فعال است. برخی از الگوریتم‌های رایج مورد استفاده عبارت‌اند از: K-نزدیک‌ترین همسایه (KNN)، ماشین بردار پشتیبان، ادراک چندلایه^۴ (MLP)، شبکه عصبی مصنوعی و غیره [۶].

استفاده از یک روش طبقه‌بندی برای بهبود تشخیص حمله

هوشمند ممکن است دارای یک میکروفون باشد که ممکن است سیستم اینترنت اشیا را نسبت به تهدیدات امنیتی و حریم خصوصی آسیب‌پذیر کند. یک فرد ساکن عادی ممکن است دانش و مهارت لازم برای تفسیر و محافظت از سیستم اینترنت اشیا را نداشته باشد. به همین دلیل، نیاز به یک راهکار خودکار برای حمایت از خانه‌های هوشمند در برابر تهدیدهای امنیتی و حریم خصوصی حائز اهمیت است.

در سال‌های اخیر، تلاش‌های زیادی برای تحقیق در زمینه سیستم‌های تشخیص نفوذ خودکار با استفاده از یادگیری ماشین^۱ (ML) صورت گرفته است. این سیستم‌ها بر اساس امضا، الگو و ویژگی‌های حملات، آموزش دیده و با رفتار ترافیک ورودی مقایسه می‌شوند. از الگوریتم‌های متعددی مانند جنگل تصادفی، درخت تصمیم، ماشین بردار پشتیبان^۲ (SVM) و شبکه‌های عصبی مصنوعی^۳ (ANN) برای این منظور استفاده می‌شود [۳]. ANN یکی از الگوریتم‌های یادگیری ماشین است که به دلیل ماهیت خود که یک جعبه سیاه است و نیاز به تطبیق با یک سیستم زیربنایی دارد، در موارد متعددی در حوزه‌های مختلف برای کارهای طبقه‌بندی استفاده شده است. این ویژگی باعث می‌شود تا این روش برای شناسایی نفوذ در ترافیک شبکه با مجموعه‌های داده با ابعاد بالا مناسب باشد، زیرا توانایی بالای آن در ساخت مدل و شناسایی الگوها به خوبی در این موارد مورد استفاده قرار می‌گیرد. انتخاب بهینه پارامترها در این روش در هر مجموعه داده بسیار حائز اهمیت است.

یک روش جدید برای تشخیص حملات در این مقاله پیشنهاد شده است که به نتایج بهتری نسبت به سایر روش‌های مشابه دست یافته است. انتخاب بهترین ویژگی‌ها بر مبنای PCA [۴ و ۵] هنگام کاهش حجم داده‌ها نقش مهمی در این روش ایفا می‌کند. علاوه بر این، از الگوریتم جنگل تصادفی در این روش برای طبقه‌بندی داده‌ها استفاده شده است که یک الگوریتم یادگیری ماشین است. در اینجا، یک رویکرد ساده و قابل مقیاس برای تشخیص حملات امنیتی با استفاده از دقت تشخیص، سربار محاسباتی و محدودیت زمانی ارائه شده است. دقت پیش‌بینی اهمیت زیادی دارد، زیرا هرگونه اشتباه در تشخیص حمله می‌تواند عواقب جدی به دنبال داشته باشد. سربار محاسباتی نیز نقش مهمی در این روش ایفا می‌کند، زیرا یافتن بهترین راه‌حل با استفاده از منابع محاسباتی فراوان ممکن است کارآمد و مقرون به صرفه نباشد. زمان نیز یک محدودیت در این روش است؛ زیرا تشخیص باید به سرعت صورت بگیرد و هر مدلی که برای آموزش یا تشخیص، زمان زیادی نیاز داشته باشد، برای تشخیص

¹ Machine Learning

² Support Vector Machine

³ Artificial Neural Networks

⁴ MultiLayer Perceptron

انجام شده است و چندین مجموعه داده ساخته شده است؛ مانند UNSWNB15 [۱۵] و IoTID20 [۱۶] که مشخصات کلی داده‌های این دو مجموعه داده در قالب عادی یا حمله در جدول ۱ خلاصه شده است. جدول ۲ و ۳ نیز به ترتیب نشان‌دهنده مشخصات حملات در قالب چند کلاسه برای مجموعه داده UNSWNB15 و IoTID20 است. برای آموزش و آزمایش مدل پیشنهادی، از مجموعه داده‌های ذکر شده استفاده شده است. مجموعه داده IoTID20 جزو جدیدترین و به‌روزترین مجموعه داده‌های جمع‌آوری شده در محیط اینترنت اشیا است که می‌تواند ویژگی‌های ترافیک شبکه واقعی‌تری را نمایش دهد. باید توجه داشت که به‌تازگی این مجموعه داده جمع‌آوری شده و برای استفاده در تحقیقات به‌راحتی در دسترس قرار گرفته است، اما تعداد کمی از تحقیقات از این مجموعه داده برای آزمایش استفاده کرده‌اند. علاوه بر این، تا کنون تحقیقات موجود، طبقه‌بندی زیرمجموعه‌های این مجموعه داده را با دقت‌های پایین انجام داده‌اند. باتوجه به این وضعیت، این مجموعه داده مناسب برای تجزیه و تحلیل مدل تشخیص نفوذ IoT بوده و امکان بهبود آن وجود دارد.

جدول (۱). مشخصات مجموعه داده‌های استفاده شده (دودویی)

مجموعه داده	نوع داده	مجموعه آموزش	مجموعه آزمایش
UNSW-NB15	عادی	۵۶۰۰۰	۳۷۰۰۰
	حمله	۱۱۹۳۴۱	۴۵۳۳۲
	کل	۱۷۵۳۴۱	۸۲۳۳۲
IoTID20	عادی	۳۲۰۵۸	۸۰۱۵
	حمله	۴۶۸۲۷۴	۱۱۷۰۶۸
	کل	۵۰۰۳۳۲	۱۲۵۰۸۳

تحقیق‌هایی انجام شده است که جمع‌آوری و آزمایش مجموعه داده IoTID20 را با استفاده از چندین رویکرد یادگیری ماشین انجام می‌دهند که در آن‌ها تشخیص فعالیت‌های غیرعادی دودویی و چند کلاسه مورد توجه قرار گرفته است. برای این منظور، الگوریتم‌های مختلفی مانند Gaussian NB، SVM، Logistic Regression، Decision Tree و Random Forest برای

می‌تواند مفید باشد، چرا که الگوریتم‌های مختلف ممکن است به طرق متفاوت به مسائل پاسخ دهند. به‌عنوان مثال، استفاده از الگوریتم جنگل تصادفی و معیار توزیع میانگین می‌تواند برای حل وابستگی ویژگی‌ها و بهبود دقت در طبقه‌بندی ترافیک شبکه استفاده شود.

در مقاله [۷]، یک روش طبقه‌بندی گروهی مبتنی بر تجزیه و تحلیل خوشه‌ها و استفاده از الگوریتم ADTree و KNN ارائه شده که نشان‌دهنده دقت بالای تشخیص حمله در سیستم‌های شناسایی نفوذ است. در مقاله [۸]، روش Bagging همراه با راهکار REPTree استفاده شده است که می‌تواند منجر به دقت بالای طبقه‌بندی شود. همچنین روش Bagging زمان کمتری برای ساخت مدل نیاز دارد. در مقاله [۹]، یک رویکرد امنیتی الهام گرفته از مفاهیم زیست‌شناسی ارائه شده که بیانگر تشخیص سریع دستگاه‌های مختلف در شهرهای هوشمند هنگام استفاده از دستگاه‌های اینترنت اشیا است. امنیت در خانه هوشمند همچنین می‌تواند توسط یک سیستم نظارت کامل انجام شود. این سیستم قادر به شناسایی خرابی سیستم‌های فیزیکی، پیش‌بینی اثرات شرایط آب‌وهوایی سخت و شناسایی رویدادهای امنیتی مشکوک است. این سیستم از حسگرها و منطق فازی برای نظارت فعال بر اینترنت اشیا نسل بعدی استفاده می‌کند [۱۰]. این سیستم می‌تواند با استفاده از داده‌های حسگر فیزیکی و اطلاعات کاربر برای کنترل محیط خانه برای ایجاد راحتی و امنیت بالا استفاده شود.

در [۱۱]، هدف اصلی ساخت مدل‌های یادگیری ماشین به‌منظور شناسایی حملات در شبکه اینترنت اشیا با استفاده از درخت تصمیم است. نتایج آزمایش‌های انجام شده در کار تحقیقاتی [۱۱ و ۱۲] نشان می‌دهد که درخت تصمیم بالاترین نرخ دقت با میزان ۹۸٫۲۳٪ در بین همه مدل‌های طبقه‌بندی‌کننده از جمله SVM، Naïve Bayes و Adaboost-based دارد. سرور و همکاران [۱۳] یک رویکرد پیشرفته برای طراحی سیستم تشخیص نفوذ برای شبکه‌های IoT با طبقه‌بندی‌کننده الگوریتم بهینه‌سازی ازدحام ذرات جنگل تصادفی برای انتخاب ویژگی ارائه کردند، سیستم پیشنهادی آنها به دقت ۹۸ درصد برای طبقه‌بندی دودویی و ۸۳ درصد برای طبقه‌بندی چند کلاسه در مجموعه داده IoTID20 دست پیدا کرد. در [۱۳ و ۱۴]، نویسندگان در سه مرحله کلیدی عمل تشخیص نفوذ را انجام دادند: پیش‌پردازش، انتخاب ویژگی، مدل‌سازی و ارزیابی سیستم پیشنهادی.

۲-۲. مجموعه داده IoT

زمانی که الگوریتم‌های یادگیری ماشین استفاده می‌شوند تا حملات را تشخیص دهند، مجموعه داده‌ها بسیار اهمیت پیدا می‌کنند؛ زیرا بدون این مجموعه داده‌ها، آموزش و آزمایش این الگوریتم‌ها امکان‌پذیر نیست. در این زمینه، تحقیقات مختلفی

طبقه‌بندی چند کلاسه توسط این تحقیق صورت نگرفته است و به همین دلیل، اثربخشی این رویکرد برای طبقه‌بندی چند کلاسه در مجموعه داده IoTID20 قابل ارزیابی نیست. در مقاله دیگری نیز ارزیابی مجموعه داده‌های متقابل با در نظر گرفتن مجموعه داده‌های IoTID20 و Bot-IoT بحث شده است [۱۹].

جدول (۳). مشخصات مجموعه داده IoTID20 (چند کلاسه)

نوع داده	داده آموزشی	داده آزمایش	کل داده‌ها
Normal	۳۲۰۵۸	۸۰۱۵	۴۰۰۷۳
Mirai-UDP Flooding	۱۴۶۵۵۱	۳۶۶۳۸	۱۸۳۱۸۹
Mirai-Hostbruteforce	۹۶۹۴۳	۲۴۲۳۵	۱۲۱۱۷۸
Mirai-HTTP Flooding	۴۴۶۵۴	۱۱۱۶۴	۵۵۸۱۸
Mirai-Ackflooding	۴۴۰۹۹	۱۱۰۲۵	۵۵۱۲۴
DoS-Synflooding	۴۷۵۱۳	۱۱۸۷۸	۵۹۳۹۱
Scan Port OS	۴۲۴۵۸	۱۰۶۱۵	۵۳۰۷۳
Scan Hostport	۱۷۷۵۴	۴۴۳۸	۲۲۱۹۲
MITM ARP Spoofing	۲۸۳۰۲	۷۰۷۵	۳۵۳۷۷

با این حال، کلیه تشخیص‌ها فقط برای سطح دودویی صورت گرفته‌اند و طبقه‌بندی سطح زیرشاخه توسط نویسندگان مورد نظر قرار نگرفته است.

۳- راه کار پیشنهادی

در این بخش، راهکار پیشنهادی برای تشخیص نفوذ که ترکیبی از انتخاب بهترین ویژگی‌ها و استفاده از دسته‌بندی جنگل تصادفی برای دسته‌بندی داده‌ها است شرح داده می‌شود.

۳-۱. مراحل راه کار ارائه شده

روش پیشنهادی از سه بخش تشکیل شده است: پیش‌پردازش مجموعه داده، انتخاب بهترین ویژگی‌ها، طبقه‌بندی و دسته‌بندی داده‌ها. در ابتدا، مجموعه داده برای حذف هرگونه داده اضافی و مقادیر از دست رفته پیش‌پردازش شده است. پس از آن، روش انتخاب ویژگی مبتنی بر PCA به منظور یافتن زیرمجموعه‌ای از ویژگی‌هایی که می‌توانند به طور مؤثرتر برای طبقه‌بندی استفاده شوند، اعمال شده است.

سپس داده‌ها به جنگل تصادفی ارسال می‌شوند تا کلاس داده‌ها (عادی یا غیرعادی) مشخص شود. در مرحله نهایی، نتایج ارزیابی روش پیشنهادی برای هر دو ترافیک عادی و حمله روی مجموعه داده‌های ذکر شده در بخش قبل هم به صورت دودویی و

تشخیص حملات IoT اجرا شده و نتایج متفاوتی به دست آمده است. در حوزه امنیت، داشتن دقت بالا در تشخیص حمله بسیار حائز اهمیت است و سیستم‌های تشخیص نفوذ با دقت کمتر از ۹۰ درصد ممکن است کاملاً قابل اعتماد نباشند؛ بنابراین، می‌توان گفت که با استفاده از این مجموعه داده، فضایی برای بهبود مدل یادگیری ماشین وجود دارد تا به دقت بالاتری در تشخیص حمله مبتنی بر زیرمجموعه‌ها دست یابیم.

جدول (۲). مشخصات مجموعه داده UNSWBN15 (چند کلاسه)

نوع داده	داده آموزشی	داده آزمایش	کل داده‌ها
Normal	۵۶۰۰۰	۳۷۰۰۰	۹۳۰۰۰
Analysis	۲۰۰۰	۶۷۷	۲۶۷۷
Backdoor	۱۷۴۶	۵۸۳	۲۳۲۹
DoS	۱۲۲۶۴	۴۰۸۹	۱۶۳۵۳
Exploits	۳۳۳۹۳	۱۱۱۳۲	۴۴۵۲۵
Fuzzers	۱۸۱۸۴	۶۰۶۲	۲۴۲۴۶
Generic	۴۰۰۰۰	۱۸۸۷۱	۵۸۸۷۱
Reconnaissance	۱۰۴۹۱	۳۴۹۶	۱۳۹۸۷
Shellcode	۱۱۳۳	۳۷۸	۱۵۱۱
Worms	۱۳۰	۴۴	۱۷۴

در مقاله [۱۷]، اهمیت انتخاب ویژگی در رویکردهای تشخیص نفوذ بر اساس تحلیل ناهنجاری به شدت مورد توجه قرار گرفته و انتخاب ویژگی با استفاده از الگوریتم جنگل تصادفی بر روی داده‌های مجموعه IoTID20 انجام گرفته است. نویسندگان، نتایج را تنها برای سه دسته حمله که دقت بالایی بیش از ۹۹٫۹٪ دارند ارائه داده‌اند. دقت در تشخیص برخی از دسته‌های حمله بسیار بالا است که نشان می‌دهد رویکرد انتخاب ویژگی مورد نظر توانایی بالایی دارد و تنها از زیرمجموعه‌ای از ویژگی‌های انتخاب شده استفاده می‌کند. با این حال، استفاده از چنین رویکردی تا زمانی که نتایج کاملی برای تمامی دسته‌ها و زیرمجموعه‌های حمله ارائه و تجزیه و تحلیل نشود قابل توجیه نیست.

در مقاله [۱۸] پیشنهاد شده است که در تشخیص بهتر ناهنجاری در مجموعه داده‌های نامتعادل مانند IoTID20، از یک رویکرد چندمرحله‌ای استفاده شود. نویسندگان در این مقاله از روش‌هایی مانند نمونه‌برداری و خوشه‌بندی K-means++ استفاده کرده‌اند تا عدم تعادل در بین نمونه‌ها برای دسته‌های مختلف تأثیر کمتری در نتایج داشته باشد. نتایج این مقاله صرفاً برای طبقه‌بندی دودویی دقت مناسبی را نشان داده است. اما

$$Z = \frac{X - \mu}{\sigma} \quad (۱)$$

۳-۳. مرحله انتخاب بهترین ویژگی‌ها

باتوجه به ابعاد بالای مجموعه داده، PCA به پایین آوردن ابعاد مجموعه به سطح مطلوب کمک می‌کند. PCA [۴ و ۵] به عنوان یک تبدیل خطی متعامد تعریف می‌شود که داده‌ها را به یک سیستم مختصات جدید تبدیل می‌کند، به طوری که بیشترین واریانس توسط برخی از پیش‌بینی‌های اسکالر داده‌ها روی مختصات اول (موسوم به اولین جزء اصلی) قرار می‌گیرد، دومین واریانس بزرگ در مختصات دوم و الی آخر که الگوریتم آن در شکل ۱ نشان داده شده است.

ماتریس داده X که دارای n سطر و p ستون است با میانگین تجربی صفر در ستون در نظر بگیرد (میانگین نمونه هر ستون به صفر منتقل شده است) که در آن هر یک از n ردیف نشان‌دهنده تکرار متفاوت آزمایش است و هر یک از ستون‌های p نوع خاصی از ویژگی‌ها را (مثلاً نتایج به دست آمده از یک آزمون) نشان می‌دهد. از نظر ریاضی، تبدیل با مجموعه‌ای از اندازه l از بردارهای p -بعدی از وزن‌ها یا ضرایب $w_k = (w_1, \dots, w_p)_{(k)}$ تعریف می‌شود که هر بردار ردیف X_i از X را به بردار جدیدی از مؤلفه اصلی نگاشت می‌کند:

$$t_k(i) = x_i \cdot w_k, i = 1, \dots, n, k = 1, \dots, l \quad (۲)$$

به گونه‌ای که متغیرهای منفرد $t_1, \dots, t_l$ از t در نظر گرفته شده بر روی مجموعه داده، به طور متوالی حداکثر واریانس ممکن را از X به ارث می‌برند، با هر بردار ضریب w محدود به یک بردار واحد (که در آن l معمولاً به شدت کمتر از p در نظر گرفته می‌شود) برای کاهش ابعاد انتخاب می‌شود

Algorithm PCA

Parameters: Step size η , Number of epoch S , Number of Dimension K

Input: Data matrix $X = (x_1, \dots, x_N) \in \mathbb{R}^{M \times N}$

Initial Matrix $W_0 = (w_1, \dots, w_K)$

(diagonal element is 1, otherwise 0) $\in \mathbb{R}^{M \times K}$

Diagonal matrix $D = \text{Diag}(K, \dots, 1) / 10^5 \in \mathbb{R}^{K \times K}$

Output: Eigenvectors $W \in \mathbb{R}^{K \times M}$

for $t = 1, 2, \dots, S$ do

$$W_t' = W_{t-1} + \frac{\eta}{N} \sum_{i=1}^N x_i (x_i^T W_{t-1} D)$$

$$W_t = \text{qr}(W_t')$$

end for

شکل (۱). الگوریتم شبه کد PCA [۵]

جدول (۴). سطح تغییرات مؤلفه‌های اصلی انتخابی مجموعه داده‌ها

مجموعه داده	کمترین سطح نسبی تغییرات	بیشترین سطح نسبی تغییرات
UNSW-NB15	۰,۰۱۱	۰,۲۴۵

هم چند کلاسه نشان داده می‌شود. نکته مهم این است که نه تنها دقت پیش‌بینی، بلکه سربار محاسباتی و پیچیدگی زمانی نیز برای اعمال این رویکرد در سناریوی خانه هوشمند که نیاز به شناسایی بلادرنگ دارد و ممکن است به سرویس‌دهنده‌های پیشرفته دسترسی نداشته باشد در نظر گرفته می‌شوند. بخش‌های زیر به تشریح فازهای تشخیص حمله می‌پردازد.

۳-۲. مرحله پیش‌پردازش

برای دو مورد مجموعه داده‌ای که در این پژوهش مورد استفاده قرار گرفته‌اند، مرحله پیش‌پردازش به تبدیل ویژگی‌ها می‌پردازد. در اینجا، عادی‌سازی و استانداردسازی داده‌ها به عنوان مرحله پیش‌پردازش اعمال شده است. باتوجه به ناتوانی الگوریتم‌های یادگیری ماشین در مدیریت ویژگی‌های دسته‌بندی، در وهله اول، عادی‌سازی داده‌ها باید انجام شود. از آنجایی که هر دو مجموعه داده UNSWNB15 و IoTID20 دارای ۳ ویژگی از ویژگی‌های طبقه‌بندی (به عنوان مثال نوع پروتکل، سرویس و پرچم) از مجموع ۴۲ و ۸۳ ویژگی خود هستند، برچسب‌گذاری برای تبدیل مقادیر غیر عددی به شکل عددی استفاده شده است. روشی که برای این تبدیل عددی استفاده می‌شود این گونه است که برای هر مقدار از مقادیر ممکن برای ویژگی‌ها یک بعد جداگانه در نظر می‌گیرد. اما مشکل در مورد این نوع برچسب‌گذاری این است که تعداد زیادی از ابعاد جدید را در زمان تخصیص داده‌های اسمی به بردارهای دودویی ایجاد می‌کند. در نتیجه، شکل ورودی مدل برای انجام برخی عملیات بسیار پیچیده می‌شود؛ بنابراین به عنوان یک راهکار پیشگیرانه برای این مشکل، به هر ویژگی طبقه‌بندی شده در مجموعه داده، یک مقدار اعشاری منحصر به فرد اختصاص داده شده است. به این نحو که در ابتدا، برای هر ویژگی اسمی، مقادیر آن بر اساس حروف الفبا مرتب می‌شوند و پس از آن، مقادیر اسمی مرتب شده با تخصیص مقادیر اعشاری خاص به هر متغیر در بازه ۱ تا طول لیست، به شکل عددی تبدیل می‌شوند.

در حال حاضر، هدف از مقیاس‌بندی ویژگی این است که اطمینان حاصل شود که همه ویژگی‌های مجموعه داده در مقیاس تقریباً یکسانی هستند و آن را برای اکثر تحلیل‌ها آسان‌تر می‌کند. در این کار از استانداردسازی برای هدف مقیاس‌بندی ویژگی‌ها استفاده شده است که نتایج مؤثری در مقایسه با رویکرد نرمال‌سازی حداقل - حداکثر را ارائه می‌کند. پس از انجام برچسب‌گذاری، استانداردسازی صورت می‌گیرد به طوری که تمام ویژگی‌ها مجدداً مقیاس می‌شوند تا میانگین و انحراف معیار به ترتیب ۰ و ۱ تضمین شود. رابطه ۱ مربوط به این استانداردسازی است که در آن μ میانگین مقادیر ویژگی‌ها و σ انحراف معیار ویژگی‌ها است و بر اساس مقادیر قبلی ویژگی‌ها و رابطه زیر، مقادیر جدید پس از استانداردسازی به دست می‌آیند.

مبنای دسته‌بندی مجموعه داده به چند بخش و به دست آوردن نمونه‌های آموزشی با نمونه‌برداری از مجموعه داده اصلی، می‌توان به تقویت نمونه‌های مورد استفاده برای آموزش دست یافت. باتوجه به اینکه مجموعه داده‌ها اغلب دارای ویژگی‌هایی با مقادیر غیر متوازن (اغلب مقادیر یک ویژگی از یک کلاس خاص هستند) هستند، چنانچه از رویکردهای مشابه برای دسته‌بندی تمام ویژگی‌ها استفاده نماییم، در آن صورت تمام ویژگی‌ها با یک رویکرد یکسان بررسی می‌شوند و این موضوع باعث می‌شود که در نهایت، روش پیشنهادی، کار آیی لازم را نداشته باشد. این رویکرد به طور مؤثر احتمال بیش از حد برازش را کاهش می‌دهد. مجموعه داده به دست آمده توسط عملیات فوق به یک مدل درخت تصمیم برای آموزش وارد می‌شود و مدل نهایی برای تولید یک مدل طبقه‌بندی جنگل ترکیب می‌شود. این مدل نتایج را با بالاترین نرخ صحیح پیش‌بینی می‌کند.

با این حال، هنوز جای زیادی برای بهبود در مدل جنگل تصادفی وجود دارد. این مسئله شامل بهبود توانایی طبقه‌بندی درخت تصمیم در جنگل، بهینه‌سازی بیشتر همبستگی بین درخت‌های تصمیم‌گیری در مدل جنگل ترکیبی و بهینه‌سازی روش رأی‌گیری اتخاذ شده در فرآیند تعیین نتیجه است. یک مدل ترکیبی نسبتاً خوب باید ویژگی‌های زیر را داشته باشد: توانایی تصمیم‌گیری خوب در طبقه‌بندی‌کننده‌ها و همبستگی اندک بین طبقه‌بندی‌کننده‌ها. در این مقاله ابتدا طبقه‌بندی‌کننده‌هایی با عملکرد بهتر انتخاب شده‌اند که این مشخصه با بررسی سطح زیر نمودار ROC به دست آمده است. سپس عمل بهینه‌سازی درختان با حذف شاخه‌های زائد انجام شد. در نهایت، درختان بهینه به دست آمده از این مرحله با یکدیگر ترکیب شده و آماده انجام دسته‌بندی داده‌ها هستند. الگوریتم استفاده شده در شکل ۲ نشان داده شده است.

جدول (۵). ماتریس درهم‌تنیدگی

کلاس واقعی	کلاس پیش‌بینی	
	عادی	غیرعادی
عادی	TN	FP
غیرعادی	FN	TP

۴- ارزیابی روش پیشنهادی

همان‌طور که قبلاً ذکر شد، کار صورت گرفته در این مقاله با استفاده از دو مجموعه داده برای تشخیص نفوذ در شبکه‌های خانه هوشمند انجام شده است. مشاهدات تجربی برای برچسب‌های موجود در هر یک از این مجموعه داده‌ها انجام شده است. برای

۰,۲۵۷	۰,۰۳۲	IoTID20
-------	-------	---------

پس از به دست آوردن مؤلفه‌های اصلی ویژگی‌ها، حداقل تعداد ویژگی‌ها بر اساس سطح تغییراتی که دارند انتخاب می‌شوند. به دلیل اینکه PCA روشی بدون نظارت است برای یافتن اجزای اصلی از برچسب کلاس استفاده نمی‌کند که این باعث می‌شود واریانس داده‌ها به حداکثر مقدار ممکن برسد. در نتیجه، در هر دو مجموعه داده، مؤلفه‌های به دست آمده و سطح تغییرات متناظر آن‌ها بدون تغییر باقی می‌مانند.

در ابتدا، نسبت تغییرات توضیح داده شده برای هر دو مجموعه داده UNSW-NB15 و IoTID20 به ترتیب با استفاده از ۴۲ و ۸۳ ویژگی واقعی به دست آمده است. اهمیت یک جزء با مقدار تغییراتی که نشان می‌دهد افزایش می‌یابد. مجموع سطح تغییرات تمام ویژگی‌های یک مجموعه داده برابر یک می‌شود. ۱۷ مؤلفه اصلی اول بیش از ۹۰ درصد از داده‌های مجموعه داده UNSW-NB15 را در بر می‌گیرند. برای مجموعه داده IoTID20، ۲۸ مؤلفه اصلی اول بیش از ۸۵ درصد داده‌ها را شامل می‌شود. به عبارت دیگر، این مؤلفه‌های اصلی انتخاب شده چیزی حدود ۰,۸۵ یا بیشتر به عنوان نماینده مجموعه داده متناظرشان خلاصه می‌شوند. کمترین و بیشترین سطح نسبی تغییرات مؤلفه‌های انتخاب شده برای هر دو مجموعه داده در جدول ۴ نشان داده شده است. بنابراین، این ۱۷ و ۲۸ ویژگی به عنوان ویژگی‌های بهینه برای جنگل تصادفی انتخاب شده‌اند.

Algorithm Random Forest
Precondition: A training set $S := (x_1, y_1), \dots, (x_n, y_n)$, features F , and number of trees in forest B .

```

function RANDOMFOREST( $S, F$ )
     $H \leftarrow \emptyset$ 
    for  $i \in 1, \dots, B$  do
         $S^{(i)} \leftarrow$  A bootstrap sample from  $S$ 
         $h_i \leftarrow$  RANDOMIZEDTREELEARN( $S^{(i)}, F$ )
         $H \leftarrow H \cup \{h_i\}$ 
    end for
    return  $H$ 
end function

function RANDOMIZEDTREELEARN( $S, F$ )
    At each node:
         $f \leftarrow$  very small subset of  $F$ 
        Split on best feature in  $f$ 
    return The learned tree
end function

```

شکل (۲). الگوریتم جنگل تصادفی برای دسته‌بندی داده‌ها

۳-۳. مرحله دسته‌بندی داده‌ها با استفاده از جنگل

تصادفی

الگوریتم جنگل تصادفی در مقایسه با سایر الگوریتم‌های طبقه‌بندی از دقت تشخیص نسبتاً بالایی برخوردار است و نسبت به نمونه‌های دارای نویز تحمل‌پذیری بیشتری دارد که منجر به انجام مطالعات نظری و تجربی متعدد با تمرکز بر استفاده از این الگوریتم شده است. به عنوان یک الگوریتم طبقه‌بندی‌کننده بر

۴. **نرخ منفی نادرست (FNR):** این معیار نسبت نمونه‌های حمله‌ای را که به اشتباه به‌عنوان نمونه‌های عادی توسط مدل برچسب‌گذاری شده‌اند، ارزیابی می‌کند.

$$FNR = \frac{FN}{TP+FN} \quad (۶)$$

۵. **خاص بودن (Specificity):** این معیار همچنین به عنوان نرخ منفی واقعی (TNR) شناخته می‌شود و بیانگر بخشی از داده‌های عادی است که به درستی توسط مدل برچسب‌گذاری شده است.

$$Specificity = \frac{TN}{TN+FP} \quad (۷)$$

۶. **دقت منفی (Negative Precision):** این معیار نرخ نمونه‌های عادی را در تمام نمونه‌های مجموعه آزمایش نشان می‌دهد.

$$Negative\ Precision = \frac{TN}{TN+FN} \quad (۸)$$

۷. **نرخ خطا (Error Rate):** این معیار، اندازه‌گیری تشخیص نادرست توسط مدل است.

$$Error\ Rate = \frac{FP+FN}{TP+FP+TN+FN} \quad (۹)$$

۸. **ضرب همبستگی ماتیو (MCC):** این معیار برای ارزیابی اختلاف بین مقادیر پیش‌بینی شده و واقعی استفاده می‌شود.

$$MCC = \frac{TN*TP-FN*FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (۱۰)$$

۲-۴. معیارهای ارزیابی در طبقه‌بندی چند کلاسه

برخلاف طبقه‌بندی دودویی با دو برچسب (Normal و Attack)، طرح تشخیص چند کلاسه شامل تعدادی برچسب تشخیص است. به‌عنوان مثال، UNSW-NB15 شامل ۹ نوع حمله مختلف با داده‌های برچسب عادی اضافی است. درحالی‌که IoTID20 در مجموع دارای ۸ ویژگی برچسب‌گذاری شده به‌جز رفتار عادی ترافیک است؛ بنابراین معیارهای ارزیابی برای تشخیص چند کلاسه در مقایسه با مدل تشخیص دودویی معمولی تا حدودی متفاوت است. با پرداختن به چنین تفاوت‌هایی، در مجموع ۴ معیار ارزیابی را برای عملکرد تشخیص چند کلاسه مدل خود بیان کرده‌ایم. در زیر به آن‌ها و معادله‌های آن‌ها اشاره شده است.

۱. **دقت وزن‌دار (Weighted Precision):** این معیار، میانگین وزنی دقت را با در نظر گرفتن وزن‌ها برابر با احتمال کلاس تفسیر می‌کند.

$$Precision_w = \frac{\sum_{i=1}^l Precision_i * TestSet_i}{Test\ Set} \quad (۱۱)$$

۲. **فراخوانی وزن‌دار (Weighted Recall):** این معیار، میانگین وزنی فراخوانی را با در نظر گرفتن وزن‌ها برابر با احتمال کلاس تفسیر می‌کند.

$$Recall_w = \frac{\sum_{i=1}^l Recall_i * TestSet_i}{Test\ Set} \quad (۱۲)$$

طبقه‌بندی، هر دو مجموعه داده حاوی برچسب‌های کلاس یکسانی هستند. یکی نشان‌دهنده ترافیک عادی است و دیگری نشان‌دهنده ترافیک غیرعادی (حمله) است. با این حال، UNSW-NB15 دارای ۱۰ برچسب طبقه‌بندی شده برای تشخیص چند کلاسه در مقایسه با IoTID20 است که دارای ۹ کلاس برچسب‌گذاری شده است. برای هر کلاس برچسب‌گذاری شده، معیارهای عملکرد روش پیشنهادی با و بدون در نظر گرفتن PCA برای نشان دادن اثر انتخاب بهترین ویژگی‌ها در مجموعه‌های داده ذکر شده ارزیابی شده‌اند. در زیر بخش ۴-۱، چشم‌انداز مختصری از معیارهای ارزیابی برای طبقه‌بندی دودویی ارائه شده است که برای ارزیابی تشخیص استفاده می‌شود. درحالی‌که بخش ۴-۲ معیارها را برای تشخیص چند کلاسه نشان می‌دهد.

۴-۱. معیارهای ارزیابی در طبقه‌بندی دودویی

همان‌طور که در جدول ۵ نشان داده شده است، ۴ نتیجه‌ی ممکن وجود دارد که از یک ماتریس درهم‌تنیدگی (برای طبقه‌بندی دودویی) قابل دستیابی هستند. بر اساس این ۴ نتیجه از ماتریس مورد نظر یعنی مثبت واقعی (TP)، منفی واقعی (TN)، مثبت کاذب (FP) و منفی کاذب (FN)، نتایج کلی با استفاده از ۸ معیار ارزیابی شناخته شده تجزیه و تحلیل شده‌اند.

TP (True Positive) ویژگی‌های حمله که به‌درستی توسط مدل برچسب‌گذاری شده‌اند.

FP (False Positive) ویژگی‌های غیر حمله که توسط مدل به‌اشتباه برچسب‌گذاری شده‌اند.

TN (True Negative) ویژگی‌های غیر حمله که به‌درستی توسط مدل برچسب‌گذاری شده‌اند.

FN (False Negative) ویژگی‌های حمله که توسط مدل به‌اشتباه برچسب‌گذاری شده‌اند.

در زیر تعریف کوتاهی از هر یک از این ۱۰ معیار به همراه معادله‌ی مربوط به آنها ارائه شده است.

۱. **دقت (Accuracy):** این معیار در واقع کسری شامل تقسیم نمونه‌های درست طبقه‌بندی شده بر تمام نمونه‌های مجموعه آزمایش است.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (۳)$$

۲. **امتیاز F1 (F1 Score):** این معیار، میانگین هارمونیک درستی (Recall) و فراخوانی (Precision) را برای یک مدل خاص تفسیر می‌کند.

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (۴)$$

۳. **نرخ هشدار نادرست (FAR):** این معیار به نسبت نمونه‌های عادی اشاره دارد که به اشتباه به‌عنوان نمونه‌های حمله توسط مدل شناسایی شده‌اند.

$$FAR = \frac{FP}{TN+FP} \quad (۵)$$

کرد که به استثنای Specificity برای مجموعه داده UNSWNB15 و Negative Precision برای مجموعه داده IoTID20، بقیه‌ی مقادیر پارامتر برای مجموعه‌ها تا حدودی افزایش یافته است. با این حال، در کنار کاهش پیچیدگی، به دلیل کاهش اطلاعات، خطر کاهش عملکرد در هنگام استفاده از PCA وجود دارد اما از سویی، میزان زمان محاسباتی نیز کاهش یافته است.

جدول (۶). نتایج آزمایش دو کلاسه روی مجموعه داده‌ها

IoTID20		UNSW-NB15		معیار (%)
بدون PCA	با PCA	بدون PCA	با PCA	
۹۸٫۸۷	۹۹٫۰۲	۹۴٫۵۴	۹۹٫۷۳	Accuracy
۸۶٫۵۴	۹۰٫۰۶	۹۰٫۰۶	۹۸٫۳۷	F1-Score
۱۸٫۲۲	۱۶٫۴۳	۱۷٫۳۷	۱۴٫۲۲	FAR
۹۰٫۳۶	۹۶٫۵۵	۹۴٫۶۹	۹۲٫۸۷	Specificity
۱۱٫۲۷	۸٫۳۵	۱۰٫۵۳	۶٫۴۲	FNR
۹۸٫۳۲	۹۶٫۶۵	۹۳٫۹۱	۹۹٫۴۷	Negative Precision
۱٫۱۳	۰٫۹۸	۵٫۴۶	۰٫۲۷	Error Rate
۱۰٫۷۲	۸٫۳۶	۷۷٫۴۳	۸۹٫۴۶	MCC

جدول (۷). زمان محاسبات آموزش و آزمایش دودویی

زمان آموزش (ثانیه)	زمان آزمایش (ثانیه)	مجموعه داده	
۹۴۳	۳۴۲	با PCA	UNSWNB15
۲۱۱۵	۹۰۶	بدون PCA	
۱۷۹۴	۸۳۰	با PCA	IoTID20
۳۲۵۸	۲۰۸۳	بدون PCA	

زمان محاسباتی صرف شده برای هر یک از مجموعه داده‌ها برای دسته‌بندی دودویی در جدول ۷ نشان داده شده است. در اینجا، کاهش قابل توجهی را می‌توان پس از کاهش ابعاد از طریق PCA مشاهده کرد. از آنجایی که مجموعه داده IoTID20 مقایسه با UNSW-NB15 شامل ویژگی‌های بیشتری است، زمان محاسبات بیشتری برای این مدل صرف می‌شود.

در داده کاوی، منحنی ROC، یک روش گرافیکی برای نشان دادن تعادل بین نرخ مثبت واقعی و نرخ مثبت کاذب مدل

۳. F1-Score وزن دار (Weighted F1 Score): این معیار میانگین وزنی امتیاز F1 محاسبه شده را با در نظر گرفتن میانگین نمره هر کلاس F1 در نظر گرفته شده تعیین می‌کند.

$$F1\ Score_w = \frac{\sum_{i=1}^l F1\ Score_i * TestSet_i}{Test\ Set} \quad (۱۳)$$

۴. نرخ خطای کلی (Overall Error Rate): این معیار بیانگر ارزیابی کل پیش‌بینی‌های نادرست مجموعه آزمون ساخته شده توسط مدل است.

$$Overall\ Error\ Rate = \frac{FP+FN}{Test\ Set} \quad (۱۴)$$

۵- نتایج آزمایش

در این مقاله تمامی شبیه‌سازی‌ها با استفاده از نرم‌افزار شبیه‌سازی متلب انجام شده است. تمامی وظایف عملیاتی این کار در محیط سخت‌افزاری زیر اجرا شده است: پردازنده مرکزی Intel core i7-3400U (۷۰.۳ گیگاهرتز)، ۴ گیگابایت رم نصب شده و سیستم عامل ویندوز ۱۰.

کار آیی یک مدل در تشخیص نفوذ به مقادیر به دست آمده از معیارهای ارزیابی آن بستگی دارد. به منظور مشاهده بهتر، نتایج کلی این مقاله در ۲ مرحله‌ی جداگانه نمایش داده شده است. مرحله‌ی اول نشان‌دهنده‌ی ویژگی‌های عملکرد روش پیشنهادی برای طبقه‌بندی دودویی است در حالی که مرحله‌ی دوم، عملکرد معیارهای مختلف را برای تشخیص چند کلاسه نشان می‌دهد. در هر دو مرحله برای نشان دادن اثر انتخاب بهترین ویژگی‌ها در بهبود نتایج، آزمایش‌ها یکبار با کل مجموعه داده و بار دیگر با بخشی از مجموعه داده که شامل ویژگی‌های انتخاب شده است انجام گرفته است.

۵-۱- مرحله اول: دسته‌بندی دودویی با کل یا بخشی از ویژگی‌ها

میزان کار آیی هر مدل تشخیص نفوذ را می‌توان بر اساس نتایج معیارهای مختلف ارزیابی آن بررسی کرد. جایی که مقادیر بالاتر دقت، امتیاز F1، خاص بودن، دقت منفی، MCC و برعکس مقادیر پایین‌تر FAR، FNR و نرخ خطا نشان‌دهنده‌ی عملکرد تشخیص بهتر است. نتایج کلی طبقه‌بندی دودویی با مجموعه داده‌های UNSWNB15 و IoTID20 در جدول ۶ نشان داده شده است. برای تجزیه و تحلیل با رویکرد مقایسه‌ای، نتایج معیارهای به دست آمده با ابعاد اصلی و ابعاد کاهش یافته‌ی مجموعه داده‌ها انجام شده است.

هدف اصلی استفاده از PCA در هر مدل، کاهش پیچیدگی آن در سطح بیشتر و در عین حال حفظ الگوهای داده‌های اصلی است. از آنجایی که مزایای PCA در یافتن متغیرهای نامرتبط در داده‌ها است این نتیجه حاصل شده است که به کارگیری آن در مجموعه داده مؤثرتر عمل می‌کند. از جدول ۶، می‌توان مشاهده

نشان داده شده است.

جدول (۸). نتایج آزمایش چند کلاسه روی مجموعه داده

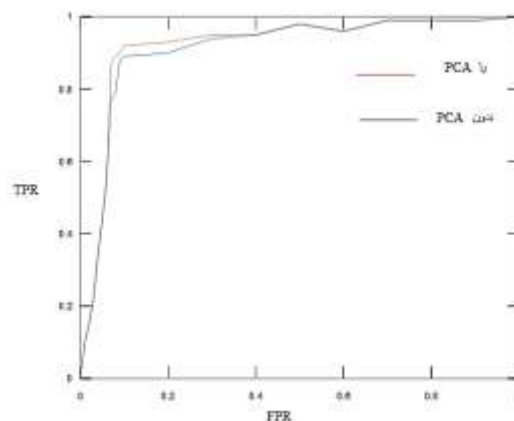
IoTID20		UNSW-NB15		معیار (%)
بدون PCA	با PCA	بدون PCA	با PCA	
۹۰٫۷۵	۹۳٫۰۹	۹۲٫۰۱	۹۸٫۴۶	Weighted Precision
۸۲٫۳۲	۸۸٫۱۶	۹۳٫۱۸	۹۷٫۴۱	Weighted Recall
۹۱٫۹۵	۹۷٫۴۸	۹۰٫۵۲	۹۱٫۴۳	Weighted F1 Score
۱۷٫۸۸	۱۲٫۰۴	۱۶٫۲۲	۱۰٫۲۸	Overall Error Rate

جدول (۹). زمان محاسبات آموزش و آزمایش چند کلاسه

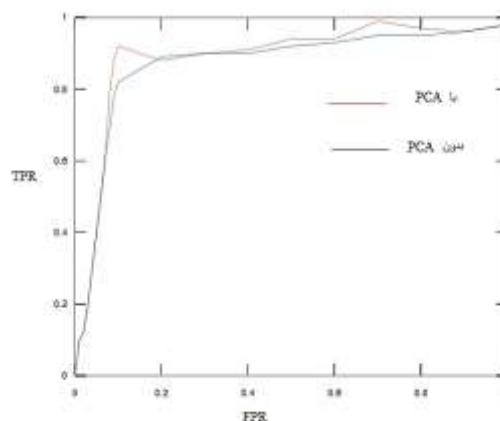
مجموعه داده		زمان آموزش (ثانیه)	زمان آزمایش (ثانیه)
UNSWNB15	با PCA	۱۲۹۳	۴۸۳
	بدون PCA	۳۰۱۴	۱۲۹۱
IoTID20	با PCA	۱۹۳۲	۱۳۱۵
	بدون PCA	۴۲۱۹	۲۹۴۶

یکی از نگرانی‌های اصلی که با مجموعه داده‌های چند کلاسه مرتبط است، توزیع نامتعادل ویژگی‌های کلاسی مختلف است. به عنوان مثال UNSW-NB15، ۵۸۸۷۱ حمله Generic و تنها ۱۷۴ برچسب حمله Worm را در خود جای داده است. به طور مشابه، برای IoTID20، در مقابل ۱۸۳۱۸۹ حمله Mirai-UDP Flooding، تنها ۲۲۱۹۲ حمله Hostport Scan در مجموعه داده وجود دارد. مدل در نتیجه چنین تعادل نابرابری، عملکرد متفاوتی را برای کلاس‌های مختلف نشان می‌دهد. به خصوص، برای کلاس‌هایی که داده‌های آموزشی کمتری دارند، مدل منجر به عملکرد ضعیف‌تری می‌شود. برای رفع این مشکل، از معیار ارزیابی F1-Score وزن‌دار نیز در نظر گرفته شده است. در اینجا عملکرد کلاس فردی به همراه تعداد رکوردهای کلاس مربوطه برای به دست آوردن عملکرد کلی مدل در نظر گرفته شده است. با بررسی جدول ۸، مشخص می‌شود که مجموعه داده

است. منحنی ROC بهینه دارای نرخ مثبت واقعی بالاتر همراه با نرخ مثبت کاذب کمتر است. در واقع، ناحیه زیر این منحنی، میزان مفید بودن مدل را نشان می‌دهد. در شکل ۳، منحنی ROC برای UNSW-NB15 و IoTID20 نشان داده شده است. با بهره‌گیری از نتایج به دست آمده و بررسی شکل ۳، می‌توان نتیجه‌گیری کرد که برای هر دو مجموعه داده، استفاده از PCA در بهبود عملکرد مدل مفید واقع می‌شود یعنی سطح زیر نمودار بیشتری در نمودار خواهیم داشت.



(الف) نمودار ROC روی مجموعه داده UNSWNB15



(ب) نمودار ROC روی مجموعه داده IoTID20

شکل (۳). نمودارهای ROC روی مجموعه داده‌های دودویی

۲-۵. مرحله دوم: دسته‌بندی چند کلاسه با کل یا

بخشی از ویژگی‌ها

همان‌طور که قبلاً هم ذکر شد بر خلاف مدل دسته‌بندی دودویی، مدل تشخیص چند کلاسه تعداد بیشتری از برچسب‌های کلاس را در خود جای می‌دهد؛ بنابراین، معیارهای عملکرد برای چنین مدل طراحی شده حاوی پارامترهای تکاملی اضافی است. در اینجا، مقادیر بالاتر مربوط به دقت وزن‌دار، امتیاز F1 وزن‌دار به همراه مقادیر پایین‌تر نرخ خطا، نشان‌دهنده عملکرد تشخیص بهتر هستند. خلاصه‌ای از عملکرد کلی برای مجموعه داده‌های چند کلاسه UNSWNB15 و IoTID20 در جدول ۸

برتری روش پیشنهادی در اغلب پارامترهای ارزیابی کارایی سیستم تشخیص نفوذ پیشنهادی دارد.

جدول (۱۰). مقایسه نتایج به دست آمده از روش پیشنهادی با برخی دیگر از روش‌های تشخیص نفوذ

روش	نوع دسته‌بندی	حداکثر دقت (%)
کومار و همکاران [۲۰]	دودویی	۹۹,۳
سرور و همکاران [۱۳]	دودویی و چند کلاسه	۹۸
گابریل و همکاران [۱۲]	دودویی و چند کلاسه	۹۸,۳
محمدرضایی و همکاران [۴]	دودویی و چند کلاسه	۹۹,۶
روش پیشنهادی مقاله	دودویی و چند کلاسه	۹۹,۷۳

۶- نتیجه‌گیری

به منظور مقابله با خطرات فزاینده علیه حریم خصوصی داده‌ها از طریق نقض‌های موجود در معماری شبکه، سیستم‌های تشخیص نفوذ در طول سال‌ها به اولویت اصلی تبدیل شده‌اند. اگرچه مطالعات زیادی در این زمینه انجام شده است، اما اغلب این مطالعات از اطلاعات ترافیک شبکه به درستی استفاده نمی‌کنند. باتوجه به این مشکل، در این مقاله، ما یک مطالعه جامع از تشخیص نفوذ شبکه با استفاده از یک مدل ترکیبی ارائه کردیم. مجموعه داده IoTID20 در کنار مجموعه داده معروف UNSW-NB15 برای رسیدگی به ناهنجاری‌های ترافیک شبکه مورد بررسی قرار گرفته است. ماهیت انفجاری داده‌های با ابعاد بالا اغلب می‌تواند فرآیند محاسباتی را در حین تجزیه و تحلیل الگوهای واقعی به خطر بیندازد. به عنوان یک اقدام مناسب در برابر چنین مشکلی، از روش تحلیل مؤلفه اصلی (PCA) استفاده شد که منجر به کاهش ابعاد و درعین حال حفظ ویژگی‌های مفید داده‌ها شده است. در اینجا آزمایش‌ها از دو منظر مختلف، دسته‌بندی دودویی و چند کلاسه انجام شد. نتایج مشاهدات نشان داد که با اتخاذ سیاست‌های درست در مرحله پیش‌پردازش داده‌ها، پیچیدگی مدل به میزان قابل توجهی کاهش می‌یابد و همراه با بهره‌گیری از سرعت اجرای الگوریتم جنگل تصادفی که یک الگوریتم مبتنی بر یادگیری ماشین است نتایج بادقت بالاتر از روش‌های قبلی به دست می‌آید

۷- مراجع

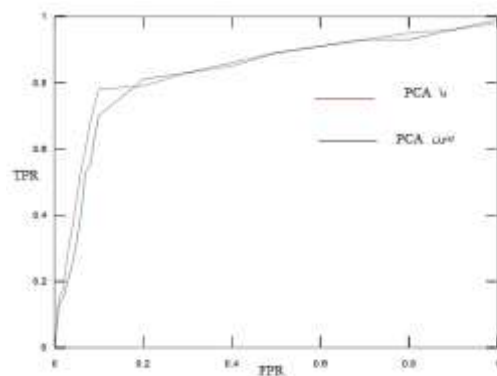
- [1] A. Dolan, L. Ray, and S. Majumdar, "Proactively extracting iot device capabilities: An application to smart homes", Data and applications security and privacy Conference, 2020, DOI:10.1007/978-3-030-49669-2_3.

UNSW-NB15 نسبت به مجموعه داده IoTID20 عملکرد بالاتری با ابعاد کاهش یافته دارد.

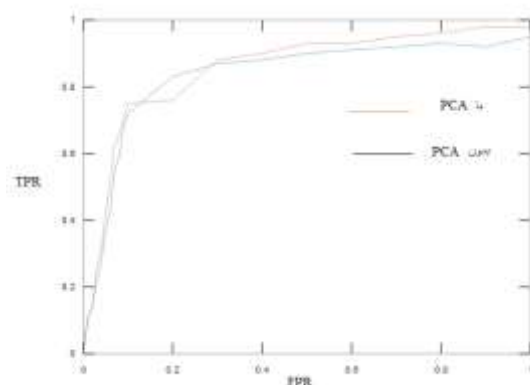
زمان محاسباتی صرف شده برای دسته‌بندی‌های چند کلاسه در جدول ۸ نشان داده شده است. در اینجا نیز مانند قبل، برای مجموعه داده‌های UNSW-NB15 و IoTID20 در زمان استفاده از PCA زمان محاسبات بسیار کمتر از زمانی است که کاهش ابعاد نداریم.

منحنی‌های ROC مربوط به عملیات چند کلاسه به منظور ارائه عملکرد خلاصه‌سازی در شکل ۴ نمایش داده شده‌اند که مانند حالت دودویی، در زمان استفاده از کاهش ابعاد، سطح زیر نمودار بیشتر و در واقع کارایی بهتر را داریم.

در حالت کلی، نتایج نشان می‌دهند که با بهره‌گیری از روش کاهش ابعاد PCA برای انتخاب ویژگی‌های مؤثرتر و استفاده از جنگل تصادفی برای دسته‌بندی داده‌ها، می‌توان زمان محاسبات را کاهش و دقت تشخیص را به طرز چشمگیری افزایش داد.



الف) نمودار ROC روی مجموعه داده UNSWNB15



ب) نمودار ROC روی مجموعه داده IoTID20

شکل (۴). نمودارهای ROC روی مجموعه داده‌های چند کلاسه

۵-۳. مقایسه با کارهای پیشین

در جدول ۱۰، مقایسه‌ای بین دقت روش پیشنهادی و برخی دیگر از کارهای انجام شده‌ی پیشین صورت گرفته است. نتایج نشان از

- detection system for IoT environment using machine learning techniques", *Procedia Comput. Sci.*, vol. 171, pp.2372-2379,2020, DOI:10.1016/j.procs.2020.04.257.
- [12] T. Gaber, A. El-Ghamry, and A. E. Hassanien, "Injection attack detection using machine learning for smart IoT applications", *Phys. Commun.*, vol. 52, pp. 101685-101695, 2022, DOI:10.1016/j.phycom.2022.101685.
- [13] A. Sarwar, S. Hasan, W. U. Khan, S. Ahmed, and S. N. K. Marwat, "Design of an Advance Intrusion Detection System for IoT Networks", 2nd International Conference on Artificial Intelligence (ICAI), 2022, DOI: 10.1109/ICAI55435.2022.9773747.
- [14] M. Bagaa, T. Taleb, J. B. Bernabe, and A. Skarmeta, "A machine learning security framework for iot systems", *IEEE Access*, vol. 8, pp. 114066- 114077, 2020, DOI: 10.1109/ACCESS.2020.2996214.
- [15] N. Moustafa, and J. Slay, "Unsw-nb15: A comprehensive data set for network intrusion etection systems (unsw-nb15 network data set)", military communications and information systems conference (MilCIS), 2015, DOI: 10.1109/MilCIS.2015.7348942.
- [16] I. Ullah, and Q. Mahmoud, "A scheme for generating a dataset for anomalous activity detection in iot networks", Canadian conference on AI, 2020, DOI:10.1007/978-3-030-47358-7_52.
- [17] P. Maniriho, E. Niyigaba, Z. Bizimana, V. Twiringiyimana, L.J. Mahoro, L. J, and T. Ahmad, "Anomaly-based intrusion detection approach for iot networks using machine learning", international conference on computer engineering, network, and intelligent multimedia (CENIM), 2020, DOI: 10.1109/CENIM51130.2020.9297958
- [18] R. Qaddoura, A.M. Al-Zoubi, I. Almomani, and H. Faris, "A multi-stage classification approach for iot intrusion detection based on clustering with oversampling", *Appl. Sci.*, vol. 11, pp. 3022, 2021,mdpi .com /2076 -3417 /11 /7 /3022.
- [19] A. Farah, "Cross dataset evaluation for IoT network intrusion detection" (Ph.D. thesis), 2020.
- [20] P. Kumar, G.P. Gupta, R. Tripathi, S. Garg, and M.M. Hassan, "DLTIF: Deep learning-driven cyber threat intelligence modeling and identification framework in IoT-enabled maritime transportation systems", *IEEE Trans. Intell. Transp. Syst.*, vol. 24, pp. 1–10, 2021, DOI:10.1109/tits.2021..3122368
- [2] H. Tanha, and M. Abbasi, "Identify malicious traffic on IoT infrastructure using neural networks and deep learning", *Electron. Cyber Def.*, vol. 11, no.2. pp. 1-13, 2023,(in pershian),dor: https://dor.isc.ac/dor/20.1001.1.23224347.1402.11.2.1.4
- [3] M. Choras, and M. Pawlicki, "Intrusion detection approach based on optimised artificial neural network", *Neurocomputing*, vol. 452, pp. 705–715, 2021, DOI:10.1016/j.neucom.2020.07.138.
- [4] M. Mohammadrezaei, "Detecting Fake Accounts on Social networks using Principal Components Analysis and Algorithm Kernel Density Estimation (A case study on the Twitter social network)," *Electron. Cyber Def.*, vol. 9, no.3 pp. 109-123, 2021,(in pershian) dor: 20.1001.1.23224347.1400.9.3.9.0
- [5] K. Keerthi Vasan, and B. Surendiran, "Dimensionality reduction using Principal Component Analysis for network intrusion detection", *Perspect. Sci.*, vol. 8, pp. 510-512, 2016, DOI:10.1016/j.pisc.2016.05.010.
- [6] F. B. Islam, R. Akter, D.S. Kim, and J.M. Lee, "Deep learning based network intrusion detection for industrial internet of things", vol. 8, pp. 418–421, 2020,journal-home.s3.ap-northeast-2.amazonaws.com/site/2020kics/presentation/0669.pdf.
- [7] M.A. Jabbar, R. Aluvalu, and S.S.S Reddy, "Cluster based ensemble classification for intrusion detection system", *Proceedings of the 9th international conference on machine learning and computing*, 2017, DOI:10.1145/3055635.3056595.
- [8] D. Gaikwad, and R.C. Thool, "Intrusion detection system using bagging ensemble method of machine learning", international conference on computing communication control and automation, 2015, DOI: 10.1109/ICCUBE.2015.61.
- [9] M.P. Kantipudi, R. Aluvalu, and S. Velamuri, "An intelligent approach of intrusion detection in mobile crowd sourcing systems in the context of iot based smart city", *Smart Science*, vol. 11, pp. 234–240, 2022, DOI:10.1080/23080477.2022.2117889.
- [10] M. Woźniak, A. Zielonka, A. Sikora, M.J. Piran, and A. Alamri, "6g-enabled iot home environment control using fuzzy rules", *IEEE INTERNET THINGS*, vol. 8, pp.5442–5452,2020, DOI: 10.1109/JIOT.2020.3044940.
- [11] K. S. Kiran, R. K. Devisetty, N. P. Kalyan, K. Mukundini, and R. Karthi, "Building a intrusion