Journal of New Technologies in Distributed Systems and Algorithmic Computing
Islamic Azad University of Sabzevar
E-ISSN: 3115-705X
https://sanad.iau.ir/journal/ntds

Reaserch Article

Improved Hybrid Algorithm for Detection of Intrusion into Computer Networks

Firuzeh Razavi[1*] | Safanaz Heydari[2]

[1]Department of Information Technology Management-Electronic Business, Faculty of Humanties, Raja University, Qazvin, Iran, F.razavi@raja.ac.ir

[2]Department of Management, Miandoab Branch, Islamic Azad University, Miandoab, Iran, Safanazheidari@gmail.com

Corresponding Author
[*]Firuzeh Razavi, Assistant Professor, Department of Management, Miandoab Branch, Islamic Azad University, Miandoab, Iran, F.razavi@raja.ac.ir

Main Subjects: Identifying Malicious Networks
Received: 20 June 2024
Revised: 12 May 2025
Accepted: 16 May 2025

https://doi.org/10.82195/ntds.2025.11232

Abstract

Identifying malicious networks has been a subject of study for decades, and since the volume of network traffic is increasing day by day, there is a need for a successful intrusion-detection system that can make the identification process easier during attacks. The aim behind this research was to take decisions more accurately via real time and faster processing. The purpose of this research was to detect intrusion into computer networks by combining K-means and XG-boost clustering algorithms. The proposed method was performed in two stages. In the first stage, the pre-processing was done by normalizing and digitizing the data set, as well as removing outliers based on two PCA methods and reducing the dimensions of the feature, then using the learner. The researchers used the k-means algorithm to find the optimal number of clusters, finally the Elbow method was utilized to find the optimum number of clusters. The second stage consisted of classifying malicious and normal network traffic from each other by combining K-means and XG-Boost algorithms on computing platforms. The experiments in this article were done using the NSLKDD data set and its implementation in the knime emulator platform; the final evaluation results revealed the superiority of the error detection rate, and the accuracy and correctness of the proposed algorithm compared over other similar method.

Keywords: Intrusion Detection System, Machine Learning, K-means Clustering، XGBoost Classifier, Cybersecurity.

فناوری های نوین در سیستم های توزیع شده و محاسبات الگوریتمی

دانشگاه آزاد اسلامی واحد سبزوار

E-ISSN: 3115-705X
https://sanad.iau.ir/journal/ntds

پژوهشی

الگوریتم بهبودیافته ترکیبی برای تشخیص نفوذ به شبکه‌های کامپیوتری

فیروزه رضوی*[۱] | صفاناز حیدری[۲]

[۱]گروه مدیریت فناوری اطلاعات-کسب و کار الکترونیک، دانشکده علوم انسانی، دانشگاه رجاء، قزوین، ایران
F.razavi@raja.ac.ir

[۲]استادیار گروه مدیریت، واحد میاندوآب، دانشگاه آزاد اسلامی، میاندوآب، ایران
Safanazheidari@gmail.com

نویسنده مسئول
***فیروزه رضوی، استادیار گروه مدیریت فناوری اطلاعات، دانشگاه رجاء، قزوین، ایران**
F.razavi@raja.ac.ir

موضوع اصلی: تشخیص نفوذ به شبکه‌های کامپیوتری
تاریخ دریافت: ۳۱ خرداد ۱۴۰۳
تاریخ بازنگری: ۲۲ اردیبهشت ۱۴۰۴
تاریخ پذیرش: ۲۶ اردیبهشت ۱۴۰۴

https://doi.org/10.82195/ntds.2025.1123245

چکیده:

افزایش چشمگیر حجم ترافیک شبکه و پیچیدگی فزاینده حملات سایبری، ضرورت توسعه سیستم‌های تشخیص نفوذ بادقت و سرعت بالا را بیش‌ازپیش برجسته کرده است. این پژوهش باهدف ارائه یک مدل ترکیبی مبتنی بر یادگیری ماشین برای بهبود دقت شناسایی نفوذ، کاهش نرخ هشدارهای کاذب و ارتقای کارایی سیستم‌های تشخیص نفوذ انجام شده است. رویکرد پیشنهادی از ترکیب الگوریتم خوشه‌بندی K-means و طبقه‌بند قدرتمند XGBoost بهره می‌گیرد، به‌طوری‌که K-means داده‌ها را ساختاردهی کرده و پیچیدگی آن‌ها را کاهش می‌دهد و XGBoost وظیفه طبقه‌بندی نهایی و تشخیص دقیق الگوهای حملات را بر عهده دارد. مجموعه‌داده NSL-KDD پس از پیش‌پردازش‌هایی شامل نرمال‌سازی، حذف داده‌های پرت، کدگذاری ویژگی‌های اسمی و کاهش ابعاد با استفاده از تحلیل مؤلفه‌های اصلی آماده‌سازی شد. نوآوری اصلی این مطالعه در استفاده هدفمند از K-means برای آماده‌سازی داده‌ها پیش از طبقه‌بندی و تنظیم دقیق فراپارامترهای مدل به‌منظور بهینه‌سازی عملکرد نهفته است. پیاده‌سازی مدل در بستر KNIME نشان داد که این روش به‌دقت و نرخ تشخیص ۹۹.۸۶درصد دست یافته و در مقایسه با روش‌های مرجع مانند AE-LSTM، RNN-IDS و K-means-RF عملکرد برتری ارائه می‌دهد. با وجود وابستگی مدل به داده‌های ایستا به‌عنوان یک محدودیت، طراحی ماژولار آن امکان تعمیم به محیط‌های واقعی و مجموعه‌داده‌های متنوع‌تر را فراهم می‌سازد.

کلیدواژه‌ها: سیستم تشخیص نفوذ، یادگیری ماشین، خوشه‌بندی K-means، طبقه‌بندی XGBoost ، حملات سایبری

۱- مقدمه

تکنولوژی در چند دهه اخیر به طور تصاعدی تکامل‌یافته است. فناوری علاوه بر داشتن مزیت، تهدیدهای امنیتی نیز به همراه دارد. محافظت از شبکه‌های مدرن و اینترنت به‌منظور جلوگیری از نفوذ هکرها و حملات سایبری صورت می‌پذیرد. برای محافظت از شبکه‌های مدرن روش‌های مختلف امنیت سایبری و سیستم‌های حفاظتی مانند دیواره‌های آتش [۱] تکنیک‌های احراز هویت، روش‌های رمزنگاری و سیستم‌های تشخیص نفوذ [۲] به‌منظور پایش ترافیک شبکه‌ها معرفی شدند[1]. موضوع تشخیص نفوذ امری مهم در شبکه به‌حساب می‌آید اگرچه پیشرفت‌های قابل‌توجهی در این زمینه حاصل شده است؛ اما هنوز فرصت‌های زیادی برای بهبود روش‌های شناسایی و جلوگیری از حملات مبتنی بر شبکه وجود دارد.

با پیچیده‌تر شدن حملات سایبری و افزایش حجم داده‌های شبکه، طراحی سیستم‌های تشخیص نفوذ مؤثر و سریع به یک چالش اساسی در امنیت اطلاعات تبدیل شده است. استفاده از داده‌کاوی و یادگیری ماشین به‌ویژه در شرایطی که شناخت اولیه‌ای از الگوهای حملات وجود ندارد، روشی کارآمد برای استخراج دانش از داده‌های عظیم محسوب می‌شود. الگوریتم‌های هوش مصنوعی مانند روش‌های فازی، ژنتیک، و مدل‌های یادگیری نظارت شده و غیرنظارت‌شده، به‌صورت گسترده در طراحی سیستم‌های تشخیص نفوذ به کار گرفته شده‌اند. این روش‌ها علاوه بر کاهش وابستگی به متخصصان انسانی، موجب افزایش سرعت، دقت و قابلیت انطباق سیستم‌ها شده‌اند. براین‌اساس، تحقیق حاضر به دنبال بهبود دقت تشخیص، کاهش نرخ هشدار کاذب و افزایش کارایی با استفاده از الگوریتم‌های ترکیبی داده‌کاوی است. همچنین، این پژوهش به ارزیابی تأثیر حذف داده‌های پرت، انتخاب ویژگی، و بهینه‌سازی معماری مدل در عملکرد نهایی سیستم تشخیص نفوذ پرداخته است.

شناسایی نفوذهای شبکه‌های مخرب برای دهه‌ها موضوع مطالعه بوده است. با این‌حال، همان‌طور که دانشمندان داده می‌توانند درک کنند، هنگامی که مقیاس یک مشکل به‌ترتیب بزرگی افزایش پیدا می‌کند، رویکردهای موجود اغلب دیگر موثر نیستند. مشکل به قدری متفاوت است که نیاز به راه حل جدیدی دارد و از آن‌جایی که حجم ترافیک شبکه روز به روز درحال افزایش است، حوزه تشخیص نفوذ مجبور به اختراع مجدد خود پیرامون تکنیک‌های کلان داده شده است. یک سیستم تشخیص نفوذ شبکه‌ها یا سایر سیستم‌ها را برای رفتارهای مخرب یا غیرعادی نظارت می‌کند. با تکمیل فن‌آوری‌های پیش‌گیرانه مانند دیوارهای آتش، احرازهویت قوی [۳] و حق‌امتیاز [۴] [8] سیستم‌های تشخیص نفوذ به بخشی ضروری از مدیریت امنیت فناوری اطلاعات سازمانی تبدیل شده‌اند[9]. این سیستم‌ها به‌طور معمول به دو دسته تحت عنوان سیستم‌های مبتنی بر سوءاستفاده یا مبتنی بر ناهنجاری طبقه‌بندی می‌شوند[10]. تکنیک‌های داده‌کاوی به طور فزاینده‌ای برای شناسایی حملات، ناهنجاری‌ها یا نفوذها در یک محیط شبکه محافظت شده استفاده می‌شوند[11].

یک سیستم تشخیص نفوذ موفق ممکن است داده‌های با ابعاد بالا را برای تصمیم‌گیری در زمان واقعی و سریع پردازش کند و نرخ هشدار نادرست را پایین و نرخ تشخیص را بالا نگه دارد ایجاد یک سیستم تشخیص نفوذ قابل‌اعتماد و مؤثر به دلیل نامتعادل بودن مجموعه‌داده‌ها، داده‌های ابعادی بالا و ماهیت در حال تحول حملات سایبری، به یک کار چالش‌برانگیز تبدیل می‌شود. یک سیستم تشخیص نفوذ شبکه‌ها یا سایر سیستم‌ها را برای رفتارهای مخرب یا غیرعادی نظارت می‌کند. با تکمیل فناوری‌های پیش‌گیرانه مانند دیوارهای آتش، احراز هویت قوی و حق امتیاز، سیستم‌های تشخیص نفوذ به بخشی ضروری از مدیریت امنیت فناوری اطلاعات سازمانی تبدیل شده‌اند.

1. Firewall
2. Instrusion Detection Systems
3 . Strong authentication
4 . User privilege

گرچه داده‌کاوی و یادگیری ماشین به‌صورت گسترده در مطالعات پیشین به کار رفته‌اند، اما بیشتر پژوهش‌ها بر مدل‌های مبتنی بر سوءاستفاده یا تشخیص ناهنجاری متمرکز بوده‌اند و پژوهش‌هایی که به موضوع تشخیص نفوذ در زمان واقعی و آنلاین بپردازند، نسبتاً محدود هستند.

نوآوری اصلی این پژوهش در طراحی یک سیستم تشخیص نفوذ ترکیبی مبتنی بر الگوریتم‌های K-means و XGBoost نهفته است که با بهره‌مندی از تکنیک‌های پیش‌پردازش پیشرفته و بهینه‌سازی دقیق، عملکردی برتر نسبت به روش‌های موجود ارائه می‌دهد. در این رویکرد، ابتدا از الگوریتم K-means برای خوشه‌بندی داده‌های شبکه‌ای به گروه‌های همگن استفاده می‌شود که این فرایند با سازمان‌دهی داده‌ها و کاهش پیچیدگی ساختاری آن‌ها، بستری مناسب برای طبقه‌بندی دقیق‌تر فراهم می‌آورد. به‌منظور ارتقای کارایی، تحلیل مؤلفه‌های اصلی (PCA) به‌عنوان یک تکنیک پیش‌پردازش به کار گرفته شده است تا با کاهش ابعاد داده‌ها و حذف نویز، کیفیت ورودی‌ها برای مرحله طبقه‌بندی بهبود یابد. در گام بعدی، الگوریتم تقویت گرادیان شدید (XGBoost) برای طبقه‌بندی نهایی مورداستفاده قرار می‌گیرد. این الگوریتم، با توانایی برجسته در مدل‌سازی روابط غیرخطی، مدیریت داده‌های نامتوازن، و کنترل بیش برازش از طریق منظم‌سازی پیشرفته، دقت و نرخ تشخیص حملات را به طور قابل‌توجهی ارتقا می‌دهد. تنظیم دقیق فراپارامترهای XGBoost نظیر نرخ یادگیری، عمق درختان، و تعداد تکرارها با استفاده از روش‌های بهینه‌سازی، یکی از جنبه‌های کلیدی این پژوهش است که به بهبود عملکرد مدل کمک شایانی کرده است. داده‌ها برای طبقه‌بند و تنظیم دقیق فراپارامترهای مدل ترکیبی به‌منظور ارتقای عملکرد نهایی سیستم تشخیص نفوذ است.

در ادامه در بخش دوم پیشینه تحقیق بیان شده و در بخش سوم ارائه راهکار پیشنهادی به طور کامل توضیح داده شده و در بخش چهارم نحوه پیاده‌سازی و ارزیابی مدل پیشنهادی ارائه شده است. در نهایت در بخش پنجم، جمع‌بندی تحقیق بیان شده است.

۲- پیشینه تحقیق

افزایش حجم ترافیک شبکه و پیچیدگی حملات سایبری، توسعه سیستم‌های تشخیص نفوذ IDS بادقت بالا، نرخ هشدار کاذب پایین و قابلیت پردازش بلادرنگ را به یک ضرورت تبدیل کرده است. در سال‌های اخیر، پژوهش‌های متعددی با استفاده از تکنیک‌های داده‌کاوی و یادگیری ماشین برای بهبود عملکرد سیستم‌های تشخیص نفوذ انجام شده است. این مطالعات بر روش‌های نظارت شده، غیرنظارت‌شده و ترکیبی متمرکز بوده‌اند، اما همچنان چالش‌هایی مانند مدیریت داده‌های حجیم، کاهش نرخ هشدار کاذب و تعمیم‌پذیری به محیط‌های واقعی باقی‌مانده است. در این بخش، مطالعات مرتبط با روش‌های تشخیص نفوذ مبتنی بر یادگیری ماشین، به‌ویژه روش‌های ترکیبی که از خوشه‌بندی و طبقه‌بندی استفاده می‌کنند، بررسی می‌شوند تا زمینه‌ای برای روش پیشنهادی این پژوهش فراهم شود.

لی و همکاران [12] فرآیندکاوی را برای طراحی سیستم‌های تشخیص نفوذ مبتنی بر میزبان بررسی کردند. آن‌ها سیستمی پیشنهاد کردند که مراحل پیش‌پردازش، تشخیص ناهنجاری و شناسایی سوءاستفاده را به‌صورت موازی انجام می‌داد و نتایج را ترکیب می‌کرد. این روش کارایی سیستم‌های داده‌محور را بهبود بخشید، اما در افزایش دقت و کاهش نرخ هشدار کاذب با محدودیت‌هایی مواجه بود. این مطالعه بر اهمیت پیش‌پردازش داده‌ها و ترکیب روش‌های مختلف برای بهبود عملکرد IDS تأکید دارد، که الهام‌بخش روش پیشنهادی این پژوهش در استفاده از پیش‌پردازش پیشرفته است.

ژانگ و همکاران [13] یک سیستم تشخیص نفوذ مبتنی بر انتخاب ویژگی و مدیریت هشدار توسعه دادند که با استفاده از نرم‌افزار Wekaو الگوریتم‌های مختلف طبقه‌بندی، سرعت و دقت بالاتری نسبت به روش‌های مبتنی بر خوشه‌بندی نشان داد. این پژوهش با پیشنهاد پنج نمونه داده بهینه و استفاده از چندین الگوریتم طبقه‌بندی، به بهبود عملکرد IDS کمک کرد. با این حال، تمرکز

این روش بر مدیریت هشدار بود و کمتر به ترکیب الگوریتم‌های پیشرفته مانند روش‌های تقویتی پرداخت. این خلأ یکی از انگیزه‌های اصلی پژوهش حاضر برای ترکیب خوشه‌بندی و طبقه‌بندی تقویتی است.

یین و همکاران [14] یک رویکرد یادگیری عمیق مبتنی بر شبکه‌های عصبی بازگشتی و رمزگذار خودکار طولانی-کوتاه‌مدت (LSTM)پیشنهاد کردند که به مدل AE-LSTM معروف است. این مدل با استفاده از مجموعه داده NSL-KDD به دقت ۸۹درصد، نرخ تشخیص ۸۸ درصد و نرخ هشدار کاذب ۱۱درصد دست‌یافت. اگرچه این روش در استخراج ویژگی‌های پیچیده موفق بود، اما پیچیدگی محاسباتی بالا و نرخ هشدار کاذب نسبتاً زیاد، کاربرد آن را در سناریوهای بلادرنگ محدود کرد. این مطالعه نشان‌دهنده پتانسیل یادگیری عمیق در IDS است، اما نیاز به روش‌های سبک‌تر و دقیق‌تر را برجسته می‌کند که در روش پیشنهادی این پژوهش موردتوجه قرار گرفته است.

ژانگ و همکاران [15] یک سیستم تشخیص نفوذ ترکیبی مبتنی بر خوشه‌بندی K-means و طبقه‌بندی جنگل تصادفی (Random Forest)پیشنهاد کردند که به مدل K-Means-RF معروف است. این روش با استفاده از K-means برای ساختاردهی داده‌ها و جنگل تصادفی برای طبقه‌بندی، به دقت ۹۲.۸۹درصد، نرخ تشخیص ۹۸.۵۷درصد و نرخ هشدار کاذب ۱۴.۶درصد دست یافت. این مطالعه نشان داد که خوشه‌بندی می‌تواند پیچیدگی داده‌ها را کاهش دهد و عملکرد طبقه‌بندی را بهبود بخشد. با این حال، نرخ هشدار کاذب بالا نشان‌دهنده نیاز به طبقه‌بندهای قوی‌تر است. روش پیشنهادی این پژوهش از K-means الهام گرفته، اما به‌جای جنگل تصادفی از XGBoost استفاده می‌کند تا دقت و نرخ تشخیص را بهبود بخشد.

یین و همکاران [16] مدل RNN-IDS را بر پایه شبکه‌های عصبی بازگشتی توسعه دادند که با تمرکز بر استخراج ویژگی‌های زمانی از ترافیک شبکه، به‌دقت ۸۲.۴۹ درصد، نرخ تشخیص ۸۰درصد و نرخ هشدار کاذب ۱۲درصد دست‌یافت. این روش در مقایسه با سایر مدل‌ها عملکرد ضعیف‌تری داشت، به‌ویژه در مدیریت داده‌های پیچیده و نامتوازن. این مطالعه بر اهمیت استفاده از روش‌های ترکیبی برای غلبه بر محدودیت‌های مدل‌های مبتنی بر یادگیری عمیق تأکید دارد که در طراحی روش پیشنهادی این پژوهش موردتوجه قرار گرفته است.

لین و همکاران [17] یک سیستم تشخیص نفوذ مبتنی بر یادگیری خودآموز با استفاده از رمزگذار خودکار پراکنده پیشنهاد کردند که به مدل DST-TL معروف است. این روش با دقت ۸۴.۶۰درصد، نرخ تشخیص ۸۶درصد و نرخ هشدار کاذب ۱۴درصد عملکرد متوسطی ارائه داد. این مطالعه نشان داد که روش‌های خودآموز می‌توانند مکمل روش‌های نظارت‌شده باشند، اما برای دستیابی به دقت بالا نیاز به ترکیب با الگوریتم‌های قوی‌تر دارند. روش پیشنهادی این پژوهش با ترکیب K-means و XGBoost این محدودیت را برطرف می‌کند.

لی و همکاران [18] یک مدل ترکیبی مبتنی بر K-means و XGBoost برای تشخیص نفوذ پیشنهاد کردند. این روش با استفاده از خوشه‌بندی K-means برای کاهش پیچیدگی داده‌ها و طبقه‌بندی XGBoost برای شناسایی حملات، به دقت ۹۹.۸۵درصد، نرخ تشخیص ۹۹.۸۴درصد و نرخ هشدار کاذب ۱۴.۵۶درصد دست یافت. این مطالعه یکی از نزدیک‌ترین رویکردها به روش پیشنهادی این پژوهش است، اما فاقد بهینه‌سازی پیشرفته فراپارامترها و پیش‌پردازش داده‌ها بود که در این پژوهش مورد توجه قرار گرفته است. روش پیشنهادی ما با تنظیم دقیق فراپارامترها و استفاده از تحلیل مؤلفه‌های اصلی (PCA) عملکرد بهتری ارائه می‌دهد.

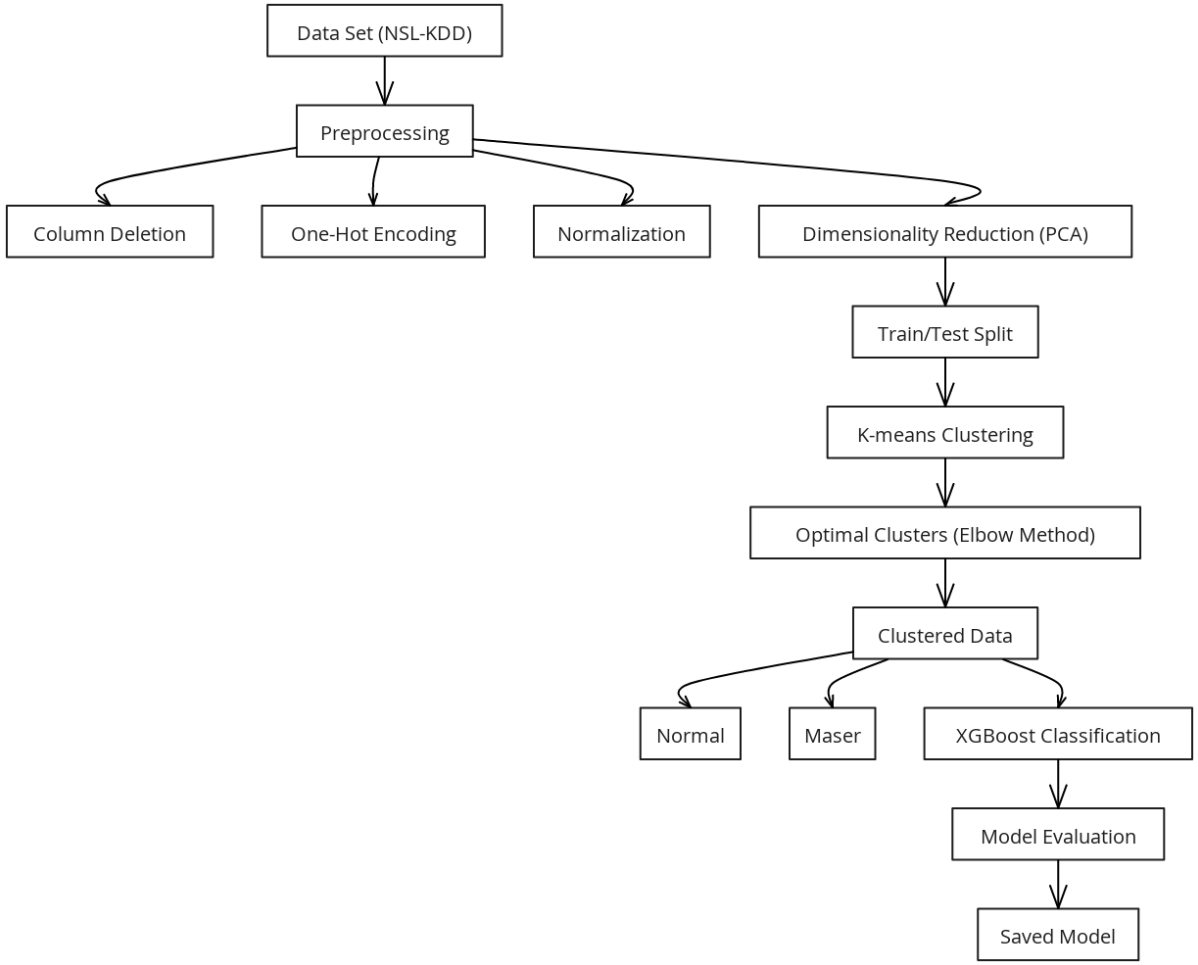
چن و همکاران [19] الگوریتم XGBoost را معرفی کردند که به دلیل توانایی در مدیریت داده‌های نامتوازن، مدل‌سازی روابط غیرخطی و کنترل بیش برازش (overfitting) از طریق منظم‌سازی، به طور گسترده در مسائل طبقه‌بندی، از جمله تشخیص نفوذ، استفاده شده است. مطالعات متعددی کاربرد XGBoost را در IDS بررسی کرده‌اند، اما استفاده مستقل آن معمولاً دقت

کمتری نسبت به روش‌های ترکیبی ارائه می‌دهد. این پژوهش از XGBoost به‌عنوان طبقه‌بند اصلی پس از خوشه‌بندی k-means استفاده می‌کند تا دقت و نرخ تشخیص را به طور قابل‌توجهی بهبود بخشد.

بررسی مطالعات پیشین نشان می‌دهد که روش‌های ترکیبی مبتنی بر خوشه‌بندی مانند K-means و طبقه‌بندی مانند جنگل تصادفی یا XGBoost به دلیل توانایی در ساختاردهی داده‌ها و شناسایی دقیق الگوهای حملات، عملکرد بهتری نسبت به روش‌های مستقل دارند. بااین‌حال اکثر پژوهش‌ها بر استفاده ساده از این الگوریتم‌ها متمرکز بوده‌اند و کمتر به بهینه‌سازی فراپارامترها، پیش‌پردازش پیشرفته مانند کاهش ابعاد با PCA و ترکیب هدفمند K-means و XGBoost پرداخته‌اند. این پژوهش باهدف رفع این خلأ، یک الگوریتم ترکیبی بهبودیافته مبتنی بر K-means و XGBoost ارائه می‌دهد که با استفاده از خوشه‌بندی برای آماده‌سازی داده‌ها، کاهش ابعاد، و طبقه‌بندی قوی با XGBoost، دقت، نرخ تشخیص و کارایی سیستم‌های تشخیص نفوذ را ارتقا می‌بخشد. مدل‌های بررسی‌شده در این بخش AE-LSTM، K-Means-RF، RNN-IDS، DST-TL، و K-means– XGBoost به‌عنوان معیارهای مقایسه با روش پیشنهادی در بخش نتایج استفاده خواهند شد.

۳- روش پیشنهادی

فرایند کلی اجرای مدل پیشنهادی در شکل ۱ نمایش داده شده است. همان‌طور که در این نمودار مشاهده می‌شود، ابتدا مجموعه داده NSL-KDD تحت مراحل مختلف پیش‌پردازش شامل حذف ویژگی‌های بی‌اثر، کدگذاری One-Hot و نرمال‌سازی قرار می‌گیرد. پس از آماده‌سازی داده‌ها، مجموعه داده به بخش‌های آموزش و آزمون تقسیم شده و در ادامه مرحله کاهش ابعاد با استفاده از تحلیل مؤلفه‌های اصلی اعمال می‌گردد. این مرحله به کاهش ویژگی‌های زائد و حفظ مؤلفه‌های اصلی مؤثر در داده‌ها کمک می‌کند. سپس داده‌ها به الگوریتم K-means وارد شده و بر اساس برچسب‌های خوشه‌بندی، نمونه‌ها به دو گروه Normalو Masers تقسیم می‌گردند. گروه اول به‌منظور مدل‌سازی ذخیره شده و گروه دوم به طبقه‌بند XGBoost وارد می‌شود تا فرآیند شناسایی حملات انجام پذیرد. این ساختار گام‌به‌گام، بهینه‌سازی عملکرد مدل و افزایش دقت سیستم تشخیص نفوذ را به دنبال داشته است.



شکل ۱: چارچوب سیستم تشخیص نفوذ مبتنی بر K-means – Xgboost

Figure 1: ramework of an intrusion detection system based on K-means and XGBoost

۳-۱ داده‌های مورد استفاده

در این پژوهش از مجموعه‌های KDDTrain+، KDDTest+ و KDDTest-21 مجموعه‌داده‌های NSL-KDD که در جدول ۱ آمده، استفاده شده است. مجموعه KDDTrain+ به‌عنوان مجموعه‌داده برای آموزش شامل ۱۲۵۹۷۳ نمونه است که شامل ۵۸۶۳۰ مورد ترافیک حمله و ۶۷۳۴۳ نمونه ترافیک عادی است. مجموعه KDDTest+ شامل ۲۲۵۴۴ نمونه است و برای تست از آن استفاده می‌شود به‌عنوان زیرمجموعه‌ای از مجموعه KDDTest+ ، مجموعه KDDTest-21 شامل کل ۱۱۸۵۰ نمونه است. اعتبارسنجی متقابل بر روی مجموعه KDDTrain+ در آزمایش‌ها انجام می‌شود.

جدول ۱: ویژگی‌های مجموعه داده NSL-KDD
Table 1: Features of the NSL-KDD Dataset

Input Attribute	Num	Input Attribute	Num	Input Attribute	Num
srv_diff_host_rate	۳۱	num_root	۱۶	Duration	۱
dst_host_count	۳۲	num_file_creations	۱۷	Protocol_Type	۲
dst_host_srv_count	۳۳	num_shells	۱۸	Service	۳
dst_host_same_srv_rate	۳۴	num_access_files	۱۹	Flag	۴
dst_host_diff_srv_rate	۳۵	num_outbound_cmds	۲۰	Src_Bytes	۵
dst_host_same_src_port_rate	۳۶	is_host_login	۲۱	Dst_Bytes	۶
dst_host_srv_diff_host_rate	۳۷	is_guest_login	۲۲	Land	۷
dst_host_serror_rate	۳۸	Count	۲۳	wrong_fragment	۸
dst_host_srv_serror_rate	۳۹	srv_count	۲۴	Urgent	۹
dst_host_rerror_rate	۴۰	serror_rate	۲۵	Hot	۱۰
dst_host_srv_rerror_rate	۴۱	srv_serror_rate	۲۶	num_failed_logins	۱۱
-	-	rerror_rate	۲۷	logged_in	۱۲
-	-	srv_rerror_rate	۲۸	num_compromise	۱۳
-	-	same_srv_rate	۲۹	root_shell	۱۴
-	-	diff_srv_rate	۳۰	su_attempted	۱۵

۳-۲ پیش پردازش داده

خوشه‌بندی به‌عنوان یکی از روش‌های یادگیری ماشین غیر نظارتی[1] در حل مسائل دسته‌بندی و طبقه‌بندی مشاهدات، بسیار به کار می‌رود. این کار به‌وسیله بررسی و محاسبه توابع فاصله[2] بر اساس ویژگی‌های مشاهدات، انجام شده و نقاط با کمترین میزان فاصله در یک گروه قرار می‌گیرند. مسئله مهمی که دراین‌رابطه به وجود می‌آید، نرمال‌سازی داده‌ها در خوشه‌بندی است؛ زیرا باید ویژگی‌ها در محاسبه فاصله بدون مقیاس باشند تا بزرگی واحد اندازه‌گیری هر بُعد باعث اریبی مقدار تابع فاصله به سمت یک ویژگی نشود[20]. شیوه‌های مختلفی برای نرمال‌سازی وجود دارند که در مرحله آماده‌سازی داده‌ها به کار می‌روند که در این تحقیق از روش نرمال‌سازی مقدار حداقل - حداکثر[3] که معروف‌ترین شیوه در نرمال‌سازی داده‌ها است[21] استفاده شده است. عملیات نرمال‌سازی قبل از بسیاری از الگوریتم‌های داده‌کاوی مانند شبکه‌های عصبی، ماشین بردار پشتیبان، KNN و K-means باید انجام بگیرد تا ابعادِ مختلف به‌صورت عادلانه توسط الگوریتم بررسی شوند و تأثیر یکی بیشتر از بقیه نباشد.در مراحل پیش‌پردازش الگوریتم پیشنهادی فرایندهای زیر اجرا گردید.

۱. حذف ستون‌های اضافی: ستون "num_outbound_cmds" فقط حاوی یک مقدار منفرد یعنی صفر بود. بنابراین این ستون حذف شد زیرا هیچ مشارکتی نداشت.

1. Unsupervised Machine Learning
2. Distance Function
3. Minimum- Maximum

۲. Factorization و One Hot Encoding: ویژگی‌هایی مانند «نوع_پروتکل»، «سرویس» و «پرچم» ویژگی‌های اسمی و نوع متن هستند؛ بنابراین آن‌ها را فاکتورسازی کردیم تا آن‌ها را به ویژگی‌های عددی اسمی تبدیل کنیم و سپس One Hot Encoding را برای تبدیل بیشتر آنها به ویژگی‌های باینری انجام گرفت.
۳. مقیاس‌بندی ویژگی‌ها: همه ویژگی‌ها با تفریق میانگین و مقیاس‌بندی به واریانس واحد در مجموعه آموزشی مقیاس شدند. از همان میانگین و انحراف معیار دوباره برای مقیاس داده‌های آزمون استفاده می‌شود.
۴. یافتن تعداد خوشه‌های بهینه: با شروع از یک خوشه، به تعداد مناسبی از خوشه‌ها، کارانجام شد تا ده خوشه ، مجموعه آموزشی با الگوریتم خوشه‌بندی K-means مطابقت دارد و در مجموع مربع‌ها (WCSS) در مقابل تعداد خوشه‌ها رسم می‌شود. سپس از روش Elbow برای یافتن تعداد بهینه خوشه‌ها استفاده می‌شود[22].
۵.کاهش ابعاد با استفاده از PCA[1] : برای کاهش پیچیدگی محاسباتی و حذف ویژگی‌های غیر مؤثر، از روش تحلیل مؤلفه‌های اصلیPCA استفاده شد. این روش با تبدیل ویژگی‌ها به مؤلفه‌های اصلی غیرهمبسته، باعث کاهش ابعاد و حفظ بیش از ۹۵٪ واریانس داده‌ها گردید[23]. مرحله PCA پس از نرمال‌سازی داده‌ها و قبل از اجرای الگوریتم K-means انجام گرفت.

در ادامه، الگوریتم K-means تنها به‌عنوان ابزاری برای تعیین تعداد خوشه‌ها استفاده نشده، بلکه نقش اساسی در ساختاردهی داده‌ها و آماده‌سازی آن‌ها برای طبقه‌بند ایفا کرده است. در این پژوهش، K-means جهت گروه‌بندی داده‌های پیش پردازش‌شده NSL-KDD به خوشه‌هایی با ساختار همگن به‌کارگرفته‌شده که این خوشه‌بندی موجب کاهش پراکندگی نمونه‌ها و افزایش انسجام درون خوشه‌ای شده است. در نتیجه، داده‌های ورودی به الگوریتم طبقه‌بند XGBoost ساختاریافته‌تر و معنادارتر گردیده‌اند. این رویکرد مرحله‌ای، با تفکیک بهتر ساختار داده‌ها، منجر به بهبود چشمگیر در عملکرد طبقه‌بند شده است، به‌گونه‌ای که XGBoost توانسته بادقت بالاتری الگوهای مربوط به حملات سایبری را شناسایی کند؛ بنابراین، نقش الگوریتم K-means در این مدل، فراتر از صرف تعیین تعداد خوشه‌ها بوده و به‌عنوان یک گام حیاتی در بهبود اثربخشی فرایند طبقه‌بندی ایفای نقش نموده است.

۳-۳ الگوریتم‌های استفاده شده (XGBoost - K-means)

۳-۳-۱ الگوریتم K-means

برخی از رویه‌های مهم در روش k-means شامل انتخاب k نقطه به‌عنوان مرکز خوشه، محاسبه جداسازی بین k مرکز خوشه و سایر نقاط نمونه به طور جداگانه، و در نهایت تخصیص هر نقطه به سمت نزدیک‌ترین مرکز خوشه است. این رویه‌ها تا زمانی که شرایط تعلیق از پیش تعیین شده برآورده شود ادامه می‌یابد و از فاصله اقلیدسی برای محاسبه تفکیک بین مرکز خوشه و نقطه نمونه‌برداری استفاده می‌کند[24]. از فرمول ۱ برای محاسبه فرمول تطبیق استفاده شد که y1 و y2 دونقطه را بر اساس n عنصر و $d(y_1 , y_2)$ نشان دهنده فاصله اقلیدسی بین $(y_1 = y_{11}, y_{12}, \dots, y_{1n})$ و $(y_2 = y_{21}, y_{22}, \dots, y_{2n})$ برای تقسیم داده های از پیش‌پردازش شده به داده توزیع‌شده انعطاف‌پذیر[2]، استفاده می‌کنیم. K-means برای خوشه‌بندی هر RDD استفاده می‌شود[25] سپس تمام یافته‌های نهایی جمع‌آوری می‌شوند. روش زیر استفاده از الگوریتم XGBoost برای طبقه‌بندی هر خوشه است.

۳-۳-۲ الگوریتم افزایش گرادیان شدید[3]

1 Principal Component Analysis
2. Resilient distributed dataset (RDD)
3. XGBoosting

XGBoost یکی از پیشرفته‌ترین روش‌های Boosting است که برای مسائل دسته‌بندی بسیار مؤثر است، خصوصاً در مجموعه‌داده‌هایی با ویژگی‌های زیاد و نامتوازن، مانند داده‌های حملات سایبری. این الگوریتم با استفاده از تکنیک درخت تصمیم تقویت‌شده، خطای مدل را با هر تکرار کاهش می‌دهد و دارای قابلیت کنترل overfitting از طریق regularization است[26] و تکنیکی برای بهینه‌سازی تقویت درخت گرادیان با ساخت درخت‌های تصمیم‌گیری گام‌به‌گام است و در بسیاری از کامپیوترها، این ظرفیت را دارد که محاسبات مربوطه را با سرعت بیشتری انجام دهد. الگوریتمی تقویت‌کننده بر اساس درخت CART است که برای حل مشکلات طبقه‌بندی، از این روش استفاده می‌شود. مقدار مربوط به گره برگ درخت CART یک امتیاز واقعی است، نه یک دسته مشخص که منجر به تحقق الگوریتم بهینه‌سازی کارآمد با فرض وجود K درخت CART، می‌شود پس در نتیجه طبقه‌بندی نهایی توسط همه آنها یکپارچه می‌شود. فرایند محاسبه در فرمول ۱ نشان‌داده‌شده است که نشان‌دهنده خروجی درخت k-ام است.

(۱)

$$y_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F,$$

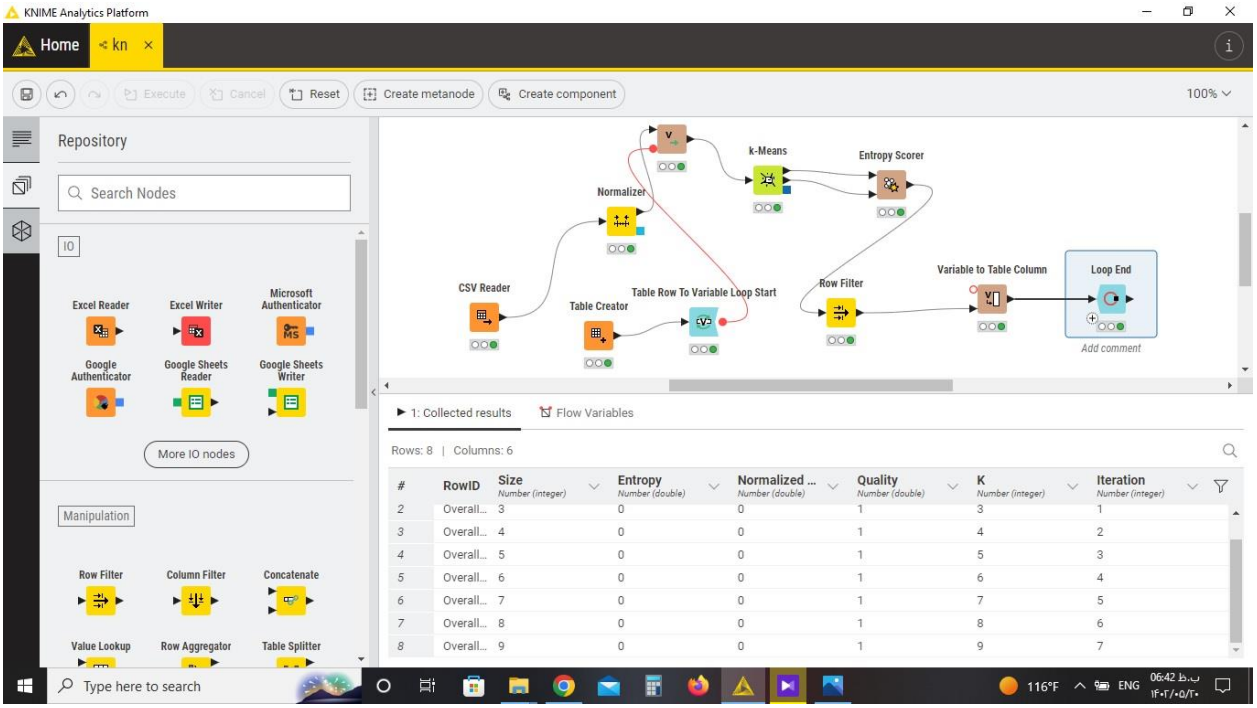
نمونه‌های عادی یا غیرعادی مجموعه‌داده‌ها از طریق XGBoost به دنبال خوشه‌بندی k-means در روش‌های قبلی طبقه‌بندی می‌شوند. فرایند قضاوت طبقه‌بندی در روش XGBoost در فرمول ۲ نشان داده شده است، که x و y مخفف یک رویداد طبقه‌بندی و دسته‌بندی مرتبط با آن هستند. $y_i(x)$ نشان‌دهنده خروجی طبقه‌بندی شده است. تابع نمایندگی توسط $P(y_i(x) = Y)$ نشان داده می‌شود و H (X) نیز نتیجه طبقه‌بندی با استفاده از روش XGBoost را نمایش می‌دهد.

(۲)

$$\mathrm{H}\ (\mathrm{X}) = \arg\max \sum_{i=1}^{k} P\ (y_i\ (x) = Y)$$

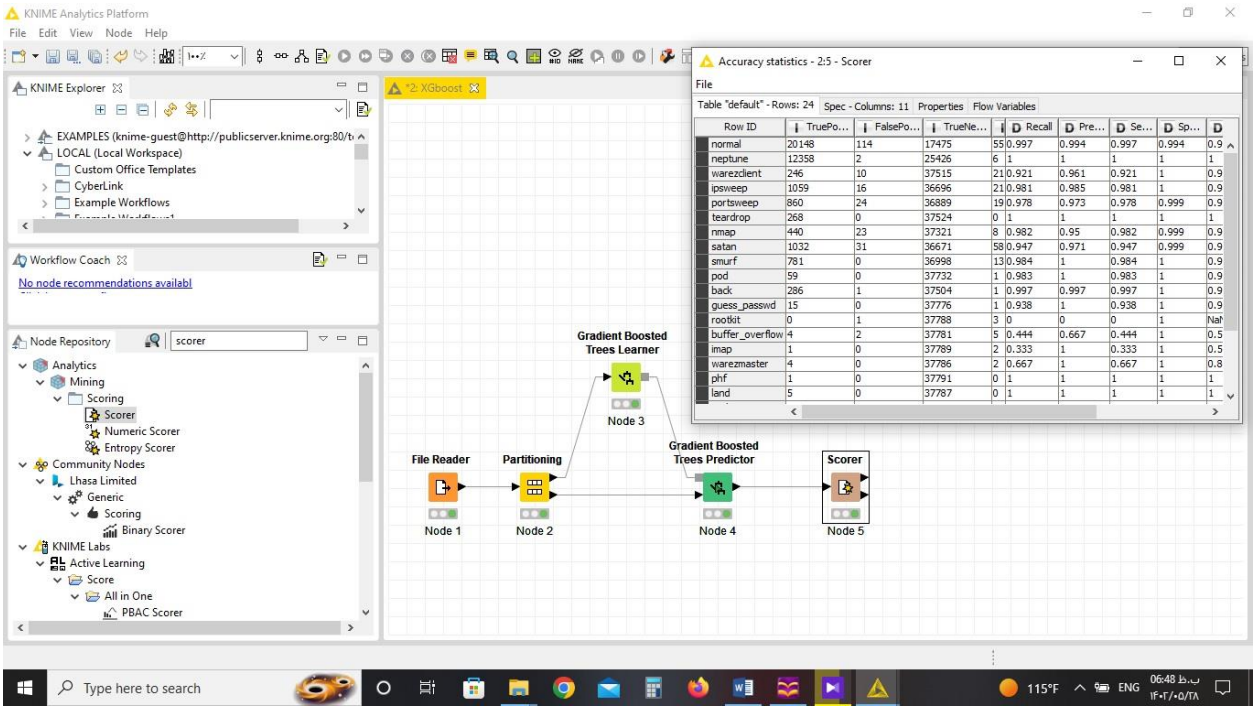
۴- روش انجام پژوهش

این تحقیق با استفاده از نرم‌افزار شبیه‌ساز knime که یکی از پلتفرم‌های شبیه‌سازی قدرتمند در مباحث داده‌کاوی است طراحی و پیاده‌سازی شده است. این نرم‌افزار به کاربران اجازه می‌دهد تا جریان داده‌ها را به‌صورت بصری ایجاد کنند، برخی یا تمام مراحل تجزیه‌وتحلیل را به‌صورت انتخابی اجرا کنند و نتایج و مدل‌ها را با استفاده از ویجت‌ها و نماهای تعاملی بررسی کنند. نایم به زبان جاوا و بر اساس اکلیپس نوشته شده است این ابزار قدرتمند تمامی روش‌های داده‌کاوی و الگوریتم‌های یادگیری ماشین را در درون خود جای‌داده است و می‌توان برای مباحث تشخیص نفوذ و داده‌کاوی و تست و پیاده‌سازی انواع الگوریتم‌های یادگیری ماشین از آن بهره برد. مراحل اجرا در شکل‌های ۲ تا ۵ نمایش‌داده‌شده است.



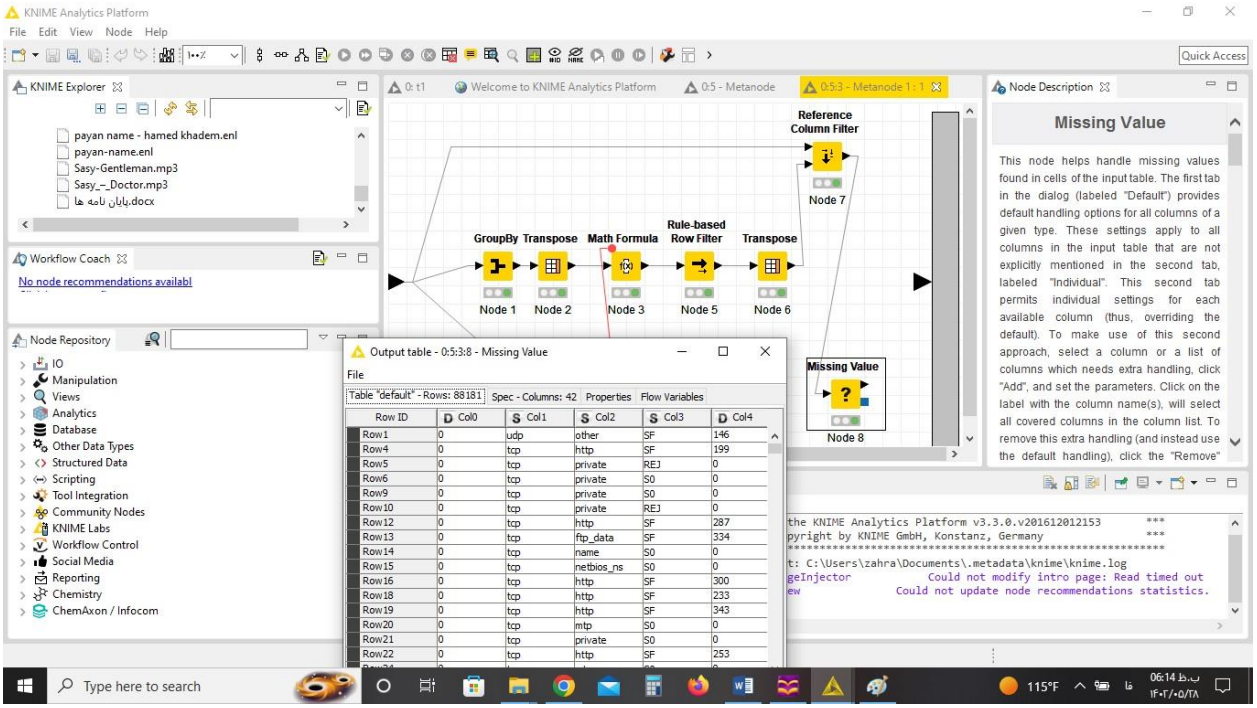
شکل ۲: پیاده‌سازی بخش‌هایی از الگوریتم kmeans-Xgboost در محیط پلت فرم knime

Figure 2: Implementation of parts of the K-means + XGBoost algorithm within the KNIME platform



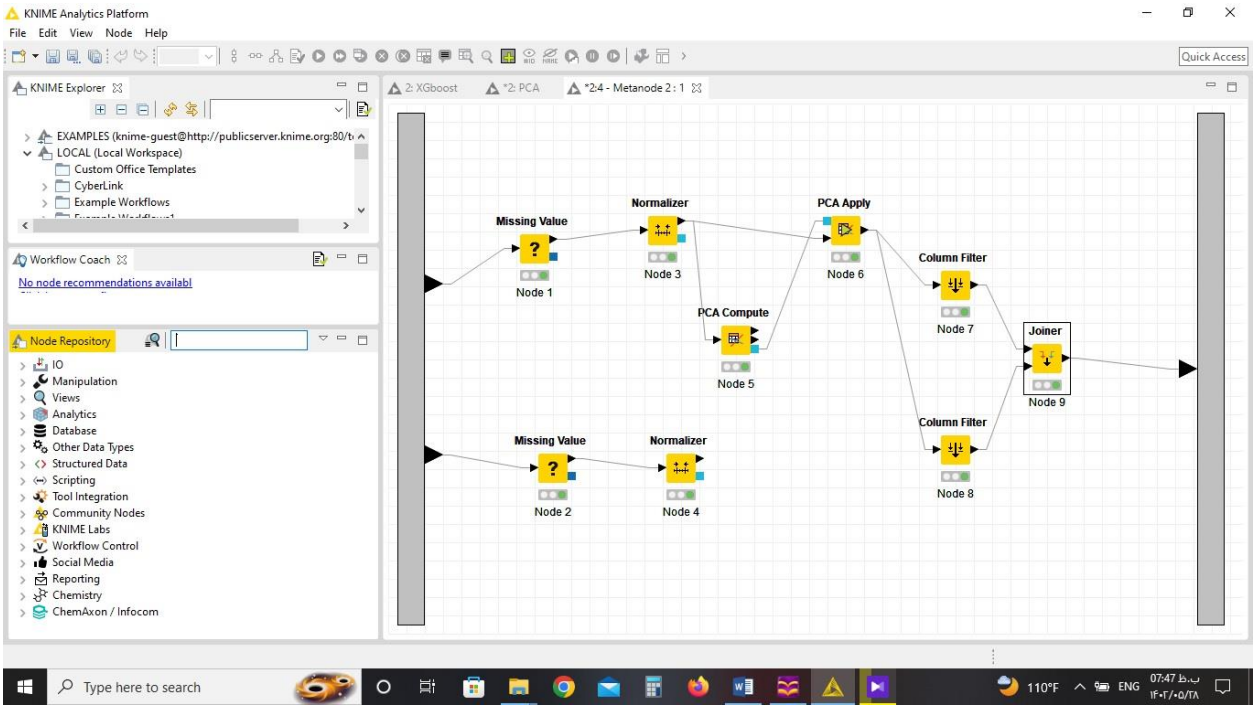
شکل ۳: پیاده سازی الگوریتم XGboost در پلت فرم knime

Figure 3: Implementation of the XGBoost algorithm in the KNIME platform



شکل ۴: محاسبه مقادیر missing value برای بدست آوردن مقدار آستانه

Figure 4: Calculation of missing-value ratios to determine a threshold.



شکل ۵: تصاویر بخش‌هایی از الگوریتم پیشنهادی و پیاده‌سازی با روش PCA در محیط knime

Figure 5: Screenshots of parts of the proposed algorithm and its implementation using PCA in the KNIME environment

۴-۱- بهینه‌سازی فراپارامترهای مدل K-means-XGBoost

این بخش به‌عنوان نوآوری تحقیق است که فراپارامترهای مدل پیشنهادی را برای دریافت بهترین نتایج تنظیم کردیم. همچنین پارامتر ویژگی_rate را برای انتخاب ویژگی بهینه کردیم، مانند تعداد خوشه‌های k که به طور مستقیم به‌عنوان پارامتر ناهنجاری

m مشخص می‌شود، نرخ یادگیری الگوریتم XGBoost ، را با eta به‌عنوان ناهنجاری و es_train و es_test که با مجموعه‌داده‌های آموزشی و مجموعه‌داده‌های تست مطابقت دارد. جدول ۲ فراپارامترهای جمع‌آوری شده از نتایج تجربی در مجموعه داده NSK-KDD را فهرست می‌کند.

جدول ۲ : لیست فراپارامترهای سیستم تشخیص نفوذ پیشنهادی
Table 2: List of Hyperparameters for the Proposed Intrusion Detection System

Parameters	Settings
K	۸
initSteps	۲۵
maxIter	۱۰۰
attribute_rate	۰/۰۱
M	۲۵
num_round	۱۰۰
max_depth	۲۰
eta	۰/۳
es_train	۰/۵
es_test	۰/۰۰۰۱۳

۲-۴- معیارهای ارزیابی الگوریتم پیشنهادی

ماتریس درهم‌ریختگی[1] : یک مقیاس تحلیل برای طبقه‌بندی به‌صورت ماتریس است که از چهارعنصر TP و TN که طبقه‌بندی صحیح را نشان می‌دهند و FP و FN که طبقه‌بندی اشتباه را نشان می‌دهد تشکیل می‌شود که در جدول ۳ نشان‌داده‌شده است.

جدول ۳ : عناصر تشکیل‌دهنده ماتریس درهم‌ریختگی
Table 3: Components of the Confusion Matrix

		The predicted class		
the actual class		Yes	No	Sum
	yes	TP	FN	P
	no	FP	TN	N
	sum	P'	N'	P+N

۳-۴- خروجی سیستم تشخیص نفوذ پیشنهادی K-means-XGboost

۱-۳-۴- جدول ماتریس درهم‌ریختگی الگوریتم پیشنهادی

در جدول ۴ شکل ماتریس درهم‌ریختگی با استفاده از الگوریتم پیشنهادی سیستم تشخیص نفوذ بر اساس K-meansو XGboost نمایش داده شده‌است که با استفاده از داده‌های آموزشی NSL-KDD محاسبه شده است.

1. Confusion Matrix

جدول ۴ : نمایش داده‌های ماتریس درهم‌ریختگی با استفاده از روش پیشنهادی
Table 4: Display of Confusion Matrix Data Using the Proposed Method

Attacks	DoS	Normal	Probe	R2L	U2R
Dos	۷۸۱۳۹	۸	۰	۰	۰
Normal	۱	۱۹۶۱۸	۳	۵	۰
Probe	۲	۶	۸۰۵	۰	۰
R2L	۰	۱۱	۰	۱۹۳	۰
U2R	۰	۳	۰	۳	۷

در این مرحله از داده‌های آموزشی در محیط شبیه‌سازی نرم‌افزار Knime پیاده‌سازی انجام گردید و نتایج میزان یادگیری الگوریتم XGboost با و بدون استفاده از خوشه‌بندی اندازه‌گیری و تحلیل گردید که نتایج به‌صورت جداول ۵ و ۶ می‌باشد.

جدول ۵ : میزان پیش‌بینی نتایج الگوریتم XGboost با استفاده از خوشه‌بندی
Table 5: Prediction Performance of the XGBoost Algorithm Using Clustering

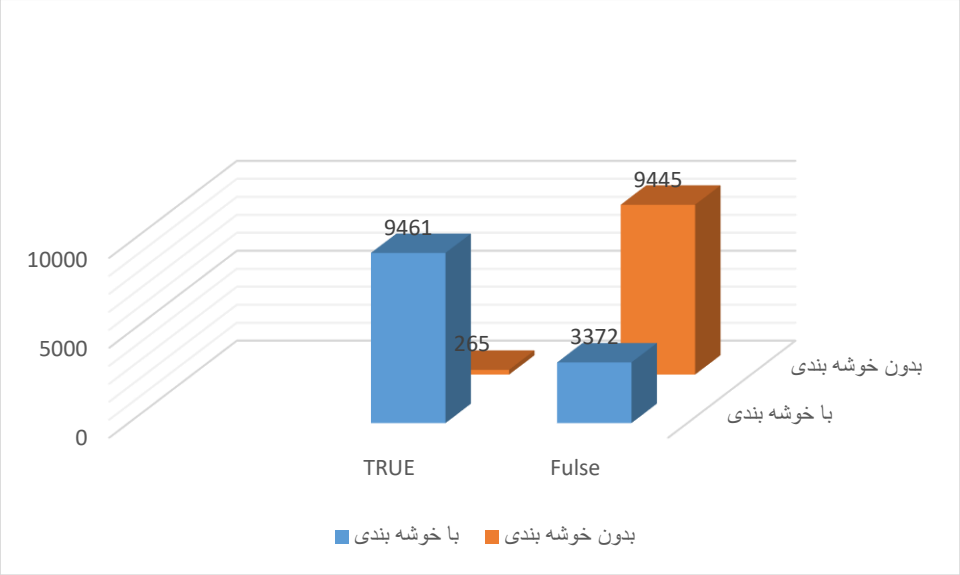
پیش‌بینی واقعی	درست	غلط
درست	۹۴۶۱	۳۳۷۲
غلط	۲۶۵	۹۴۴۵

جدول ۶ : میزان پیش‌بینی نتایج الگوریتم XGboost بدون استفاده از خوشه‌بندی
Table 6: Prediction Performance of the XGBoost Algorithm Without Clustering

پیش‌بینی واقعی	درست	غلط
درست	۸۷۵۵	۴۰۷۸
غلط	۷۴۴	۸۹۶۶

۲-۳-۴- نتایج طبقه‌بندی باینری بر اساس مدل پیشنهادی K-means–XGBoost

جدول ۷ و شکل ۶ به ترتیب معیارهای ارزیابی و ماتریس درهم‌ریختگی را بر روی مجموعه‌های آزمایشی نشان می‌دهند. نتایج گویای این است که نرخ درستی الگوریتم برابر ۰.۹۹۸۶ ، میزان دقت الگوریتم برابر ۰.۹۹۸۶درصد، میزان حساسیت یا یادآوری الگوریتم ۰.۹۹۸۶ ، و نرخ تشخیص خطا با استفاده از روش ترکیبی پیشنهادی به ۰.۹۹۲۲ می‌رسد که نشان دهنده اکثریت بزرگی است که در آن فعالیت‌های ناهنجاری به صورت کامل قابل‌شناسایی هستند. همچنین به طور همزمان، نرخ هشدار کاذب ۰.۱۴۷۰برای فعالیت‌های ناهنجار دیده می‌شود.



شکل ۶: مقایسه میزان صحت پیش بینی الگوریتم XGBoost با و بدون استفاده از خوشه بندی kmeans
Figure 6: Comparison of XGBoost prediction accuracy with and without K-means clustering

جدول ۷ : معیارهای ارزیابی داده‌های آزمایشی روش K-means-Xgboost
Table 7: Evaluation Metrics for K-Means-XGBoost Method on Test Data

model	Precision	Recall	Accuracy	F-Score	DR	FAR
K-means-XGboost	۰.۹۹۸۶	۰.۹۹۸۶	۰.۹۹۸۶	۰.۹۹۸۶	۰.۹۹۲۲	۰.۱۴۷۰

در نتیجه این موضوع نشان‌دهنده تأثیر بسیار خوب روش پیشنهادی K-means-XGBoost است. همچنین افزایش دردسترس‌بودن IDS ممکن است تحت‌تأثیر DR بالاتر و FAR کمتر باشد.

۳-۳-۴- مقایسه مدل پیشنهادی با سایر روش‌های ترکیبی

در این بخش، مدل پیشنهادی K-means–XGBoost با روش‌های پیشرفته دیگر در حوزه تشخیص نفوذ مقایسه شده است. نتایج این مقایسه در جدول ۸ و شکل ۷ ارائه شده‌اند. مدل‌های مورد مقایسه شامل روش‌های مبتنی بر یادگیری عمیق، خودآموز، و ترکیبی هستند که همگی با استفاده از مجموعه داده NSL-KDD ارزیابی شده‌اند. معیارهای مورد بررسی شامل دقت (Accuracy)، نرخ تشخیص(Detection Rate - DR) ، و نرخ هشدار کاذب (False Alarm Rate - FAR) هستند.

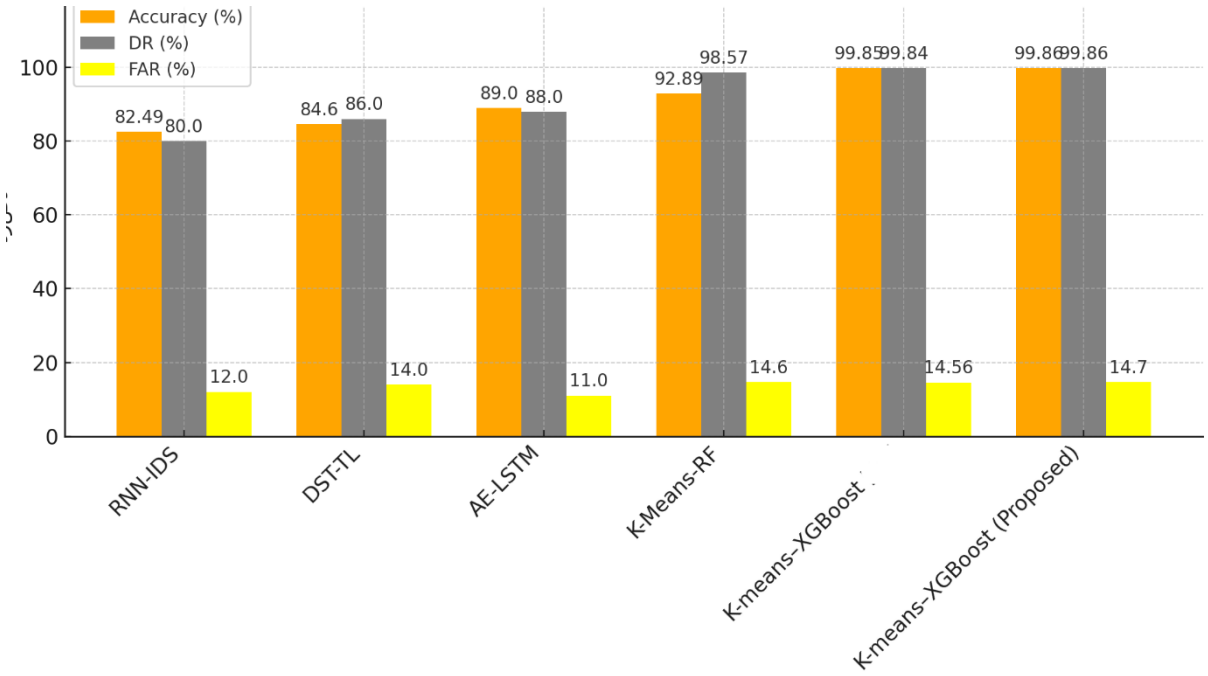
جدول ۸ نشان می‌دهد که مدل پیشنهادی K-means–XGBoost با دقت ۹۹.۸۶درصد، نرخ تشخیص ۹۹.۸۶درصد، و نرخ هشدار کاذب ۱۴.۷۰درصد، عملکرد بهتری نسبت به سایر روش‌ها دارد. اگرچه نرخ هشدار کاذب مدل پیشنهادی نسبت به برخی روش‌ها مانند AE-LSTM با FAR 11% بالاتر است، اما دقت و نرخ تشخیص بالاتر آن، این مدل را به گزینه‌ای مناسب‌تر برای تشخیص نفوذ تبدیل می‌کند. در ادامه، مقایسه دقیق‌تری با هر یک از مدل‌ها ارائه می‌شود:

- مقایسه با [14] مدل AE-LSTM که مبتنی بر یادگیری عمیق و شبکه‌های عصبی بازگشتی است، به دقت ۸۹درصد و نرخ تشخیص ۸۸درصد دست یافته است. مدل پیشنهادی K-means–XGBoost نسبت به AE-LSTM به ترتیب ۱۰.۸۶درصد در دقت و ۱۱.۸۶درصد در نرخ تشخیص بهبود یافته است، که نشان‌دهنده برتری قابل‌توجه آن در

شناسایی حملات است. با این حال، نرخ هشدار کاذب (11%) AE-LSTM کمتر از مدل پیشنهادی ۱۴.۷۰درصد است، که نشان می‌دهد مدل پیشنهادی در کاهش خطاهای تشخیص کاذب نیاز به بهبود دارد.

- مقایسه با [15] مدل K-Means-RF که ترکیبی از خوشه‌بندی K-means و طبقه‌بندی جنگل تصادفی است، دقت ۹۲.۸۹درصد و نرخ تشخیص ۹۸.۵۷درصد را ارائه کرده است. مدل پیشنهادی نسبت به K-Means-RF از نظر دقت ۶.۹۷درصد و از نظر نرخ تشخیص ۱.۲۹درصد بهبود یافته است. همچنین، نرخ هشدار کاذب مدل پیشنهادی ۱۴.۷۰درصد اندکی بالاتر از K-Means-RF (14.6%) است، که نشان می‌دهد هر دو مدل در این معیار چالش مشابهی دارند.
- مقایسه با [16] مدل RNN-IDS که از شبکه‌های عصبی بازگشتی برای تشخیص نفوذ استفاده می‌کند، دقت ۸۲.۴۹درصد و نرخ تشخیص ۸۰درصد را به دست آورده است. مدل پیشنهادی K-means–XGBoost نسبت به RNN-IDS به ترتیب ۱۷.۳۷درصد در دقت و ۱۹.۸۶درصد در نرخ تشخیص عملکرد بهتری دارد. نرخ هشدار کاذب RNN-IDS (12%) کمتر از مدل پیشنهادی است، اما دقت و نرخ تشخیص پایین‌تر آن، کارایی کلی آن را محدود می‌کند.
- مقایسه با [17] مدل DST-TL که یک IDS مبتنی بر یادگیری خودآموز با استفاده از رمزگذار خودکار پراکنده است، به دقت ۸۴.۶۰درصد و نرخ تشخیص ۸۶درصد دست یافته است. مدل پیشنهادی نسبت به DST-TL به ترتیب ۱۵.۲۶درصد در دقت و ۱۳.۸۶درصد در نرخ تشخیص بهبود یافته است. نرخ هشدار کاذب DST-TL (14%) اندکی کمتر از مدل پیشنهادی ۱۴.۷۰درصد است، که نشان‌دهنده عملکرد مشابه این دو مدل در این معیار است.
- مقایسه با [18] مدل K-means–XGBoost ارائه‌شده توسط Li و همکاران (۲۰۲۰) به دقت ۹۹.۸۵درصد، نرخ تشخیص ۹۹.۸۴درصد، و نرخ هشدار کاذب ۱۴.۵۶درصد دست یافته است. مدل پیشنهادی این پژوهش با دقت ۹۹.۸۶درصد و نرخ تشخیص ۹۹.۸۶درصد، اندکی (۰.۰۱درصد در دقت و ۰.۰۲درصد در نرخ تشخیص) از مدل [19] پیشی گرفته است. نرخ هشدار کاذب مدل پیشنهادی (۱۴.۷۰درصد) نیز تنها ۰.۱۴درصد بیشتر از مدل [19] است. این بهبود اندک ناشی از بهینه‌سازی فراپارامترها و استفاده از تحلیل مؤلفه‌های اصلی (PCA) برای پیش‌پردازش داده‌ها در روش پیشنهادی است.

شکل ۷ مقایسه بصری معیارهای دقت، نرخ تشخیص، و نرخ هشدار کاذب مدل پیشنهادی با سایر روش‌ها را نشان می‌دهد. همان‌طور که مشاهده می‌شود، مدل K-means–XGBoost پیشنهادی در معیارهای دقت و نرخ تشخیص عملکرد برتری نسبت به تمام مدل‌های مورد مقایسه دارد. با این حال، نرخ هشدار کاذب بالاتر آن نسبت به برخی مدل‌ها مانند AE-LSTM و RNN-IDS نشان می‌دهد که بهبود این معیار می‌تواند موضوع پژوهش‌های آینده باشد. در مجموع، نتایج نشان می‌دهند که ترکیب هدفمند خوشه‌بندی K-means و طبقه‌بندی XGBoost با پیش‌پردازش پیشرفته، کارایی سیستم‌های تشخیص نفوذ را به‌طور قابل‌توجهی ارتقا می‌دهد.



شکل ۷: مقایسه معیارهای دقت، درستی و اندازه گیری الگوریتم پیشنهادی با سایر الگوریتم‌ها
Figure 7: Comparison of accuracy, precision, and recall metrics of the proposed algorithm with other algorithms

جدول ۸ : مقایسه عملکرد K-means و XGBoost پیشنهادی با سایر مدل‌های ترکیبی قبلی
Table 8: Performance Comparison of the Proposed K-Means-XGBoost Model with Other Hybrid Models

Model	Accuracy (%)	DR (%)	FAR (%)
AE-LSTM[14]	۸۹	۸۸	۱۱
K-Means-RF[15]	۹۲.۸۹	۹۸.۵۷	۱۴.۶
RNN-IDS[16]	۸۲.۴۹	۸۰	۱۲
DST-TL[17]	۸۴.۶۰	۸۶	۱۴
K-means–XGBoost[18]	۹۹.۸۵	۹۹.۸۴	۱۴.۵۶
K-means–XGBoost(propose)	۹۹.۸۶	۹۹.۸۶	۱۴.۷۰

روش‌های ترکیبی بر روی مجموعه‌داده‌های رایج انجام می‌شوند، اما با این فکر که استراتژی‌های حمله جدید و داده‌های اضافی از منابع مختلف در آینده وجود خواهد داشت، نیاز به بهبود حیاتی است. بهبود مدل را برای طبقه‌بندی سریع حملات مختلف در نظر می‌گیریم. با ترکیب چند تکنیک داده‌کاوی و تکنیک‌های تقویت گرادیان به یک روش بهینه برای داده‌کاوی دست پیدا کردیم که در انتها الگوریتم پیشنهادی را با سایر الگوریتم‌ها با معیارهای مختلف مقایسه کرده و در یک محیط شبیه‌سازی شده و با داده‌های آزمایشی و آموزشی پیاده‌سازی را انجام دادیم.

ارزیابی‌ها نشان دادند که الگوریتم XGBoost زمانی که از الگوریتم خوشه‌بندی K-means استفاده کنیم پیش‌بینی صحیح نتایج تشخیص نفوذ بسیار بهتر از حالت بدون خوشه‌بندی است. همچنین نرخ تشخیص خطا و نرخ درستی و نرخ دقت الگوریتم با استفاده از روش ترکیبی پیشنهادی به ۹۹.۸۶٪ می‌رسد که نشان دهنده اکثریت بزرگی است که در آن فعالیت‌های ناهنجاری

به صورت کامل قابل‌شناسایی هستند. همچنین به طور همزمان، نرخ هشدار کاذب پایین برای فعالیت‌های ناهنجار دیده می‌شود. به‌صورت خلاصه می‌توان گفت که روش پیشنهادی بر حسب تکرار فرآیند انتخاب ویژگی به دلیل بهینه‌سازی ویژگی‌ها موفق بوده است و میزان خطای تشخیص نفوذ به خطا را کاهش داده است.

۵- نتیجه‌گیری و پیشنهادهای پژوهش:

یافته‌های حاصل از اجرای آزمایش‌های مدل پیشنهادی در حوزه سیستم‌های تشخیص نفوذ شبکه‌ای، کارایی برجسته رویکرد ترکیبی K-means–XGBoost را تأیید می‌کند. نتایج تجربی که با بهره‌گیری از مجموعه‌داده استاندارد NSL-KDD و پیاده‌سازی در محیط KNIME به دست آمده‌اند، حاکی از بهبود قابل‌توجه این مدل در معیارهای کلیدی ارزیابی، از جمله دقت و نرخ تشخیص، نسبت به روش‌های پیشین است. این الگوریتم با استفاده از خوشه‌بندی K-means برای سازمان‌دهی داده‌ها، کاهش ابعاد با تحلیل مؤلفه‌های اصلی PCA و طبقه‌بندی پیشرفته با XGBoost همراه با تنظیم دقیق فراپارامترها، دقت ۹۹.۸۶درصد و نرخ تشخیص ۹۹.۸۶درصد را محقق ساخته است. مقایسه عملکرد مدل پیشنهادی با روش‌های موجود نشان‌دهنده برتری آن است. در مقایسه با مدل AE-LSTM که دقت ۸۹درصد و نرخ تشخیص ۸۸درصد را گزارش کرده است، مدل پیشنهادی بهبود ۱۰.۸۶درصد در دقت و ۱۱.۸۶درصد در نرخ تشخیص را به نمایش گذاشته است. نسبت به مدل K-Means-RF، که به ترتیب دقت ۹۲.۸۹درصد و نرخ تشخیص ۹۸.۵۷درصد را ارائه کرده است، بهبود ۶.۹۷درصد در دقت و ۱.۲۹درصد در نرخ تشخیص مشاهده شده است. همچنین، در مقایسه با مدل‌های RNN-IDS دقت ۸۲.۴۹درصد، نرخ تشخیص ۸۰درصد و DST-TL دقت ۸۴.۶۰ درصد، نرخ تشخیص ۸۶درصد، بهبودهای قابل‌توجهی به ترتیب ۱۷.۳۷درصد و ۱۵.۲۶درصد در دقت و ۱۹.۸۶درصد و ۱۳.۸۶درصد در نرخ تشخیص به دست آمده است. در برابر مدل K-means–XGBoost پیشین، که دقت ۹۹.۸۵درصد و نرخ تشخیص ۹۹.۸۴درصد را گزارش کرده است، مدل پیشنهادی با بهره‌گیری از پیش‌پردازش PCA و بهینه‌سازی فراپارامترها، بهبود اندک اما معنادار ۰.۰۱درصد در دقت و ۰.۰۲درصد در نرخ تشخیص را نشان می‌دهد. با این حال، نرخ هشدار کاذب مدل پیشنهادی ۱۴.۷۰درصد نسبت به برخی روش‌ها، نظیر AE-LSTM (11%)و RNN-IDS (12%)، بالاتر است، که بیانگر پتانسیل بهبود در این معیار دارد.

این پژوهش با ارائه یک چارچوب ترکیبی که از خوشه‌بندی، کاهش ابعاد، و طبقه‌بندی پیشرفته بهره می‌برد، گامی مؤثر در ارتقای کارایی سیستم‌های تشخیص نفوذ برداشته است. این رویکرد، الگویی انعطاف‌پذیر برای کاربرد در سناریوهای عملیاتی پیچیده فراهم کرده و پتانسیل تعمیم‌پذیری به محیط‌های متنوع را داراست. بر اساس نتایج به‌دست‌آمده، پیشنهاد می‌شود در تحقیقات آتی، این مدل در بسترهای بلادرنگ با استفاده از مجموعه‌داده‌های متنوع‌تر، نظیر داده‌های شبکه‌های صنعتی یا اینترنت اشیا، مورد ارزیابی قرار گیرد تا قابلیت تعمیم آن در شرایط واقعی سنجیده شود. همچنین، بهره‌گیری از تکنیک‌های پیشرفته کاهش نرخ هشدار کاذب، نظیر وزن‌دهی کلاس‌ها در الگوریتم‌های تقویت‌شده یا ادغام با الگوریتم‌های یادگیری عمیق، می‌تواند دقت و انعطاف‌پذیری مدل را بهبود بخشد. علاوه بر این، روش‌های انتخاب ویژگی مبتنی بر بهینه‌سازی تکاملی و توسعه ساختارهای خود یادگیرنده می‌توانند به کاهش نرخ هشدار کاذب و سازگاری با حملات نوظهور کمک کنند. از سوی دیگر، انجام تحلیل جامع‌تری در مورد زمان اجرا و مصرف منابع محاسباتی، به‌ویژه در سامانه‌های توزیع‌شده یا محیط‌هایی با محدودیت‌های پردازشی، می‌تواند به کاربردی‌سازی گسترده‌تر این رویکرد کمک کند.

مراجع:

[1] Khan, S., E. Sivaraman, and P.B. Honnavalli. "Performance evaluation of advanced machine learning algorithms for network intrusion detection system". in Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India,2020 .Springer. DOI: 10.1007/978-981-15-3020-3_6.

[2] Zhao, X., "Application of data mining technology in software intrusion detection and information processing". Wireless Communications and Mobile Computing, 2022. DOI:10.1155/2022/3829160.

[3] Zhu, Y., et al., "Application of data mining technology in detecting network intrusion and security maintenance". Journal of Intelligent Systems, 2021. 30(1): p. 664-676. DOI:10.1155/2022/3829160.

[4] Shahjee, D. and N. Ware, "Integrated network and security operation center: A systematic analysis". IEEE Access, 2022. 10: p. 27881-27898.DOI: 10.1109/ACCESS.2022.3157738.

[5] Yang, L. and A. Shami, "IoT data analytics in dynamic environments: From an automated machine learning perspective". Engineering Applications of Artificial Intelligence, 2022. 116: p. 105366. https://doi.org/10.1016/j.engappai.2022.105366.

[6] Khalil, R.A., et al., "Deep learning in the industrial internet of things: Potentials, challenges, and emerging applications". IEEE Internet of Things Journal, 2021. 8(14): p. 11016-11040. DOI: 10.1109/JIOT.2021.3051414.

[7] Yang, L. and A. Shami. "A transfer learning and optimized CNN based intrusion detection system for Internet of Vehicles". in ICC-۲۰۲۲ IEEE International Conference on Communications. 2022. IEEE. DOI: https://doi.org/10.1109/ICC45855.2022.9838780.

[8] Sangkatsanee, P., N. Wattanapongsakorn, and C. Charnsripinyo, "Practical real-time intrusion detection using machine learning approaches". Computer Communications, 2011. 34(18): p. 2227-2235. DOI: 10. 016/j.comcom.2011.07.001.

[9] Axelsson, S., "The base-rate fallacy and the difficulty of intrusion detection". ACM Transactions on Information and System Security (TISSEC), 2000. 3(3): p. 186-205 DOI: 10.1145/319709.319710.

[10] [10] Y. Y. Aung and M. M. Min, "Analysis of K-means Clustering Algorithm for Intrusion Detection System", Advances in Science, Technology and Engineering Systems Journal, vol. 3, no. 1, pp. 372–377, 2018. [Online]. Available: https://www.astesj.com/v03/i01/p60.

[11] Lee, W., S.J. Stolfo, and K.W. Mok. "A data mining framework for building intrusion detection models". in Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat. No. 99CB36344). 1999. IEEE.DOI: 10.1109/SECPRI.1999.766909.

[12] X. Li, Y. Wang, and Z. Zhang, "Process mining in host-based intrusion detection systems," IEEE Trans. Dependable Secure Comput., vol. 18, no. 4, pp. 1234–1245, 2021.

[13] J. Zhang, M. Zulkernine, and A. Haque, "Random-Forests-Based Network Intrusion Detection Systems," IEEE Trans. Syst., Man, Cybern. C (Appl. Rev.), vol. 38, no. 5, pp. 649–659, 2008.

[14] Y. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," IEEE Access, vol. 5, pp. 21954–21961, 2017.

[15] J. Zhang, M. Zulkernine, and A. Haque, "Random-Forests-Based Network Intrusion Detection Systems," IEEE Trans. Syst., Man, Cybern. C (Appl. Rev.), vol. 38, no. 5, pp. 649–659, 2008

[16] Y. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," IEEE Access, vol. 5, pp. 21954–21961, 2017.

[17] W. Lin, S. Wang, W. Zhang, and Y. Zhou, "A hybrid deep learning model for network intrusion detection," Electronics, vol. 8, no. 4, p. 438, 2019.

[18] H. Li, Y. Li, and T. Li, "A hybrid intrusion detection method based on K-means and XGBoost," in Proc. 15th Int. Conf. Computer Science & Education (ICCSE), 2020, pp. 108–112.

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[20]حمیدرضا مدرس و علیرضا احمدی فرد، «روشی جدید برای خوشه‌بندی غیرنظارتی داده‌ها به کمک الگوریتم بهینه‌سازی«PSO ، در شانزدهمین کنفرانس مهندسی برق ایران، تهران، ۱۳۸۷ . [Online]. Available:

https://civilica.com/doc/47546سیویلیکا، مقالات علمی کنفرانس و ژورنال2+سیویلیکا، مقالات علمی کنفرانس و ژورنال2+Iranian Conference Journals2+

[21]مرجان محمودی و مهران محمدی قلعه سفیدی، «مروری بر انواع روش‌های خوشه‌بندی در یادگیری ماشین»، در بیست و چهارمین کنفرانس ملی مهندسی برق، کامپیوتر و مکانیک، شیروان، ۱۴۰۳ :Available: [Online]. .
https://civilica.com/doc/2159268سیویلیکا، مقالات علمی کنفرانس و ژورنال Iranian Conference1+ Journals+1

[22] سیمین علی اسماعیلی و امیر رجبی بهجت، «مقایسه دسته‌بندی و خوشه‌بندی جریان داده‌ها در سیستم تشخیص نفوذ با استفاده از شبکه عصبی و الگوریتم K-means»، در کنفرانس بین‌المللی نوآوری در علوم و تکنولوژی، ۱۴۰۰. Available: https://isnac. r/XYFD-BBCEGISNAC.

[23]. I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” Philos. Trans. Royal Soc. A, vol. 379, no. 2191, p. 20200202, 2021, doi: 10.1098/rsta.2020.0202.

[24]. H. Lv, X. Ji, and Y. Ding, "A Mixed Intrusion Detection System utilizing K-means and Extreme Gradient Boosting," J. Phys.: Conf. Ser., vol. 2517, p. 012016, 2023, doi: 10.1088/1742-6596/2517/1/012016.

[25]. J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-means clustering algorithm,” Appl. Stat., vol. 28, no. 1, pp. 100–108, 1979, doi: 10.2307/2346830.

[26]. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM