Karafan

Quarterly Scientific Journal of Technical and Vocational University

Autumn 2023, Vol. 20, No. 3, p. 509-529

Journal Homepage: https://karafan.tvu.ac.ir/?lang=en

TVU

10.48301/KSSA.2023.389398.2481

Network Intrusion Detection Using Thermal Exchange Optimization and Seagull Optimization Algorithm

Mona Emadi[1*], Mahmoud Niaei[2]

[1]Assistant Professor, Department of Computer Engineering, Payame Noor University, Tehran, Iran.
[2]PhD of Information Technology Management, Environment Protection Agency, Tehran, Iran.

ARTICLE INFO

Article Type:
Original Research

Received: 04.01.2023
Revised: 07.01.2023
Accepted: 09.05.2023

Keyword:
Intrusion Detection
Thermal Exchange Optimization
Seagull Optimization Algorithm
Random Forest
CICIDS-2017

***Corresponding Author:**
Mona Emadi
Email: *emadi.mona@pnu.ac.ir*

ABSTRACT

The increasing growth of computers and the internet has provided a new and widely used platform for providing network services. This has significantly increased the provision of administrative, social, financial, educational and recreational services on the network, particularly the internet. The expansion of the use of Internet applications creates an opportunity to abuse the network and its information for criminal purposes. Based on this, intrusion into the network and unauthorized access to information have become the main concerns of network users as well as network managers. Intrusion detection systems include a set of tools and mechanisms for monitoring computer systems and network traffic. Various methods are used for intrusion detection, such as statistical techniques, cognitive-based methods, and machine learning methods. In the present research, a method for intrusion detection using machine learning algorithms was reviewed and proposed. The proposed model is a multi-class method that, in addition to intrusion detection, also determines the type of attack. This method is a hybrid model in which the combination of the Seagull optimization algorithm, thermal exchange optimization algorithms and random forest algorithm are used. CICIDS-2017 dataset was used for analysis in this research. The proposed method was compared with several different algorithms and the accuracy value of the proposed method was equal to 98.8, which is higher than that of many machine learning methods.

E-ISSN: 2538-4430
ISSN: 2382-9796

EXTENDED ABSTRACT

Introduction

Various methods have been proposed by researchers to detect intrusion in the network, but greater research was needed to increase the accuracy and speed of detection. In this research, a network intrusion detection method was introduced. In the proposed method, the hybrid of two optimization algorithms, seagull optimization algorithm and thermal exchange optimization were used to select features and reduce the dimensions of the data set. In addition, the classifying attacks in the proposed method was performed by the random forest algorithm. After testing different methods, the hybrid of the above algorithms was selected due to appropriate performance and the level of accuracy, precision, readability, high F1-Score and low training time. CICIDS2017 dataset which is a new dataset was used in this research. The strengths of the method proposed can be summarized as follows:

1- Using a new feature selection method in intrusion detection.
2- The proposed method is multi class.
3- Using the new CICIDS2017 dataset.

Methodology

The proposed method used a new optimization algorithm for feature selection. This is a hybrid algorithm which combines two optimization algorithms: Seagull Optimization Algorithm (SOA) and Thermal Exchange Optimization Algorithm (TEO). The SOA algorithm has a good global search ability, while the TEO algorithm has a strong local search ability. In order to improve the search ability of both algorithms, the hybrid optimization algorithm was used to select the features. In addition, Random Forest Algorithm was applied for classifying attacks. The Forest Algorithm was used to increase the accuracy and speed of detection. In this research, the CICIDS-2017 dataset was used as one of the most recent datasets. The CICIDS-2017 dataset was designed by the Cyber Security Institute of Canada to solve the problem of lack of up-to-date and valid datasets for intrusion detection and collected for IPS and IDS design.

Results and discussion

The evaluation criteria used in this research were accuracy, recall, F1-score and time. After data preprocessing steps, the data were divided into two categories of features (X) and labels (Y). Then 10 percent of the data set were randomly and proportionately selected so that there was an equal proportion of labels, named X_train, y_train, X_test and y_test. Then, the learning algorithms were applied to X_train and y_train. ACC results for LR, DT, KNN, RF, GB, NN, RNN, GRU and LSTM algorithms were recorded as per Table 1.

Table 1. Accuracy of learning algorithms on primary data.

Algorithm	LSTM	GRU	RNN	NN	GB	RF	KNN	DT	LR
ACC	0/781	0/79	0/816	0/742	0/896	0/912	0/89	0/522	**0/761**

The execution time of the mentioned algorithms (in seconds) is shown in Table 2.

Table 2. Algorithm execution time (seconds) on primary data.

Algorithm	LSTM	GRU	RNN	NN	GB	RF	KNN	DT	LR
Training time	3002	2647	1808	584	1623	**19/4**	471	59	378
Testing time	42	39	36	18	12	0/17	183	0/9	0/3

As can be seen in Table 2, the training time of the random forest algorithm was 19.4 seconds, which was the fastest method under the same conditions as other algorithms. In Table 3, the arithmetic mean of the F1-Score, readability and accuracy criteria for the investigated methods can be seen on the primary dataset.

Table 3. Arithmetic mean of F1-Score/ readability /accuracy with primary data.

Algorithm	LSTM	GRU	RNN	NN	GB	RF	KNN	DT	LR
F1-Score	0/13	0/17	0/2	0/1	0/72	**0/79**	0/8	0/409	0/2
Readability	0/11	0/16	0/19	0/13	0/78	**0/819**	0/806	0/429	0/21
Accuracy	0/1	0/37	0/1	0/11	0/69	**0/81**	0/8	0/54	**0/18**

CICIDS-2017 dataset has 70 features after pre-processing. Feature selection by hybrid of thermal exchange optimization and seagull optimization algorithm resulted in the selection of 32 features. The features selection are presented in Table 4.

Table 4. Feature selection by hybrid of seagull optimization algorithm and heat exchange algorithm.

Features selection									
12	1	2	5	6	7	12	15	18	19
24	26	27	44	30	34	37	39	41	44
47	48	49	52	53	55	56	57	58	59
62	64	65	67						

As can be seen in Table 4, 32 selected features have higher importance than other features.

Table 5. Accuracy of learning algorithms on final data.

Algorithm	LSTM	GRU	RNN	NN	GB	RF	KNN	DT	LR
ACC	0/839	0/83	0/84	0/82	0/97	**0/988**	0/949	0/78	0/81

As can be seen in Table 5, the accuracy of random forest algorithm was higher than other methods. In addition, F1-Score, accuracy and readability of algorithms are shown in Table 6.

Table 6. Arithmetic mean F1-Score, readability and accuracy with final data.

Algorithm	LSTM	GRU	RNN	NN	GB	RF	KNN	DT	LR
F1-Score	0/28	0/34	0/28	0/32	0/92	**0/92**	0/86	0/61	0/49
Readability	0/3	0/36	0/41	0/29	0/89	**0/98**	0/89	0/49	0/52
Accuracy	0/28	0/39	0/24	0/23	0/94	**0/96**	0/9	0/67	0/26

As can be seen in Table 7, the F1-Score, accuracy and readability values in the random forest algorithm were higher than other methods.

Conclusion

In this research, the intrusion detection method was proposed by using the combination of thermal exchange optimization and seagull optimization algorithm for feature selection. Furthermore, random forest algorithm was used for attack classification.

This method was compared with several machine learning and deep learning methods and the results demonstrated that it has higher accuracy, precision, readability and F1-Score criteria than other algorithms. Moreover, the proposed method has a higher speed than other algorithms investigated in this research. To compare with previous research, some similar articles were selected and compared with the proposed method. The comparison showed that the results of this research were higher than the selected studies.

کارافن
فصلنامه علمی دانشگاه فنی و حرفه‌ای
پاییز ۱۴۰۲، دوره ۲۰، شماره ۳، ۵۲۹-۵۰۹
آدرس نشریه: https://karafan.tvu.ac.ir/
doi:10.48301/KSSA.2023.389398.2481

تشخیص نفوذ در شبکه با استفاده از الگوریتم بهینه‌سازی ترکیبی تبادل حرارتی و مرغ دریایی

منا عمادی[1*]، محمود نیائی[2]

۱- استادیار، گروه مهندسی کامپیوتر، دانشگاه پیام نور، تهران، ایران.
۲- دکترای مدیریت فناوری اطلاعات، سازمان محیط زیست، تهران، ایران.

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

دریافت مقاله: ۱۴۰۲/۰۱/۱۲
بازنگری مقاله: ۱۴۰۲/۰۴/۱۰
پذیرش مقاله: ۱۴۰۲/۰۶/۱۴

کلید واژگان:
تشخیص نفوذ
الگوریتم مرغ دریایی
الگوریتم بهینه‌سازی تبادل حرارتی
جنگل تصادفی
CICIDS2017

***نویسنده مسئول:** منا عمادی
پست الکترونیکی:
emadi.mona@pnu.ac.ir

چکیده

رشد روزافزون رایانه و اینترنت، بستر جدید و پرکاربردی برای ارائه خدمات شبکه‌ای فراهم نموده است. این امر میزان ارائه خدمات اداری، اجتماعی، مالی، آموزشی و تفریحی را بر روی شبکه و به ویژه اینترنت به طور چشمگیری افزایش داده است. گسترش استفاده از کاربردهای اینترنت، فرصتی برای سوءاستفاده از شبکه و اطلاعات آن برای مقاصد مجرمانه به وجود می‌آورد. براین‌اساس نفوذ در شبکه و دسترسی غیرمجاز به اطلاعات، به یکی از اصلی‌ترین نگرانی‌های کاربران شبکه‌ها و همچنین مدیران شبکه تبدیل شده است. سیستم‌های تشخیص نفوذ شامل مجموعه‌ای از ابزارها و سازوکارها برای نظارت بر سیستم‌های رایانه‌ای و ترافیک شبکه می‌باشد. روش‌های مختلفی برای تشخیص نفوذ مورداستفاده قرار می‌گیرد، مانند تکنیک‌های آماری، روش‌های مبتنی بردانش و همچنین روش‌های یادگیری ماشین. در این مقاله، یک روش برای تشخیص نفوذ با استفاده از الگوریتم‌های یادگیری ماشین بررسی و پیشنهاد شده است. مدل پیشنهادی، یک روش چند کلاسه می‌باشد که علاوه بر تشخیص نفوذ، نوع حمله را نیز مشخص می‌نماید. این روش یک مدل ترکیبی بوده که در آن از ترکیب الگوریتم‌های بهینه‌سازی مرغ دریایی و تبادل حرارتی و الگوریتم جنگل تصادفی استفاده شده است. به‌منظور تحلیل در این پژوهش، مجموعه‌داده CICIDS-۲۰۱۷ به‌کاررفته است. روش پیشنهادی با چندین الگوریتم مختلف مقایسه شده و مقدار دقت در روش پیشنهادی برابر با ۹۸/۸ به‌دست‌آمده که نسبت به بسیاری از روش‌های یادگیری ماشین دارای مقدار بالاتری می‌باشد.

شاپای الکترونیکی: ۲۵۳۸-۴۴۳۰
شاپای چاپی: ۲۳۸۲-۹۷۹۶

مقدمه

امروزه فناوری شبکه نقش عمده‌ای در ارتباطات و انتقال اطلاعات ایفا می‌کند به‌طوری که اکثر افراد با اعتماد به آن، به انجام فعالیت‌های فردی، اجتماعی و اقتصادی خود می‌پردازند و روزبه‌روز کاربردهای این فناوری افزایش می‌یابد. اما متأسفانه به همین نسبت، تعداد و نوع حملات و نفوذها در شبکه نیز در حال افزایش می‌باشد. ازاین‌رو دارابودن یک کانال امن برای انتقال اطلاعات، به یکی از نیازهای اجتناب‌ناپذیر برای سازمان‌ها و مدیران شبکه‌های کامپیوتری مبدل شده است. معمولاً برای ایجاد امنیت در یک شبکهٔ کامپیوتری، از دو سطح امنیتی استفاده خواهد شد. سطح اول استفاده از روش‌هایی مانند کنترل دسترسی، مکانیسم‌های احراز هویت و رمزنگاری و در سطح دوم، استفاده از سیستم‌های تشخیص نفوذ، دیوارهٔ آتش و آنتی‌ویروس‌ها است [۱].

روش‌های مختلفی به‌منظور تشخیص نفوذ در شبکه توسط محققان پیشنهاد شده است با این وجود به‌نظر می‌رسد این مسیر به تحقیقات بیشتری نیازمند باشد تا دقت و سرعت تشخیص افزایش یابد. در این پژوهش یک روش تشخیص نفوذ در شبکه معرفی شده است. در روش پیشنهاد شده، از ترکیب دو الگوریتم بهینه‌سازی مرغ دریایی [۲] و تبادل حرارتی [۳] برای انتخاب ویژگی و کاهش ابعاد مجموعه‌داده استفاده شده است. همچنین، عملیات دسته‌بندی حملات در روش پیشنهادی توسط الگوریتم جنگل تصادفی انجام می‌شود.

انتخاب ترکیب الگوریتم‌های فوق، پس از اعمال روش‌های مختلف و آزمون آنها، به دلیل عملکرد مناسب و میزان صحت، دقت، بازخوانی و مقدار F1-Score بالا و زمان آموزش پایین بوده است. در این تحقیق از مجموعه‌داده CICIDS۲۰۱۷ که یک مجموعه‌داده جدید به شمار می‌آید استفاده شده است. بدین ترتیب از نقاط قوت روش پیشنهاد شده در این تحقیق می‌توان به موارد زیر اشاره کرد:

- استفاده از یک روش ترکیبی جدید انتخاب ویژگی در تشخیص نفوذ
- چند کلاسه بودن روش پیشنهادی
- استفاده از مجموعه‌داده جدید CICIDS۲۰۱۷

در بخش دوم این مقاله پژوهش‌های پیشین ارائه می‌شوند و در بخش سوم مجموعه‌داده استفاده شده و روش پیشنهادی معرفی شده‌اند. بخش چهارم شامل نتایج پژوهش بوده و جمع‌بندی روش پیشنهادی در بخش پنجم ارائه شده است.

پژوهش‌های پیشین

سیستم‌های تشخیص نفوذ از نظر نوع تشخیص به سه گروه مبتنی بر امضا، مبتنی بر ناهنجاری و ترکیبی تقسیم می‌شوند و نظر نوع فناوری تشخیص، از طریق روش‌های آماری، روش‌های مبتنی بر دانش و روش‌های یادگیری ماشین قابل‌اجرا می‌باشند. محققان بسیار زیادی بر روی شناسایی تهدیدات، آسیب‌پذیری‌ها و حملات کار کرده‌اند.

در پژوهش [۴] که در سال ۲۰۱۹ به چاپ رسیده، استفاده از تشخیص ویژگی‌ها را برای کشف نفوذ مؤثر دانسته شده است. بدین منظور با استفاده از روش فیلترکردن FGLCC ویژگی‌ها را تشخیص داده و آنها را خوشه‌بندی می‌کند و در انتها با استفاده از روش درخت تصمیم‌گیری در ارتباط با داده‌های عادی و یا نفوذ تصمیم‌گیری به عمل می‌آید. در این روش از مجموعه‌داده KDD استفاده‌شده است. دقت نهایی در این روش ۹۵/۰۳ درصد است.

در سال ۲۰۲۰ در مقاله [۵] روش یادگیری گسترده BLS را برای کشف نفوذ به‌کار برده شده است. سیستم یادگیری گسترده BLS از قدرت یادگیری افزایشی استفاده می‌کند. یعنی بدون انباشته‌شدن ساختار لایه، شبکه‌های عصبی طراحی‌شده گره‌های عصبی را به طور گسترده گسترش می‌دهند و وزن‌های شبکهٔ عصبی را زمانی که به گره‌های اضافی نیاز است و زمانی که داده‌های ورودی به طور مداوم وارد شبکه‌های عصبی می‌شوند، به‌صورت تدریجی به‌روزرسانی

می‌کنند. ساختار شبکه طراحی شده و الگوریتم یادگیری افزایشی برای مدل‌سازی و یادگیری محیط کلان‌داده کاملاً مناسب است. در این پژوهش که از مجموعه‌داده CICIDS-۲۰۱۷ و چند مجموعه‌داده دیگر استفاده شده است.

در پژوهشی دیگر در سال ۲۰۲۰ یک سیستم تشخیص نفوذ به نام Siam-IDS به‌کاررفته است. روش شبکهٔ عصبی سیامی برای رفع مشکل عدم تعادل کلاس در مجموعه‌داده NSL-KDD پیشنهاد شده است. لازم به ذکر می‌باشد که این پژوهش نتوانسته است میزان دقت را بهبود ببخشد [۶].

در منبع دیگری [۷] محقق از روش شبکه باور عمیق و SVM استفاده نموده و دقت به‌دست‌آمده در آن ۹۶/۴۹ می‌باشد. در این پژوهش از مجموعه‌داده CICIDS۲۰۱۷ برای تحلیل استفاده شده است. همچنین در سال ۲۰۲۰ در مقاله [۸] از روشGRU برای تشخیص نفوذ استفاده شده است. در این تحقیق از مجموعه‌داده CICIDS۲۰۱۷ برای تحلیل استفاده شده و معیارهای دقت، بازخوانی و F1-Score آن محاسبه شده است.

در پژوهش [۹] که در سال ۲۰۲۱ انجام شده، پژوهشگر از روش یادگیری CNN و DNN روی مجموعه‌داده NSL-KDD به‌صورت چند کلاس استفاده کرده است. نتیجهٔ پژوهش با پنج کلاس بر روی مجموعه‌داده مورد اشاره، حداکثر صحت پژوهش ۷۷/۶۹، دقت ۷۳/۹۲، بازخوانی ۵۲/۸۶ و F1-Score برابر با ۵۳/۵۲ برای CNN بوده است.

در پژوهشی که در سال ۲۰۲۱ در مجله کارافن چاپ شد برای انتخاب ویژگی‌ها از الگوریتم ماشین بردار پشتیبان استفاده شد و تأثیر استفاده از الگوریتم‌های یادگیری ماشین را در میزان تشخیص نفوذ در سیستم مورد بررسی قرار داده است [۱۰]. یک پژوهش دیگر در ۲۰۲۱ [۱۱]، امکان نفوذ در شبکه را با شبکهٔ عصبی عمیق مورد بررسی قرار داده و روش شبکهٔ عصبی عمیق مبتنی بر پشته را پیشنهاد نموده است. محقق برای تحلیل این روش، از مجموعه‌داده CICIDS۲۰۱۷ استفاده نموده و ادعا کرده که توانسته دقت و صحت سیستم را تا اندازه‌ای ارتقا دهد. مقدار دقت در این پژوهش ۸۹/۹۷ گزارش شده است.

در منبع [۱۲] پژوهشگر یک سیستم تشخیص نفوذ با روش یادگیری شبکهٔ عصبی عمیق پیشنهاد داده است. این پژوهش که از چند مجموعه‌داده به همراه مجموعه‌داده CICIDS-۲۰۱۷ استفاده شده بهترین نتایج در روش DNN با ۵ لایه به دست آمده است. یادگیری عمیق زیرمجموعه‌ای از یادگیری ماشین است که مبتنی بر شبکه‌های عصبی مصنوعی بوده و به‌عنوان شبکه‌های عصبی عمیق (DNN) هم شناخته می‌شود. نتیجهٔ این پژوهش صحت ۹۶/۲ اعلام شده که نسبت به صحت روش پیشنهادی، مقدار پایین‌تری می‌باشد.

روش پیشنهادی

سیستم‌های تشخیص نفوذ که با رویکردهای معمول یادگیری ماشین ارائه می‌شوند، معمولاً با مشکلاتی مواجه می‌باشند، مانند مشکلات مهندسی ویژگی که توسط افراد خبره انجام می‌شود و منجر به حذف اطلاعات قابل‌توجهی از مجموعه‌داده‌ها می‌شود، یا مشکل زمان برای مواردی که مشخصات داده‌های ترافیک شبکه بسیار پیچیده و تعداد نمونه‌ها بسیار زیاد باشد.

برای کاهش این مشکلات، روش ارائه شده در این پژوهش از یک الگوریتم فرابتکاری[1] جدید که ترکیبی از دو الگوریتم بهینه‌سازی مرغ دریایی[2](SOA) و الگوریتم بهینه‌سازی تبادل حرارتی[3] (TEO) می‌باشد برای انتخاب ویژگی استفاده کرده است و همچنین الگوریتم جنگل تصادفی را برای دسته‌بندی حملات به کار گرفته است، تا دقت تشخیص و سرعت آن را افزایش دهد. نتایج نشان می‌دهند که الگوریتم بهینه‌سازی ترکیبی ارائه شده نسبت به سایر روش‌ها، از نظر دستیابی به جستجوی سراسری و سرعت هم‌گرایی برتری دارد.

[1] Meta-heuristic Algorithms
[2] Seagull optimization algorithm (SOA)
[3] Thermal exchange optimization (TEO)

در ادامه این بخش، ابتدا مجموعه‌داده مورداستفاده معرفی شده و سپس به تشریح الگوریتم‌های بهینه‌سازی مرغ دریایی و تبادل حرارتی و جنگل تصادفی پرداخته خواهد شد و در نهایت روش پیشنهادی معرفی می‌شود.

معرفی مجموعه‌های داده

معمولاً در پژوهش‌هایی که مربوط به تشخیص و کشف نفوذ در شبکه‌ها می‌باشند، از داده هایی مانند مجموعه‌داده Kyoto (۲۰۰۶)، Sperotto (۲۰۰۸)، ISCX (۲۰۱۳)ADFA (۲۰۱۲) و KDD-۹۹ استفاده می‌شود. اما در این پژوهش، ازمجموعه‌داده CICIDS-۲۰۱۷[1] به‌عنوان یکی از جدیدترین مجموعه‌ها استفاده شده است. مجموعه‌داده CICIDS- ۲۰۱۷ به‌منظور طراحی و ساخت IPS و IDS و برای حل مشکل کمبود مجموعه‌داده به‌روز و معتبر برای تشخیص نفوذ توسط موسسهٔ امنیت سایبری کانادا طراحی شده است [۱۳]. این مجموعه‌داده حاوی داده‌های اصلی ترافیک شبکه می‌باشد که با CICFlowMeter گرفته شده و شامل ۲۸۳۰۷۴۳ رکورد با برچسب ۱۴ نوع حمله و ۷۸ ویژگی می‌باشد. حملات مورد بررسی در مجموعه‌داده CICIDS۲۰۱۷ به‌صورت زیر می‌باشند [۱۴]:

- **جستجوی پورت**[2]: این نوع حملات در اولین مرحله یک حمله صورت می‌پذیرد و معمولاً شامل بررسی پورت‌های باز و یافتن نقاط ضعف سیستم‌ها استفاده می‌شود.
- **منع سرویس**[3] : برای خارج‌کردن ماشین و منابع شبکه از دسترس کاربران می‌باشد.
- **تزریق SQL**[4]: مهاجم با استفاده از دستورات SQL، عملیات آسیب را انجام می‌دهد.
- **حملات HULK**[5]: این حمله حجم بسیار زیادی از ترافیک منحصربه‌فرد و مبهم را در یک سرور وب ایجاد می‌کند.
- **حملات GE**[6]: نوعی از حملات منع سرویس که با درخواست مداوم آدرس‌های وب و نگه‌داشتن ارتباط، برای سرریز کردن منابع سرورهای وب استفاده می‌کند.
- **نفوذ از داخل شبکه**[7]: در این حملات، از یک کاربر آسیب‌پذیر در داخل شبکه استفاده می‌شود و پس از نفوذ به سیستم کاربر، اجرای حملات دیگر در شبکه امکان‌پذیر می‌گردد.
- **بات نت**[8]: شبکه‌ای از چند کامپیوتر که مخفیانه و بدون اطلاع کاربران واقعی، برای انجام فعالیت‌های مخرب تحت کنترل قرار می‌گیرند.
- **حملات جستجوی جامع**[9]: حملاتی برای به‌دست‌آوردن رمزهای عبور سرورها استفاده می‌شود. در این روش، هر ترکیب احتمالی از اعداد، حروف و کاراکترها مورد بررسی قرار می‌گیرد.
- **FTP-Patator**: نوعی حمله جستجوی جامع که بر روی سرویس FTP انجام می‌شود.
- **SSH-Patator**: از انواع حملات جستجوی جامع که بر روی سرویس SSH انجام می‌شود. SSH یک پروتکل برای برقراری اتصال امن میان کاربر و سرور می‌باشد.

[1] Canadian Institute of Cybersecurity Intrusion Detection System
[2] Port scan
[3] DOS
[4] SQL injection
[5] HTTP Unbearable Load King
[6] GoldenEye
[7] Infiltration of the network from inside
[8] Botnet
[9] Brute force

- **حملات SL**[1]: در این حمله سعی می‌شود بعد از ایجاد یک اتصال موفق با سرور، تاحدامکان آن اتصال حفظ شود.
- **حملات آزمون Slow HTTP**: نوعی از حملات DOS که از روش‌های کار پروتکل HTTP بهره‌برداری می‌شود [۱۳].

الگوریتم جنگل تصادفی

الگوریتم جنگل تصادفی اولین‌بار در سال ۲۰۰۱ توسط بریمن معرفی گردید. این الگوریتم یک روش یادگیری تحت نظارت است که برای دسته‌بندی و رگرسیون مورداستفاده قرار می‌گیرد. این الگوریتم بر اساس ساختاری متشکل از تعدادی درخت تصمیم کار می‌کند، الگوریتم درخت تصمیم یکی از الگوریتم‌های طبقه‌بندی و از رایج‌ترین روش‌های داده‌کاوی است که سادگی و کارآمدی آن باعث شده تا علی‌رغم مشکلاتی که در اجرای الگوریتم با متغیرهای دارای نویز یا صفات فاقد مقدار وجود دارد، به شکل گسترده‌ای در کاربردهای مختلف و مسائل مربوط به یادگیری ماشین استفاده شود [۱۵].

در درخت تصمیم با دنبال‌کردن مجموعه‌ای از سؤالات مرتبط با خصوصیات داده‌ها و نگاه به داده جاری برای اتخاذ تصمیم، طبقه یا دسته آن تعیین می‌شود. CART یک درخت باینری در الگوریتم درخت تصمیم است و جنگل تصادفی مجموعه‌ای از درختان CART است که در چهار مرحله بیان می‌شود:

۱- تعداد K زیرمجموعه از نمونه‌های آموزش ($D_1,D_2,\ldots,D_K$) در میان مجموعهٔ کل نمونه‌های موجود در بخش آموزش (D) توسط روش نمونه‌برداری Bootstrap انتخاب می‌شوند. درنهایت K درخت تصمیم‌گیر ایجاد می‌شود.

۲- در N شاخص گره درخت طبقه‌بندی، m مشخصه به طور تصادفی انتخاب می‌شود و مطابق با اصل حداقل خلوص گره، بهترین مشخصه در بین M شاخص کاندید انتخاب خواهد شد. به‌این‌ترتیب درختان رشد خواهند کرد.

۳- این مرحله تکرار گام دوم است. K درخت تصمیم‌گیر تولید می‌شود.

۴- تعداد K درخت تصمیم‌گیر که به‌خوبی رشد پیدا کرده‌اند جنگل تصادفی طبقه‌بند ترکیبی را تشکیل می‌دهند. نمونه واقع در طبقهٔ نهایی جنگل تصادفی منتظر رأی اکثریت می‌ماند [۱۶].

از مزایای الگوریتم جنگل تصادفی می‌توان به پاسخگویی بهتر در داده‌های با حجم بالا، پایداری الگوریتم، قابل‌استفاده برای رگرسیون و دسته‌بندی، محدود بودن تعداد پارامترها و آسان بودن استفاده از آن نام برد.

انتخاب ویژگی با استفاده از الگوریتم‌های فراابتکاری

یک روش برای حل مسائل بهینه‌سازی بررسی تمامی جواب‌های امکان‌پذیر و سپس محاسبهٔ توابع هدف مربوط به آنها می‌باشد تا در نهایت بهترین جواب انتخاب گردد. روش شمارش کامل جواب‌ها اگر چه به جواب دقیق مسئله منتهی می‌شود؛ اما به دلیل زیاد بودن تعداد جواب‌های امکان‌پذیر در عمل غیرممکن است. باتوجه‌به مشکل شمارش کامل جواب‌ها استفاده از روش‌هایی مؤثرتر مانند روش‌های ابتکاری و فرا ابتکاری موردتوجه قرار گرفته است. این روش‌ها می‌توانند جوابی نزدیک به بهینه را در زمانی محدود برای مسئله ارائه دهند [۱۷].

انتخاب ویژگی یک مسئله بهینه‌سازی باینری می‌باشد که در آن راه‌حل‌ها به مقادیر باینری محدود می‌شوند. انتخاب ویژگی را می‌توان به‌عنوان یک مسئله بهینه‌سازی چندهدفه در نظر گرفت که در آن باید دو هدف متناقض با حداقل

[1] Slow-Loris

تعداد ویژگی و حداکثر دقت طبقه‌بندی به دست آید [۱۸]. هر چه تعداد ویژگی‌های راه‌حل کمتر باشد و دقت طبقه‌بندی بالاتر باشد راه‌حل بهتر می‌باشد. هر راه‌حل برطبق تابع برازندگی پیشنهادی ارزیابی می‌شود، که بستگی به طبقه‌بند KNN برای به دست آوردن دقت طبقه‌بندی راه‌حل و تعداد ویژگی‌های انتخاب شده در آن دارد. در روش پیشنهادی جمعیتی از ذرات در ابعاد N در فضای ویژگی ایجاد شده و سپس پارامترهای الگوریتم بهینه‌سازی تنظیم شد. سپس تابع برازندگی با استفاده از K-NN ارزیابی شد و بهترین برازندگی به‌روز رسانی شده است. باتوجه‌به منبع [۱۹] تابع برازش K-NN انتخاب شده است. در نهایت در صورت برآورده‌شدن معیار پایان فرایند انتخاب پایان می‌پذیرد و زیرمجموعه‌ای از ویژگی‌ها انتخاب می‌گردد. به‌منظور تعادل بین تعداد ویژگی‌های انتخاب شده در هر راه‌حل (حداقل) و دقت طبقه‌بندی (حداکثر) از تابع برازندگی در فرمول ۱ استفاده شده است.

$$Fitness = \alpha\gamma r(D) + \beta\frac{|R|}{|N|} \quad (۱)$$

به‌طوری‌که $\gamma r(D)$ نرخ خطای طبقه‌بندی را نشان می‌دهد که در اینجا طبقه‌بند استفاده شده KNN می‌باشد. علاوه‌بر این $|R|$، تعدا عناصر زیرمجموعه انتخاب شده و $|N|$ تعداد کل ویژگی‌های مجموعه‌داده می‌باشد. α و β دو پارامتر مرتبط با اهمیت کیفیت طبقه‌بندی و طول زیرمجموعه هستند. $\beta = (1-\alpha)$ و $\alpha \in [0,1]$ که از [۱۷] اتخاذ شده‌اند.

الگوریتم مرغ دریایی

مرغ‌های دریایی پرندگانی هستند که از نظر وزن و قد انواع مختلفی دارند و از حشرات، ماهی‌ها، خزندگان، دوزیستان و کرم‌های خاکی تغذیه می‌کنند. این پرندگان با روش‌هایی هوشمندانه طعمه را جذب می‌کنند. مدل‌های ریاضی مهاجرت و حملهٔ این شکارچیان مورد بحث قرار گرفته و شبیه‌سازی‌شده است. یک مرغ دریایی باید شرایط زیر را در حین مهاجرت رعایت کند:

برای جلوگیری از برخورد بین عوامل جستجوی مجاور از متغیر اضافه‌ای مانند A استفاده شده است تا موقعیت جدید عامل جستجو محاسبه شود.

$$C_s = A \times P_s \quad (۲)$$

C_s بیانگر موقعیت عامل جستجویی است که با عامل‌های دیگر برخوردی ندارد. P_s نشان‌دهندهٔ موقعیت عامل جستجو می‌باشد و X اندیس تکرار فعلی را نشان می‌دهد و A رفتار عامل جستجو در فضای جستجوی معین را نشان می‌دهد.

$$A = f_c - (x \times (\frac{f_c}{Max_{iteration}})) \quad (۳)$$

f_c برای کنترل تکرار متغیر A که به‌صورت خطی از f_c تا صفر کاهش می‌یابد استفاد شده است. پس از جلوگیری از برخورد بین همسایگان، عامل‌های جستجو به سمت بهترین همسایه حرکت می‌کنند.

$$M_s = B \times (P_{bs}(x) - P_s(x)) \quad (۴)$$

M_S نشان‌دهنده موقعیت های عامل جستجو P_S به سمت بهترین عامل جستجوی مناسب P_{bs} است. رفتار B به‌صورت تصادفی است که باعث ایجاد تعادل بین اکتشاف و بهره‌برداری مناسب است. B به‌صورت زیر محاسبه می‌شود:

$$B = 2 \times A^2 \times rd \quad (۵)$$

به‌طوری‌که rd یک عدد تصادفی در بازه $[۰,۱]$ می‌باشد. در نهایت عامل جستجو می‌تواند موقعیت خود را نسبت به بهترین عامل جستجو به‌صورت زیر به‌روزرسانی کند:

$$D_s = |C_s + M_s| \quad (۶)$$

D_S بیانگر فاصله بین عامل جستجو و عامل جستجو مناسب است. در زمان حمله، عمل مارپیچی در هوا انجام می‌شود. این رفتار در صفحات x ، y و z به‌صورت زیر توصیف شده است:

$$x' = r \times cos(k) \quad (۷)$$

$$y' = r \times sin(k) \quad (۸)$$

$$z' = r \times k \quad (۹)$$

$$r = u \times e^{kv} \quad (۱۰)$$

در فرمول‌های بالا r شعاع هر چرخش مارپیچ است،k عددی تصادفی در بازه $0 \leq k \leq 2\pi$ و $u.$ و v ثابت‌هایی برای تعریف شکل مارپیچی هستند و e پایه لگاریتم طبیعی است. موقعیت به‌روز شده عامل جستجو با استفاده از فرمول‌های ۶ تا ۹ محاسبه می‌شود.

$$P_s(x) = (D_s \times x' \times y' \times z') + P_{bs}(x) \quad (۱۱)$$

در فرمول بالا P_S بهترین راه‌حل را ذخیره می‌کند و موقعیت عامل‌های دیگر جستجو را به‌روز رسانی می‌کند [۲].

الگوریتم بهینه‌سازی تبادل حرارتی

الگوریتم TEO یک الگوریتم بهینه‌سازی جدید بر اساس قانون سرمایش نیوتن است که در آن نرخ اتلاف گرمای یک جسم متناسب با تفاوت دما بین جسم و محیط اطراف آن است. در این الگوریتم برخی از عوامل به‌عنوان اشیا خنک‌کننده تعریف می‌شوند و عوامل دیگر قرار است نمایانگر محیط باشند. به‌روزرسانی فرمول دما بین اشیا به‌صورت زیر تعریف می‌شود [۳]:

$$T^{env_i} = (1 - (c_1 + c_2 \times (1 - t)) \times rand) \times T'^{env_i} \quad (۱۲)$$

$$t = \frac{l}{L} \quad (۱۳)$$

که در آن c_1 و c_2 متغیرهای کنترل، T'^{env}_i دمای قبلی جسم هستند که به T^{env}_i تغییر داده می‌شود. l عدد تکرار فعلی و L حداکثر تعداد تکرار است. دمای جدید، مطابق با مراحل قبلی و رابطهٔ زیر به‌روز می‌شود:

$$T_i^{new} = T_i^{env} + (T_i^{old} - T_i^{env})\, exp(-\beta t) \quad (۱۴)$$

که در آن یک شی β را پایین می‌آورد و دما را اندکی جابه‌جا می‌کند. مقدار β برای هر شی با رابطهٔ (۱۵) محاسبه می‌شود.

$$\beta = Cost\,(object)/\,Cost\,(worst - object) \quad (۱۵)$$

هزینهٔ (شئ) ارزش فعلی شئ هدف است و هزینهٔ (بدترین شئ) مقدار ارزش شئ هدف است. برای جلوگیری از گیرافتادن دمای جسم در بهینهٔ محلی، پارامتر Pro تنظیم می‌شود. پارامتری که تعیین می‌کند آیا باید یک مؤلفه از هر شئ خنک‌کننده تغییر کند یا خیر. اگر مقدار Pro از *rand* بیشتر باشد، یک بُعد از عامل i ام به طور تصادفی انتخاب شده و مقدار آن به شکل زیر به‌روز می‌شود:

$$T_{i,j} = Tj, minj, max_{i,min} \quad (۱۶)$$

که در آن $T_{i,j}$ متغیر j امین عامل i ام است. $T_{j,}max$, min و T_j محدوده‌های پایین و بالای متغیر j ام هستند. با الهام از این فرایند، یک‌روال بهینه‌سازی قابل طراحی است [۳].

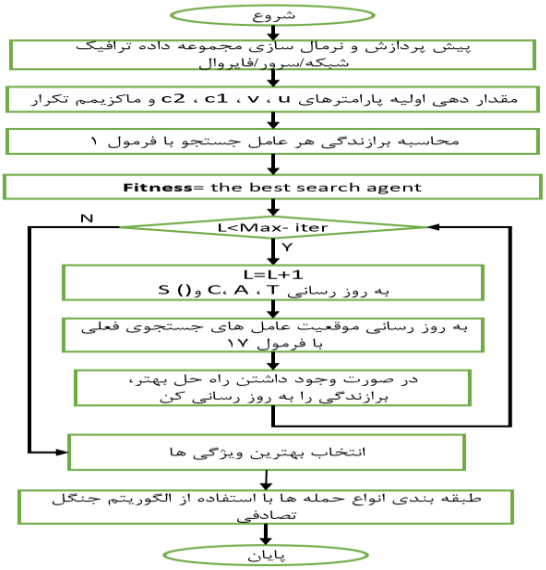
مدل پیشنهادی

در روش پیشنهادی برای انتخاب ویژگی‌ها از الگوریتم بهینه‌سازی که ترکیبی از الگوریتم‌های بهینه‌سازی مرغ دریایی و تبادل حرارتی است بهره گرفته است. الگوریتم بهینه‌سازی ترکیبی بر روی مجموعه‌داده اعمال و بهترین ویژگی‌های مجموعه‌داده انتخاب شده‌اند. الگوریتم SOA توانایی جستجوی سراسری خوبی دارد، درحالی‌که الگوریتم TEO توانایی جستجوی محلی قوی‌ای دارد. به‌منظور بهبود توانایی جستجوی هر دو الگوریتم از الگوریتم بهینه‌سازی ترکیبی جهت انتخاب ویژگی‌ها استفاده شده است.

بیان در الگوریتم TEO با فرمول حملهٔ مرغ دریایی الگوریتم SOA بهبود می‌یابد، تا در نهایت توانایی جستجوی محلی الگوریتم مرغ دریایی را بهبود دهد. از ایدهٔ تبادل حرارتی در الگوریتم TEO برای افزایش بهره‌وری مرغ دریایی استفاده می‌شود. در فرمول ۱۳، β ، دما را به کندی بین اشیا مبادله می‌کند تا به سرعت به شئ هدف نزدیک شود. بنابراین β در معادلهٔ ۳ بهبود یافته است تا مرغان دریایی بهتر بتوانند به سمت طعمه حرکت کنند. فرمول ریاضی آن به شرح زیر است:

$$M_s = B \times (P_{bs}(x) - P_s(x)) \times exp(-\beta t) \quad (۱۷)$$

پس از آن داده‌ها به دو دسته ویژگی‌ها (X) و برچسب‌ها (Y) تقسیم و تعداد ۱۵ درصد از سطرهای مجموعه‌داده به طور تصادفی و به صورتی که به‌تناسب، ترکیبی از تمام انواع برچسب‌ها انتخاب شده باشند، جدا و به‌صورت X_train، y_train، X_test و y_test نام‌گذاری شدند. در نهایت برای کلاس بندی حملات، الگوریتم جنگل تصادفی مورداستفاده قرار گرفت. فلوچارت مدل پیشنهادی در شکل ۱ نشان داده شده است.



شکل ۱. فلوچارت مدل پیشنهادی.

آزمایش‌ها و نتایج

در این بخش، روش پیشنهاد شده بر روی مجموعه‌داده CICIDS-۲۰۱۷ اعمال و به ارزیابی نتایج پرداخته خواهد شد. معیارهای ارزیابی مورداستفاده در این پژوهش، علاوه بر معیار صحت[۱]، معیارهای دقت[۲]، فراخوانی[۳]، F1-score و زمان می‌باشند [۲۰]. لازم به ذکر است معیارهای مزبور در برخی مقالات و مطالعات با عناوین دیگری مورداستفاده قرار می‌گیرند؛ لذا در جدول (۱) روابط و نام‌های متفاوت آن معیارها ذکر شده است.

جدول ۱. معیارهای ارزیابی و اسامی متفاوت آنها.

معیار ارزیابی	نام‌های دیگر مورداستفاده در مقالات	رابطه ریاضی
صحت (Accuracy)	دقت	$\frac{TP+TN}{TP+FP+TN+FN}$
دقت (Precision)	Positive Predictive Value, PPV, PR, Efficiency	$\frac{TP}{TP+FN}$
بازخوانی (Recall)	DR, True Positive, TP Rate, Sensitivity, Detection Rate, Effectiveness	$\frac{TP}{TP+FN}$
F1-Score	F-Value, F-Score, F-Measure, F1	$\frac{2.Re\,c\,all.precision}{Re\,c\,all+precision}$

[1] Accuracy
[2] Precision
[3] Recall

سیستم رایانه مورداستفاده در این پژوهش، دارای پردازنده Core i5 و مقدار ۸ گیگابایت RAM بوده و از Matlab2020b و پایتون در محیط Colab استفاده شده است.

پیش‌پردازش داده‌ها

به‌منظور پیش‌پردازش در مجموعه‌داده، ابتدا ویژگی‌های با مقدار صفر و داده‌های گم شده و همچنین داده‌هایی که مقداری برای آنها ثبت نشده بود حذف شدند. سپس حملات به‌صورت شش گروه حمله و یک گروه نرمال عددی شدند. در ادامه برای یکسان کردن مقیاس داده‌ها عملیات نرمال‌سازی بر روی داده‌ها انجام شد. این مرحله با روش کمترین - بیشترین انجام شد و تمامی داده‌ها با استفاده از رابطهٔ (۱۸) به بازهٔ [۱ و۰] منتقل شدند.

$$x_i = \frac{(x - MIN(X))}{(MAX(X) - MIN(X))} \qquad (۱۸)$$

که در آن x دادهٔ اولیه، MIN(X) و MAX(X) به ترتیب کمترین و بیشترین داده‌های مجموعه می‌باشند و xi عدد نهایی است که باید به‌جای دادهٔ اصلی در محاسبات یادگیری استفاده خواهد شد.

نتایج الگوریتم‌های مختلف بر روی مجموعه‌داده اولیه

پس از انجام تمام مراحل پیش‌پردازش داده‌ها، داده‌ها به دسته‌های ویژگی‌ها (X) و برچسب‌ها (Y) تقسیم شدند و سپس تعداد ۱۰ درصد از رکوردهای مجموعه‌داده به طور تصادفی و به شکلی که به‌تناسب، ترکیبی از تمام انواع برچسب‌ها انتخاب شده باشند، جدا شده و به‌صورت X_train، y_train، X_test و y_test نام‌گذاری شدند. سپس الگوریتم‌های یادگیری بر روی X_train و y_train اعمال شده و نتایج ACC برای الگوریتم‌های LR[1]، DT[2]، KNN[3]، RF[4]، GB[5]، NN[6]، RNN[7]، GRU[8] و LSTM[9] مطابق جدول (۲) ثبت شدند.

جدول ۲. مقادیر صحت (Accuracy) الگوریتم‌های یادگیری بر روی داده‌های اولیه.

الگوریتم	LR	DT	KNN	RF	GB	NN	RNN	GRU	LSTM
ACC	۰/۷۶۱	۰/۵۲۲	۰/۸۹	۰/۹۱۲	۰/۸۹۶	۰/۷۴۲	۰/۸۱۶	۰/۷۹	۰/۷۸۱

در جدول (۳) مقادیر زمان اجرای الگوریتم‌های مذکور (به ثانیه) نشان داده شده است.

[1] Logistic regression
[2] Decision Tree
[3] K-Nearest Neighbors
[4] Random forest
[5] Gradient boosting
[6] Neural Network
[7] Recurrent Neural Network
[8] Gated Recurrent Unit
[9] Long Short-Term Memory

جدول ۳. مدت‌زمان (ثانیه) اجرای الگوریتم‌ها بر روی داده‌های اولیه.

الگوریتم	LR	DT	KNN	RF	GB	NN	RNN	GRU	LST
زمان آموزش	۳۷۸	۵۹	۴۷۱	۱۹/۴	۱۶۲۳	۵۸۴	۱۸۰۸	۲۶۴۷	۳۰۰۲
زمان آزمایش	۰/۳	۰/۹	۱۸۳	۰/۱۷	۱۲	۱۸	۳۶	۳۹	۴۲

مطابق جدول (۳) زمان آموزش الگوریتم جنگل تصادفی ۱۹/۴ ثانیه بوده است که در شرایط کاملاً مشابه با سایر الگوریتم‌ها سریع‌ترین روش بوده است. لازم به ذکر است که در هر الگوریتم، برای رسیدن به بهترین مقدار، آرگومان‌ها بارها تکرار و بررسی شده و نتیجهٔ مشاهده شده، مربوط به بهترین آرگومان می‌باشد. در جدول (۴) میانگین حسابی نتایج معیارهای F1-Score، بازخوانی و دقت برای روش‌های بررسی شده، با داده‌های اولیه مجموعه‌داده مشاهده می‌شود.

جدول ۴. میانگین حسابی مقادیر F1-Score، بازخوانی و دقت (Precision) با داده‌های اولیه.

الگوریتم	LZ	DT	KNN	RF	GB	NN	RNN	GRU	LSTM
F1-Score	۰/۲	۰/۴۰۹	۰/۸	**۰/۷۹**	۰/۷۲	۰/۱	۰/۲	۰/۱۷	۰/۱۳
بازخوانی	۰/۲۱	۰/۴۲۹	۰/۸۰۶	**۰/۸۱۹**	۰/۷۸	۰/۱۳	۰/۱۹	۰/۱۶	۰/۱۱
دقت	**۰/۱۸**	۰/۵۴	۰/۸	**۰/۸۱**	۰/۶۹	۰/۱۱	۰/۱	۰/۳۷	۰/۱

اعمال انتخاب ویژگی بر روی مجموعه‌داده

مجموعه‌داده CICIDS-۲۰۱۷ پس از پیش‌پردازش، دارای ۷۰ ویژگی می‌باشد. عملیات انتخاب ویژگی با الگوریتم ترکیبی مرغ دریایی و بهینه‌سازی تبادل حرارتی، موجب انتخاب تعداد ۳۲ ویژگی شده است. ویژگی‌های انتخاب شده در جدول (۵) مشاهده می‌شوند.

جدول ۵. گزارش انتخاب ویژگی با الگوریتم ترکیبی بهینه‌سازی مرغ دریایی و الگوریتم تبادل حرارتی.

شماره ویژگی‌های انتخاب شده									
۱۲	۱	۲	۵	۶	۷	۱۲	۱۵	۱۸	۱۹
۲۴	۲۶	۲۷	۴۴	۳۰	۳۴	۳۷	۳۹	۴۱	۴۴
۴۷	۴۸	۴۹	۵۲	۵۳	۵۵	۵۶	۵۷	۵۸	۵۹
۶۲	۶۴	۶۵	۶۷						

چنان‌که در جدول (۵) مشاهده می‌شود تعداد ویژگی‌های انتخاب شده که دارای اهمیت بالاتری نسبت به بقیه ویژگی‌ها دارند تعداد ۳۲ ویژگی می‌باشد. در ادامه پژوهش الگوریتم‌های یادگیری بر روی این ویژگی‌ها اعمال گردید. نتیجهٔ این عملیات در جداول (۶) تا (۹) آمده است.

جدول ۶. مقادیر صحت (Accuracy) الگوریتم‌های یادگیری بر روی داده‌های نهایی.

الگوریتم	LR	DT	KNN	RF	GB	NN	RNN	GRU	LSTM
ACC	۰/۸۱	۰/۷۸	۰/۹۴۹	۰/۹۸۸	۰/۹۷	۰/۸۲	۰/۸۴	۰/۸۳	۰/۸۳۹

چنانچه در جدول (۶) ملاحظه می‌شود، صحت در الگوریتم جنگل تصادفی از بقیه روش‌ها بیشتر می‌باشد. جدول (۷) زمان اجرای یادگیری و آزمایش بر روی مجموعه‌داده نهایی را نشان می‌دهد.

جدول ۷. مدت‌زمان اجرای الگوریتم‌ها با مجموعه‌داده نهایی.

الگوریتم	LR	DT	KNN	RF	GB	NN	RNN	GRU	LSTM
زمان آموزش	۲۱۷	۵۹	۳۰۵	**۱۲/۶**	۹۶۱	۶۷۹	۸۲۴	۱۹۸۰	۲۵۳۲
زمان آزمایش	۰/۲	۰/۶	۹۶	**۰/۱**	۱۲	۹	۲۴	۲۴	۴۲

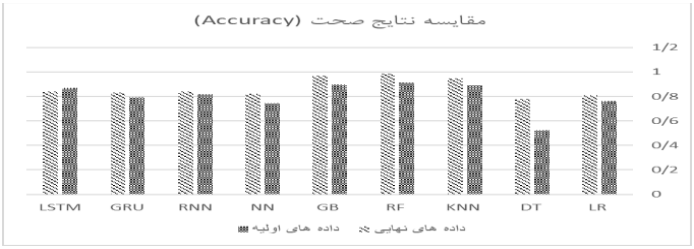
همچنین نتایج F1-Score، دقت و بازخوانی الگوریتم‌ها با داده‌های ذکر شده به‌صورت جدول (۸) بوده است.

جدول ۸. میانگین حسابی F1-Score ، بازخوانی و دقت (Precision) با داده‌های نهایی.

الگوریتم	LR	DT	KNN	RF	GB	NN	RNN	GRU	LSTM
F1-Score	۰/۴۹	۰/۶۱	۰/۸۶	**۰/۹۲**	۰/۹۲	۰/۳۲	۰/۲۸	۰/۳۴	۰/۲۸
بازخوانی	۰/۵۲	۰/۴۹	۰/۸۹	**۰/۹۸**	۰/۸۹	۰/۲۹	۰/۴۱	۰/۳۶	۰/۳
دقت	۰/۲۶	۰/۶۷	۰/۹	**۰/۹۶**	۰/۹۴	۰/۲۳	۰/۲۴	۰/۳۹	۰/۲۸

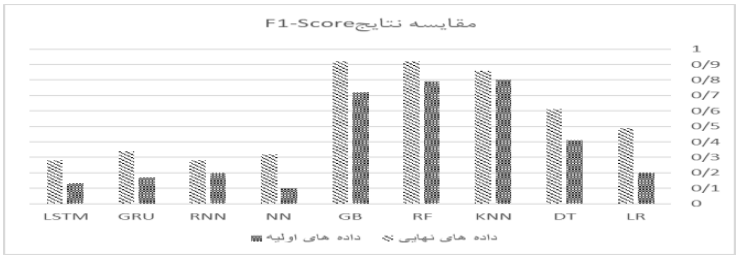
چنانچه در جدول (۸) ملاحظه می‌شود، مقادیر F1-Score، دقت و بازخوانی در الگوریتم جنگل تصادفی نسبت به سایر روش‌ها بالاتر می‌باشد.

مقایسه نتایج دقت مورد محاسبه در روش پیشنهادی با سایر روش‌های انجام شده بر روی داده‌های اولیه و داده‌های نهایی به‌صورت نمودار (۱) داده شده است.



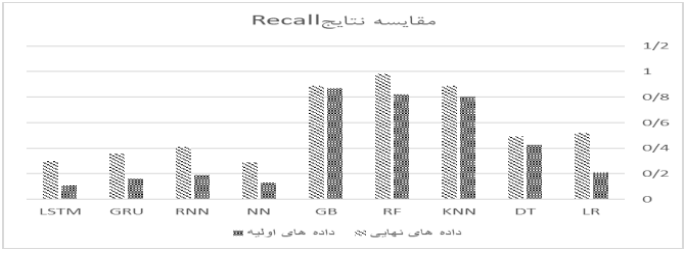
نمودار ۱. مقایسهٔ صحت (Accuracy) روش پیشنهادی با سایر روش‌ها بر روی داده‌های اولیه و داده‌های نهایی.

چنان‌که مشاهده می‌شود، مقدار صحت در روش پیشنهادی این پژوهش پس از انتخاب ویژگی بهبودیافته است و همچنین نسبت به سایر روش‌ها نیز بالاتر بوده است. نمودار (۲) مقایسهٔ نتایج F1-Score روش پیشنهادی با سایر روش‌های انجام شده بر روی داده‌های اولیه و داده‌های نهایی را نشان می‌دهد.



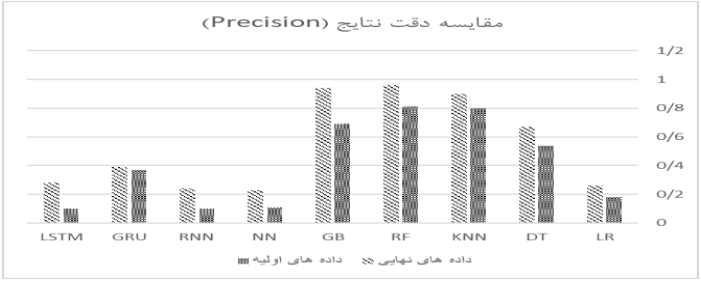
نمودار ۲. مقایسه F1-Score روش پیشنهادی با سایر روش‌ها بر روی داده‌های اولیه و داده‌های نهایی.

چنان‌که در نمودار (۲) دیده می‌شود، مقدار F1-Score نیز در روش پیشنهادی پس از انتخاب ویژگی افزایش یافته و نسبت به سایر روش‌ها بالاتر بوده است. نتایج بازخوانی روش پیشنهادی با سایر روش‌های انجام شده بر روی داده‌های اولیه و داده‌های نهایی به‌صورت نمودار (۳) داده شده است.



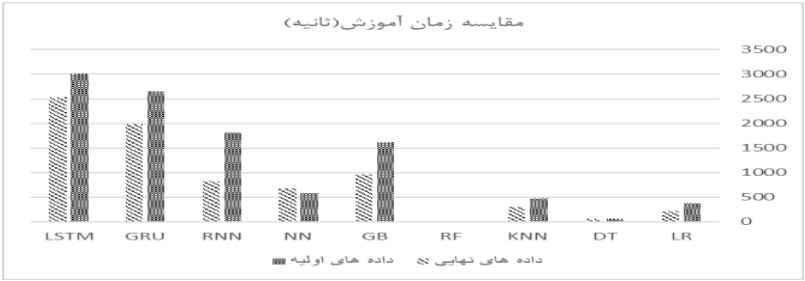
نمودار ۳. مقایسه بازخوانی روش پیشنهادی با سایر روش‌ها بر روی داده‌های اولیه و داده‌های نهایی.

نمودار (۳) نشان می‌دهد که مقدار بازخوانی در روش پیشنهادی با اعمال انتخاب ویژگی ارتقا یافته است و درضمن این روش نسبت به سایر روش‌ها نیز دارای مقدار بازخوانی بالاتری بوده است. نمودار (۴) نتایج دقت روش پیشنهادی با سایر روش‌های انجام شده را بر روی داده‌های اولیه و داده‌های نهایی مقایسه نموده است.



نمودار ۴. مقایسه دقت روش پیشنهادی با سایر روش‌ها بر روی داده‌های اولیه و داده‌های نهایی.

چنان‌که مشاهده می‌شود، مقدار دقت نیز در روش پیشنهادی این پژوهش پس از انتخاب ویژگی بهبود یافته است و چنانچه می‌توان دید، این مقدار نسبت به سایر روش‌ها بالاتر بوده است. نتایج مدت‌زمان اجرای آموزش در روش پیشنهادی با سایر روش‌های انجام شده بر روی داده‌های اولیه و داده‌های نهایی به‌صورت نمودار (۵) داده شده است.



نمودار ۵. مقایسهٔ زمان اجرای روش پیشنهادی با سایر روش‌ها بر روی داده‌های اولیه و داده‌های نهایی.

چنان‌که مشاهده می‌شود، زمان اجرای آموزش در روش پیشنهادی این پژوهش پس از انتخاب ویژگی کمتر شده است و این روش نسبت به سایر روش‌ها سریع‌تر بوده است.

مقایسه با پژوهش‌های مشابه

به‌منظور مقایسهٔ نتایج روش پیشنهادی با نتایج پژوهش‌های مشابه، برخی از این پژوهش‌ها انتخاب و نتایج کسب شده در جدول (۹) موردمطالعه قرار گرفته‌اند.

جدول ۹. مقایسهٔ نتایج پژوهش با پژوهش‌های پیشین

شماره ارجاع	روش	سال ارائه	مجموعه‌داده	معیار ارزیابی	مقدار (درصد)
[۵]	BLS	۲۰۲۰	CICIDS۲۰۱۷	صحت	۹۶/۶۳
[۷]	DBN SVM	۲۰۱۸	CICIDS۲۰۱۷	صحت	۹۶/۴۹
				دقت	۹۰/۴۰
				بازخوانی	۹۵/۶۵
				F1-Score	۹۲/۹۵
[۸]	GRU	۲۰۲۰	CICIDS۲۰۱۷	دقت	۸۲/۲۸
				بازخوانی	۸۵/۲۸
				F1-Score	۷۸/۱۴
[۱۲]	DNN	۲۰۱۹	CICIDS۲۰۱۷	صحت	۹۶/۲
روش پیشنهادی	TEO+RF	۲۰۲۳	CICIDS۲۰۱۷	صحت	۰/۹۸۸
				دقت	۰/۹۲
				بازخوانی	۰/۹۸
				F1-Score	۰/۹۶

همان‌گونه که مشاهده می‌شود در پژوهش [۵] روش یادگیری گسترده BLS و مجموعه‌داده CICIDS-۲۰۱۷ به‌کار برده شده است. پژوهشگر اعتقاد دارد این سیستم یک تغییر پارادایم کامل در یادگیری سریع و دقیق بدون ساختار عمیق است. در این پژوهش برای مجموعه‌داده CICIDS-۲۰۱۷ میزان صحت ۰/۹۶۶۳ اعلام شده است که در مقایسه با روش پیشنهادی، مقدار کمتری می‌باشد.

در مقالهٔ [۷] که از روش شبکه باور عمیق و SVM استفاده شده، مقدار دقت ۴۹ /۹۶ گزارش شده است که در مقایسه با مقدار صحت محاسبه شده در پژوهش حاضر مقدار کمتری می‌باشد.

روش GRU که در مقالهٔ [۸] پیشنهاد شده، معیارهای دقت ۸۲/۲۸، بازخوانی ۸۵/۲۸ و F۱-Score ۷۸/۱۴ محاسبه شده است. همان گونه که مشاهده می‌شود این نتایج نیز از مقدار به‌دست‌آمده از پژوهش حاضر پایین‌تر می‌باشند.

در روش یادگیری شبکهٔ عصبی عمیق که در پژوهش [۱۲] پیشنهاد شده نیز برای مجموعه‌داده CICIDS-۲۰۱۷ مقدار صحت ۹۶/۲ اعلام شده که نسبت به صحت روش پیشنهادی، مقدار پایین‌تری می‌باشد.

دلیل بهبود نتایج فوق، استفاده از روش انتخاب ویژگی ترکیبی مناسب بوده است. این روش که ترکیب الگوریتم مرغ دریایی (SOA) با توانایی جستجوی سراسری بالاتر با الگوریتم تبادل حرارتی (TEO) با توانایی جستجوی محلی بهتر، توانسته است متغیرهای با اهمیت و اولویت بالاتر را در زمان کمتری انتخاب نماید. در ادامه، پس از انتخاب مهم‌ترین ویژگی‌های مجموعه‌داده، توسط الگوریتم دسته‌بندی مناسب در خصوص نوع حمله آن تصمیم‌گیری شده است.

نتیجه‌گیری و پیشنهادها

در این پژوهش، یک روش تشخیص نفوذ با استفاده از الگوریتم بهینه‌سازی ترکیبی مرغ دریایی و تبادل حرارتی به‌عنوان انتخاب ویژگی و الگوریتم جنگل تصادفی برای دسته‌بندی ارائه شده است. با استفاده از ترکیب الگوریتم تبادل حرارتی با مرغ دریایی، توانایی بهینه‌سازی محلی الگوریتم مرغ دریایی بهبود یافت. این روش با چندین روش یادگیری ماشین و یادگیری عمیق مقایسه شده و نتایج نشان داده که معیارهای صحت، دقت، بازخوانی و F۱-Score بالاتری نسبت به سایر الگوریتم‌های اجرا شده داشته است. همچنین، سرعت اجرای روش پیشنهادی نیز نسبت به داده‌های اولیه و هم نسبت به سایر الگوریتم‌هایی که در این پژوهش اجرا شده‌اند بالاتر بوده است. به‌منظور مقایسه با پژوهش‌های پیشین، چند مقاله مشابه انتخاب و با روش پیشنهاد شده مقایسه گردید. مقایسه نشان می‌دهد که مقادیر نتایج این پژوهش بالاتر از مطالعات انتخاب شده می‌باشند.

بدیهی است باتوجه به متفاوت‌بودن سخت‌افزار و پهنای باند مورداستفاده در سایر پژوهش‌ها، مقایسهٔ زمان اجرای کدها قابل‌استناد نمی‌باشد، لیکن مقایسه زمان اجرا در الگوریتم‌های به‌کار رفته در این پژوهش که در نمودار (۵) قابل‌مشاهده است، حاکی از سرعت بالای روش پیشنهادی نسبت به الگوریتم‌های دیگر می‌باشد.

برای پژوهش‌های آینده، با توجه به‌ضرورت اعمال تغییرات آنی پس از تشخیص حملات در شبکه، پیشنهاد می‌شود سیستم توصیه‌گرِ مدیر شبکه، بر اساس خروجی روش پیشنهادی بررسی و مطالعه گردد.

References

[1] Kappagantu, M., Villamor, D. E. V., Bullock, J. M., & Eastwell, K. C. (2017). A rapid isothermal assay for the detection of Hop stunt viroid in hop plants (Humulus lupulus), and its application in disease surveys. *Journal of Virological Methods*, *245*, 81-85. https://doi.org/10.1016/j.jviromet.2017.04.002

[2] Dhiman, G., & Kumar, V. (2019). Seagull optimization algorithm: Theory and its applications for large-scale industrial engineering problems. *Knowledge-Based Systems*, *165*, 169-196. https://doi.org/10.1016/j.knosys.2018.11.024

[3] Kaveh, A., & Dadras, A. (2017). A novel meta-heuristic optimization algorithm: Thermal exchange optimization. *Advances in Engineering Software*, *110*, 69-84. https://doi.org/10.1016/j.advengsoft.2017.03.014

[4] Mohammadi, S., Mirvaziri, H., Ghazizadeh Ahsaee, M., & Karimipour, H. (2019). Cyber intrusion detection by combined feature selection algorithm. *Journal of Information Security and Applications*, *44*, 80-88. https://doi.org/10.1016/j.jisa.2018.11.007

[5] Rios, A. L. G., Li, Z., Bekshentayeva, K., & Trajković, L. (2020, October 12-14). *Detection of Denial of Service Attacks in Communication Networks.* 2020 Institute of Electrical and Electronics Engineers International Symposium on Circuits and Systems Seville, Spain. https://doi.org/10.1109/ISCAS45731.2020.9180445

[6] Bedi, P., Gupta, N., & Jindal, V. (2020). Siam-IDS: Handling class imbalance problem in Intrusion Detection Systems using Siamese Neural Network. *Procedia Computer Science*, *171*, 780-789. https://doi.org/10.1016/j.procs.2020.04.085

[7] Marir, N., Wang, H., Feng, G., Li, B., & Jia, M. (2018). Distributed Abnormal Behavior Detection Approach Based on Deep Belief Network and Ensemble SVM Using Spark. *Institute of Electrical and Electronics Engineers Access*, *6*, 59657-59671. https://doi.org/10.1109/ACCESS.2018.2875045

[8] Kurochkin, I., & Volkov, S. (2020, September 6-13). *Using GRU based deep neural network for intrusion detection in software-defined networks.* Institute of Physics Conference Series: Materials Science and Engineering 2020: XIII International Conference on Applied Mathematics and Mechanics in the Aerospace Industry, Alushta, Russia. https://doi.org/10.1088/1757-899X/927/1/012035

[9] Mulyanto, M., Faisal, M., Prakosa, S. W., & Leu, J-S. (2021). Effectiveness of Focal Loss for Minority Classification in Network Intrusion Detection Systems. *Symmetry*, *13*(1), 4. https://doi.org/10.3390/sym13010004

[10] Abdolhosseini, M., Abdollahi, R., & Rajaee, M. (2021). Designing of PIλDδ controller for PMBLDC motor using metaheuristic algorithms. *Karafan Quarterly Scientific Journal*, *17*(4), 149-165. https://doi.org/10.48301/kssa.2021.128401

[11] Tama, B. A., & Lim, S. (2021). A Stacking-Based Deep Neural Network Approach for Effective Network Anomaly Detection. *Computers Materials & Continua*, *66*(2), 2217 - 2227. https://doi.org/10.32604/cmc.2020.012432

[12] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. *Institute of Electrical and Electronics Engineers Access*, *7*, 41525-41550. https://doi.org/10.1109/ACCESS.2019.2895334

[13] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018, January 22-24). *Toward generating a new intrusion detection dataset and intrusion traffic characterization.* Proceedings of the 4th International Conference on Information Systems Security and Privacy, Funchal, Madeira, Portugal. https://doi.org/10.5220/0006639801080116

[14] Singh Panwar, S., Raiwani, Y., & Panwar, L. S. (2019, March 15). *Evaluation of network intrusion detection with features selection and machine learning algorithms on CICIDS-2017 dataset.* International Conference on Advances in Engineering Science Management & Technology 2019 Uttaranchal University, Dehradun, India. http://dx.doi.org/10.2139/ssrn.3394103

[15] Gao, D., Zhang, Y-X., & Zhao, Y-H. (2009). Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, *9*(2), 220-226. https://doi.org/10.1088/1674-4527/9/2/011

[16] Bhavani, T. T., Rao, M. K., & Reddy, A. M. (2020). Network Intrusion Detection System Using Random Forest and Decision Tree Machine Learning Techniques. In A. K. Luhach, J. A. Kosa, R. C. Poonia, X.-Z. Gao, & D. Singh (Eds.), *First International Conference*

on Sustainable Technologies for Computational Intelligence. Springer Singapore. https://doi.org/10.1007/978-981-15-0029-9_50

[17] Mirjalili, S. (2016). Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. *Neural Computing and Applications*, *27*(4), 1053-1073. https://doi.org/10.1007/s00521-015-1920-1

[18] Rohart, F., Gautier, B., Singh, A., & Lê Cao, K-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, *13*(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752

[19] Rashno, A., Nazari, B., Sadri, S., & Saraee, M. (2017). Effective pixel classification of Mars images based on ant colony optimization feature selection and extreme learning machine. *Neurocomputing*, *226*, 66-79. https://doi.org/10.1016/j.neucom.2016.11.030

[20] Joshi, R. (2016). Accuracy, Precision, Recall & F1 Score: interpretation of performance measures.