Defence Science Journal, Vol. 74, No. 2, March 2024, pp. 246-255, DOI : 10.14429/dsj.74.18953

A Comprehensive Review of Dimensionality Reduction Techniques for Real-time Network Intrusion Detection with Applications in Cybersecurity

Rohan Gondhalekar[#] and Rajan Chattamvelli[$,*]

[#]*Department of Mathematics, Vellore Institute of Technology, Vellore - 632 014*
[$]*Department of CSE, Amrita Vishwa Vidyapeetham, Amaravati - 522 503*
**E-mail: cvrajan@gmail.com*

ABSTRACT

This paper reviews popular signature and anomaly-based intrusion detection systems (IDS). Dimensionality reduction techniques (DRT) are used to increase the efficiency of such systems for real-time operation. Autoencoder-based IDS is rapidly gaining in popularity, primarily due to its inherent ability to denoise data and reduce dimensionality. In addition to the efficiency, we also look at the classification techniques used by various authors, and the overall impact of a model in terms of performance metrics. This article is written for novices in cyber security to get a jumpstart on the latest IDS algorithms. The purpose is to give useful insights into the broad and progressive view of various techniques in wide use, expose high-impact future research areas and to summarize classic IDS methods and feature selection techniques.

Keywords: Attack-vector; Autoencoder; Deep-learning; Feature-extraction; Kernel; Principal components.

1. INTRODUCTION

The Internet has drastically changed the world in multiple ways. Several businesses worldwide profoundly depend on it. The heavy dependency on the Internet is actually a double-edged sword because it gives simplicity and convenience to users but also leads to serious intrusion threats, data leaks and other cybersecurity issues. The recent cyber-attack at the AIIMS, New Delhi (November 2022) in which data about 40 million patients including "VVIPs" were compromised by hackers (who demanded ₹200 crores ransom in cryptocurrency) is an incident that could have been detected by proper malware detection software. This must be a wakeup call for healthcare agencies, military and defence departments, and other government organisations that store enormous amounts of sensitive data in digital form on networked servers.

1.1 Importance of IDS

Intrusion detection (ID) and prevention, malware prevention, firewalls, user authentication, etc. are important to reduce, if not completely preclude cyberattacks[1]. IDS is the most popular component of any security stack. They are important to safeguard sensitive data, perceive policy violations, inhibit replication of ransomware, curtail down time (of affected servers), and provide a proactive security defence.

1.2 Challenges in IDS

IDS has become a major challenge due to the rapid deployment of Internet of Things (IoT), the complexity of modern wired and wireless communication networks that connect myriads of networked devices and the dynamics of intruder behaviour[2-3]. Host-based IDS (HB-IDS) are deployed to guard powerful hosts through which millions of data packets flow incessantly, whereas network-based IDS (NB-IDS) are deployed on an entire physical network. Attackers randomly change packet contents to disguise itself as genuine. The signatures, patterns, anomalies or deviations are captured and stored in databases or in-memory lists (signature datastore) for fast future processing[4]. This datastore is updated and adapted on an ongoing basis to detect emerging novel intrusions.

Received : 14 April 2023, Revised : 14 September 2023
Accepted : 29 November 2023, Online published : 26 March 2024

1.3 Cybersecurity

Cybersecurity deals with bigdata, which are inherently high-dimensional. Bigdata used in machine learning (ML) algorithms face some critical issues like sparseness. It affects the performance of some methods, especially deep learning (DL) algorithms. Effective DRT leads to faster feature selection (FS) and feature extraction (FE)[5].

Table 1. DDoS landscape

Year	Traffic (TB)	Technique used	Target
2023	Unk	Mitel vulnerability	Cloud flare client
2023	1.4	Mirai-variant botnet	US-ISP Cloud flare client)
2021	1.02	Amplification	Imperva client
2020	2.3	CLDAP reflection	Amazon AWS
2018	1.7	Memcache	US client
2018	1.35	Memcache	Github
2017	2.5	Spoofing	Google

1.4 Motivation for Our Work

As hackers invent newer and faster ways to crack into secure networks, every device and connected systems need more cybersecurity. The Wanna Cry ransomware attack on May 12, 2017 infected thousands of unpatched Windows systems, encrypted system files, locked them down and demanded a ransom to be paid to unlock them[6]. This was an eye-opener for security professionals worldwide.

AI-based models that steal passwords by listening to the keyboard sound produced when the password is typed-in already exist, and are being used by hackers in public places. AI-driven attacks can easily sidestep security barriers such as multi-factor authentication. Hence it is of paramount importance to develop advanced techniques that goes in parallel with the innovative and sophisticated hacking techniques being developed to ensure cent-percent protection.

Real-world data packets could contain missing-values (represented as Null, NaN, blanks, etc). Most anti-malware programs either drop such packets or impute them using kNN or tree-based methods. Hackers deliberately introduce missing values randomly in data packets to disguise them as genuine. Hence dropping such data packets may result in skewed detection models.

1.5 Botnet and DDoS Attacks

A new type of attack called a "botnet attack" has recently emerged in which a distant hacker transmits remote commands to a networked device to gain control, and keeps on infecting more machines meticulously to create a network of remotely orchestrated devices under their direct control. Each "bot" in the botnet can serve as a unique vantage-point, and then initiate a local attack like a distributed denial-of-service (DDoS) attack. An HTTP DDoS attack fires a flood of instantaneous HTTP requests toward a target-website. An evolution in botnet DNA called virtual-machine (VM)-based DDoS botnet, which is up to 5000 times stronger in attack-capability, by leveraging cloud-computing capabilities has arrived (blog. cloudflare.com). Hyper-volumetric DDoS attacks are those involving more than 50 million requests per second (RPS). Several such waves of attacks have recently been launched by hacktivist groups REvil, Kill net and Anonymous Sudan Hacktivist alliance dubbed "Darknet Parliament" (Table 1). These originated from numerous cloud providers with over 30000 IP addresses. Adaptive-DDoS-detection mitigates such attacks more intelligently using unique data-traffic patterns. A new breed of hackers called "hackers-for-hire" have emerged to launch DDoS attacks on payment basis.

Intruders could mask their identity after gaining access, and transmit data packets from infected systems using spoofed IP addresses on cloud servers. Such attack patterns could evolve and mature over time and exploit new vulnerabilities, making them more difficult to detect.

The service-downtime during which a network becomes unusable varies from $300000 to $1 million per hour[7]. It can also seriously damage brand-reputation. Hence the ability to detect hitherto-unknown attacks is becoming increasingly important. Deep neural networks (DNN) that are highly scalable are ideal in such situations[8].

1.6 Classic IDS Methods

Three major ways exist to detect intruders (Anomaly-based IDS (AB-IDS), signature-based IDS (SB-IDS) and stateful protocol-based IDS (SPB-IDS)). The AB-IDS uses AI techniques to detect anomalous attacks using inconsistent or erratic resource consumption or exploitation. One example is the "virtual disk" created in RAM. Intruders create such virtual disks, zip sensitive data onto them, and transmit it to remote IP addresses using tiny ftp. The SB-IDS identifies network attacks based on known signatures. AB-IDS is more powerful in detecting unknown attacks, and can serve as a feeder-mechanism to the SB-IDS (the attack-vector of new attack-types is saved as its signature).

The SPB-IDS uses universal profiles. With rapid developments in ML models and its wide adaptation by industry, military and defence organisations, AI-based[9] and regex-pattern-matching-based[10] IDS are becoming popular. Federated Learning (FL) algorithms are used in intricate distributed networks. In Bertoli[11], *et al.* a stacked-unsupervised FL approach was introduced. Another is the lightweight DL model (DLM), which exploits the power and accuracy of DL models and effectively reduces computational cost using GPU and Google Colab.

Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), k-Nearest Neighbour (k-NN), Multi-Layer Perceptron (MLP) and DLM are popular for data classification. Hybrid or ensemble approaches are becoming widespread[12,64]. A hybrid approach combines good features of above methods for better accuracy[13].

1.7 Autoencoder-based IDS

An autoencoder is the *de facto* standard used to learn nonlinearity in features. It is an unsupervised learning model in which the number of nodes in the input and output layers are the same, but nodes in the middle layer(s) could be fewer. Several variants exist such as stacked autoencoders, deep sparse autoencoders, denoising autoencoders, variational autoencoders, etc. It is implemented as a feed-forward non recurrent NN that learns to reconstruct data from compressed versions of itself. Autoencoders are not generic (like MP3 or JPEG) because they employ lossy compression, and can only compress data similar to what they have been trained for. Thus separate autoencoders are needed for different data types[14-17,71]. Parallel deep autoencoder is used to detect intrusion in IoT[18-19]. A hybrid IDS using a sparse autoencoder and DNN is discussed in Narayanarao[20], *et al.* Hou[21], *et al.* .uses an asymmetric autoencoder that extracts the latent features of network traffic with two different CNNs. Li & Liu[22] used AE-IDS based on RF to improve the prediction accuracy. Nazir & Ahmed[41] used a recursive feature addition algorithm using RF to select the optimal subset of features. When combined, these approaches act as force-multipliers to progressively make the intrusion detection task simpler and detection accuracy better.

2. RELATED WORKS

Many reviews on IDS have appeared recently Li & Liu[22]. In references[23-26], the authors discuss FS and FE techniques in IDS. The filter-based and wrapper-based methods appear

in[27]. Ferreira[28] *, et al.* uses a supervised fitness-filter technique with a relevance redundancy technique for feature selection. The feature relevance and redundancy are measured by mutual information metric[29]. Fisher's score, which is a ratio of "between group"sums of squares over "within group" sum of squares $F = \sum_i n_i \left(\overline{x}_i - \overline{\{x\}}\right)^2 / \sum_i n_i (x_i - \overline{x}_i)^2$ may be used to discriminate important features. Large F-values indicate higher discriminatory power. Data outliers must be removed before computing F. However, outliers can't be ignored in IDS because they can be due to a novel intrusion technique or a unique signature. Outliers are analysed separately from the rest to reveal abnormalities, if any.

Sarker[30], *et al.*, discuss cybersecurity and the challenges in data collection. A review of99 selected articles appears in Suryotrisongko & Musashi[31]. They identified 24 theories applicable in cybersecurity, and reviewed 11 papers that discuss the use of advanced ML, NN, and DL.

Almaiah[32], *et al.* identified unsolved research challenges and unexplored research topics from each of the features encountered.

3. METHODOLOGY

This article reviews solutions to the following research objectives.

RO1) : Study various DRT approaches in intrusion detection based on ML.

RO2) : Study different classification techniques hybridised with DRT to identify different attacks /intrusions in a network.

RO3) : Check the present popularity of different datasets.

3.1 Dimensionality Reduction (DR)

Computational complexity of most ML algorithms depends upon the input dimension and the number of samples. Hence DRT is helpful to reduce the complexity. Data used in cybersecurity have a wide range of features (connection and content features, and statistical characteristics of network traffic). DRT uses FS and FE. FS can be performed in several ways using correlation, variance, information gain of variables, Fisher's score, Latent Discriminant Analysis (LDA), Principal Component Analysis (PCA)[32-33,63] etc. The reduction-ratio (RR) is the ideal number of principal components needed for intrusion detection.

3.1.1 Feature Selection

Useful features are extracted by selecting important variables[5,26-27]. The CyDDoS architecture used in Lopes[65], *et al.* combines a DNN with an ensemble of five ML algorithms for FS. Unsupervised meta-learning algorithms have better detection capabilities and may even outperform the classification performance of other ML algorithms when dealing with unknown attacks[34-35].

3.1.2 Feature Extraction (FE)

FE is used to acquire important information by removing redundant features to improve classification efficacy[5]. Linear and nonlinear FE methods exist, of which nonlinear methods are more popular.

The main use of PCA is to reduce data dimensionality without information loss, in minimum time with high detection accuracy and low false-positive-rates (FPR)[7].

4. ANALYSIS OF ARTICLES

Most recent articles use FS, FE, ensemble or hybrid approaches for DR. Ogundokun[9], *et al.*, used DRT by Particle Swarm Optimization algorithm (PSO). Output of PSO is fed to k-NN and DT. A confusion matrix is formed using True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN) (Table 2). The PSO+DT ensemble achieved an accuracy of 89.6 %, FP rate 0.011, and FN rate 0.038, while the PSO+k-NN ensemble attained an accuracy of 96.2 %, a lower FP rate 0.004, and FN rate 0.104 (which obviously outperformed).

Nagaraju[10], *et al.* used Kalman Filter (KF) and Salp Swarm Algorithm (SSA) for DR. KF is used for reducing the repetitive data entries, whereas SSA is used for optimum feature selection that contains maximum information. Reduced data are classified to detect intrusion using multiclass classifier Kernel Extreme Learning Machine (KELM). They used three kernels, namely, sigmoid, Gaussian and hard limit functions.

Table 2. Evaluation metrics

Precision	TP/(TP+FP)
Recall	TP/(TP+FN)
Accuracy	(TP+TN)/(TP+TN+FP+FN)
FPR	FP/(FP+TN)
F-Score	2*Precision*Recall/(Precision+Recall)

They used NSL-KDD and CICIDS 2017 for training, and Tokyo 2006+, CICIDS 2018 for testing. Dimensionality was reduced from 41 to 13(70.7 % and 86.49 % accuracy). Using CICIDS2017,22 out of 80 features were selected. NSL-KDD achieved 99.9 %, and CICIDS2017 95.68 % detection accuracy.

The Multi-objective Artificial Bee Colony Optimization algorithm[36], PSO algorithm[37], SSA, wrapper based FS[38], multi-objective Grey Wolf Optimisation algorithm[39] are all used for DR.

Doreswamy & Gad[40] used an ensemble approach for DR based on Union and Quorum Combination techniques (UaQCT). They used 5 methods for feature selection:--(i) Analysis of Variance (ANOVA), (ii) Variance Threshold, (iii) Sequential Backward Search, (iv) Recursive Feature Elimination, and (v) Least Absolute Selection with Shrinkage Operator. UaQCT are used for balancing and equalising individual feature set. Individual FS and classification were done by RF, DT, k-NN, Gaussian Naive Bayes (GNB), and Logistic Regression (LR). Performance of each approach was tested based on cross validation for feature selection, F score, accuracy, precision and recall (Table 2). This proposed approach is compared on UNSW-NB15 and NSL-KDD. The results on UNSW-NB15 shows that RF with union and quorum feature sets produce F1-Score 99 % and 99.02 %respectively with minimum 6 and 12 features, whereas results for the NSL-KDD dataset for the

same combination produces F1-Score 99.34 % and 99.21 % with a minimum of 28 and 18 features respectively.

Nazir & Ahmed[41] proposed Tabu Search-Random Forest (TS-RF) on UNSW-NB15. They chose features in two ways:- (a) TS with score 1 and feature-weight >= 0.2 and (b) TS with score 1 and feature-weight >= 0.15. The selected 16 features reduced dimensionality to 62 % with the help of TS, increased the accuracy of classification for RF by 2 % and reduced the time-complexity by 40 %. TS-RF combination gave better performance than the others.

Omkar[42], *et al.* used Sequential Feature Selector (SFS) and Extra Tree (ET) (extremely Randomised Tree) for performance evaluation. The performance was tested over UNSW-NB15 and NSL-KDD using RF, ANN, SVM, Kernel SVM (K-SVM) and k-NN. On the UNSW-NB15, the SFS+RF performed with 99.97 % accuracy, 99.92 % recall, 99.98 % precision, and F-score of 0.9995.

Susanto[43], *et al.* investigated botnet-attacks in IoT networks with N-BaIoT dataset using Random Projection (RP),k-NN, DT, RF, AdaBoost (AD), and Gradient Boost (GB). Feature number reduction by 2, 3, 5, 10, and 15 are used to study the impact caused by RP with minimum features. DRT+RP and DT have detected IoT botnet within 8.44 secs (lowest among all combinations) with accuracy 100 %. FPR is low (nearly 0) compared to k-NN, RF, AD, and GB.

Pranto[44]et. al. classified network incoming traffic as normal or anomalous with NSL-KDD dataset using correlation[66] for DR by FS. A cut-off of 0.5 for feature correlation and variance is kept to remove unwanted features. Reduced dataset is used with k-NN (Euclidean norm), DT, NB, LR, and ensemble of previous 4 classifiers based on majority-voting and RF. RF classifiers gave 99.5 % maximum accuracy with a 0.6 % false alarm rate.

Kunal[45], *et al.* used attribute evaluation and an ensemble approach using NSL-KDD. Ranker search and information gain reduced the number of features from 41 to 11. The models were evaluated using an ensemble of IBk(k-NN), RT, REP Tree, J48graft and RF classifiers. Low FPR with an accuracy of 99.72% for binary classification and 99.68% for multi-class were observed.

The above articles have used FS methods using either filter-based or wrapper-based techniques. PCA-based articles are discussed below.

Song[46], *et al.* used PCA to reduce the dimensions to 19on KDD-CUP99. Reduced dataset is given to WOA(Whale Optimization Algorithm)-XGBoost(Extreme Gradient Boosting) algorithm. The average ACC, sensitivity, and specificity are 0.9906, 0.9958, and 0.9574, respectively.

Abdulhammed[47,55], *et al.* proposed a ML model using CICIDS2017. PCA reduced 81 features to 10, which is fed into RF, Bayesian Network, LDA and Quadratic Discriminant Analysis (QDA) classifiers. The dimension decreased from 81 to 10 with an accuracy 99.6 %.

PCA performance is effectively verified with different types of attacks by Desai[48] , *et al.* They identified intrusions with IoT botnet attacks. The performance of DT, RF and SVM revealed that RF Classifier (RFC) is better in multiclass classification. PCA+RFC performance was tested with different proportions of normal data and attack-types. Among all combinations, the highest accuracy attained is 99.97 %.

Dipon[49], *et al.* used 1999 DARPA dataset from the 4th and 5th week's testing. Output of PCA is fed into density-based-clustering model to count outliers and then ranked accordingly.

Some articles have used PCA in a hybridised way with other DRT or simply used Kernel-PCA (K-PCA) for better performance. Lu[50], *et. al.* proposed a hybrid model to detect SB-IDS based on ML, NN and K-PCA for DR. Anomaly detection used an Extreme Learning Machine (ELM) with a Hybrid Kernel Function (HKELM). Parameters of HKELM are optimised with the help of a blended version of Gravitational Search Algorithm (GSA) and Differential Evolution Algorithm (DEA). These authors proposed a combination of KPCA-DEGSA-HKELM for IDs. Model was tested on 3 datasets KDD99 UNSW-NB15 and the Industrial Intrusion Detection (IID) dataset from the Tennessee Eastman. They were able to achieve an accuracy of 95.82 % which is higher than the accuracy obtained in other works.

A new trend is observed in recent works in which PCA is combined with other DRT, and FS or FE methods to improve detection accuracy. This has helped to lower the FPR.Salo[51]et. al. used ISCX2012, NSL-KDD and Kyoto 2006+. A DRT is proposed in which FS is performed using filtering method. In ISCX2012 there were 19 features out of which 9 were selected by IG; and PCA compressed it to 7. In NSL-KDD, out of 41 features, IG selected 13 whereas in Kyoto 2006+ of 24 features IG selected 10 and, in both datasets, PCA compressed it to 12 PCs. The Average-of-Probabilities (AoP) based vote classifier is used for ensemble to gear up the performance. The accuracies achieved by the models were 99.01 % (ISCX2012), 98.24 % (NSL-KDD) and 98.95 % (Kyoto 2006+) with 7, 12, and 12 features, respectively.

DRT and Tri-Light GBM classifiers were used in Apruzzese[52], *et al.* using a semi-supervised learning approach and Information gain to reduce redundant features. Models were tested on UNSW-NB15.

Zhang[53], *et al.* used semi-supervised ML approach to detect AB-IDS. They proposed Multi-Strategy Feature Filtering, PCA and stratified sampling technique based on upgraded Tri-Light Gradient Boosting Machine (Tri-LightGBM). Tri-Light GBM is used to detect intrusions. This increases Accuracy, Recall, Precision, and F-measure by 0.5%.

Some articles compare PCA with other DRTs such as Autoencoder (AE), t-SNE, Chi-square etc. Jieling[54], *et al.* compared Autoencoder and PCA. AE reduced features from 81 to 59 and PCA from 81 to 10, resulting in 99.6 % accuracy.

Varunram[56], *et al.* compared the performance of DRT on ML based IDS using CICIDS (2017) Friday-Working-Hours-Data set (which is imbalanced) with PCA, t-SNE and UMAP for DR. This is fed to different classifiers (LR, SVM, Naïve Bayes, ANN, k-NN, AdaBoost and RF). PCA+ AdaBoost has better performance, but took 150 mins. The tSNE+AdaBoost and UMAP+AdaBoost respectively took 240 and 210 mins.

Samdekar[57], *et al.* compared the performance of different DRT over intrusion detection in IoT dataset BoT-IoT with 40 features. They used Chi-square and Extra-Tree-Classifier (ETC) for FS and PCA and Firefly-Algorithm for

Table 3. Comprehensive comparison of recent intrusion detection systems

Authors (Year)	Dataset	# of features	DR technique	Reduced Dim.	Classification Technique
Ogundokun, *et al.* (2021)	KDD CUP 99	41	PSO	--	k-NN,DT
Doreswamy, *et al.* (2020)	UNSW-NB15 NSL-KDD	42 41	Ensemble	--	RF, DT, k-NN, LR, Naïve Bayes
Gavel, *et al.* (2021)	KYOTO 2006 CICIDS2018 AWS	41 80	Kalman Filter + Salp Swarm Algorithm	13 22	KELM
Nazir, *et al.* (2021)	UNSW-NB15	41	Tabu Search, Chi-square	--	RF, k-NN, Perceptron Naïve Bayes
Shende, *et al.* (2022)	UNSW-NB15 NSL-KDD	42 41	Seq. Feature Selection	15	RF, DT, k-NN, LR kernel SVM, Naïve Bayes
Susanto, *et al.* (2021)	N-BaIoT	115	Random Projection	15	RF, k-NN, DT AdaBoost
Pranto, *et al.* (2022)	NSL- KDD	41	Correlation	---	k-NN,CART,LR, RF, Logistic Regression Naïve Bayes, Ensemble
Kunal, *et al.* (2020)	NSL-KDD	41	Information gain with Ranker Search method	11	Ensemble of IBk(KNN),RT,RF, REP Tree, j48graft
Song, *et al.* (2022)	KUD-CUP99	41	PCA	19	WOA, XGBoost
Abdulhammed, *et al.* (2019)	CICIDS2017	81	PCA	10	Bayesian Network RF, LDA,QDA
Desai, *et al.* (2020)	IoT NID dataset	15	PCA	10	DT, RF, SVM
T.M. Dipon, *et al.* (2020)	1999 DARPA	---	PCA		Density Clustering
Lu Lv, *et al.* (2020)	KDD-CUP99, UNSW-NB15 IID Dataset		Kernel PCA		HKELM+ GSA+ DEA
Salo, *et al.* (2019)	ISCX2012 NSL-KDD Kyoto 2006+	19 41 24	Combination of IG and PCA	9(IG)- 7(PCA) 13(IG) 12(PCA) 10(IG) 12(PCA)	SVM, IBK(k-NN) MLP Ensemble of above 3
Zhang, *et al.* (2020)	UNSW-NB15	42	IG+PCA	---	TriLight-GBM
Zhang, *et al.* (2022)	UNSW-NB15 CICIDS2017	42 74	Multi Strategy (Fisher Score+Information Gain)+PCA	21(PCA) 2(PCA)	TriLight-GBM
Abdulhammed, *et al.* (2019)	CICIDS2017	81	Autoencode PCA	59 10	RF, LDA, QDA Bayesian Network
Varunram, *et al.* (2021)	CICIDS (2017) Friday Working Hours	78	PCA t-SNE UMAP	--- --- ---	LR, k-NN, RF, ANN, SVM Naïve Bayes, AdaBoost
Samdekar, *et al.* (2021)	BoT-IoT	40	Chi-square, PCA,FA, Extra Tree classifier	10 10 ---	Firefly Algorithm SVM

IG= Information Gain, LR=Logistic Regression, DT=Decision Tree, RF= Random Forest, LDA= Linear Discriminant Analysis, QDA= Quadratic Discriminant Analysis , k-NN= k-Nearest Neighbour, ANN= Artificial Neural Network, KELM = Kernel Extreme Learning Machine, FA= Firefly Algorithm, DEA= Differential Evolution Algorithm, GSA= Gravitational Search Algorithm, PSO= Particle Swarm Optimization, SSA=Salp Swarm Algorithm, WOA=Whale Optimization Algorithm, XGBoost=Extreme Gradient Boosting, HKELM =Hybrid Kernel Function, SVM= Support Vector Machine, IID=Industrial Intrusion Detection (from Tennessee Eastman Process)

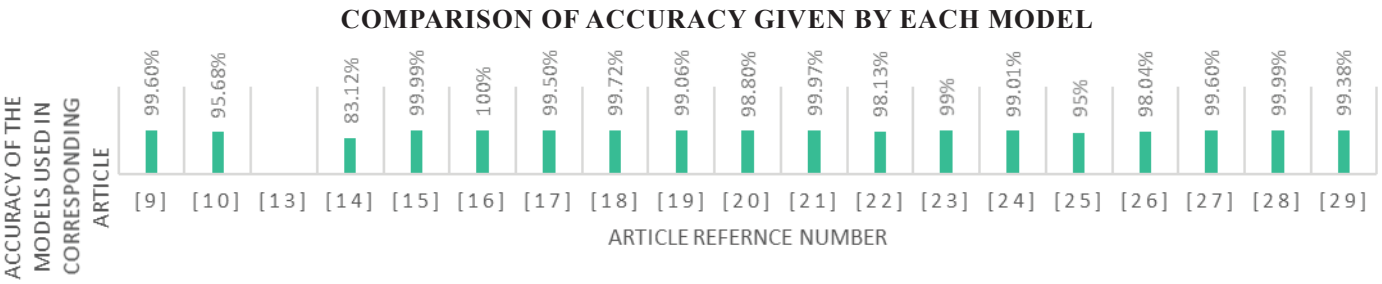

Figure 1. Accuracy comparison.

Table 4. Accuracy comparison of recent intrusion detection systems

Ref. #	Dataset	Dimensionality Reduction	Classifier	Accuracy (%)
[9]	KDD CUP 99	PSO	k-NN	99.60
[37]	CICIDS2017	Kalman Filter + SSA	kernel ELM-Based multiclass	95.68
[41]	UNSW-NB15	Tabu Search	Random Forest (RF)	83.12
[42]	NSL-KDD	Sequential Feature Selector	RF	99.99
[16]	N-BaIoT	Random Projection	DT	100
[44]	NSL- KDD	Correlation	RF	99.50
[45]	NSL- KDD	IG with Ranker Search method	Ensemble of IBk (KNN), RT, REP Tree, j48graft, RF	99.72
[46]	KDD CUP 99	PCA	WOA + XGBoost	99.06
[47]	CICIDS2017	PCA with UDBB	RF	98.80
[48]	IoT NID dataset	PCA	RF	99.97
[49]	1999 DARPA	PCA	Density Based Clustering	98.13
[50]	KDD-CUP99	Kernel PCA	Hybrid Kernel Function (HKELM)+GSA+DEA	99
[51]	ISCX2012	Combination of IG and PCA	Ensemble of IBK(k-NN), SVM, Multi Linear Perceptron	99.01
[52]	UNSW-NB15	IG + PCA	Tri Light GBM	95
[53]	CICIDS2017	Multi Strategy (Fisher Score+ IG + PCA	Tri Light GBM	98.04
[54]	CICIDS2017	PCA	RF	99.60
[56]	CICIDS (2017) Friday Working Hours	t-SNE	RF	99.99
[57]	BoT-IoT	Firefly Algorithm	SVM	99.38

data compression. Output is loaded to SVM. Firefly+SVM gave 99.38 % Accuracy, 99.30 % Recall, 100 % Precision and 99.87% F1-Score. Results are compared with Chi-square+SVM, ETC+SVM and PCA+SVM.

Fuzzy Optimised Independent Component Analysis for DR using three classifiers Hyper-heuristic SVM (HH-SVM), Hyper-Heuristic-Improved PSO-based SVM and Hyper-Heuristic Firefly-algorithm based Convolutional Neural Networks (HHFA-CNN) is used in Aswanandini & Deepa[58]. The accuracies for HH-SVM+FOCIA, HHIPSO-FOCIA and HHFA-CNN+FOCIA over NSL-KDD are 91.67, 94.45 and 97.75, and for ISCX-IDS are 88.33, 94.50 and 95.67 respectively. The accuracies of HH-SVM+FOCIA, HHIPSO-FOCIA and HHFA-CNN+FOCIA are 91.67, 93.74 and 98.89.

5. DISCUSSION

Analysis is undertaken to filter out the essential attributes for each attack-type. Quick detection is highly preferred in IDS, especially in banking, military and defence organisations. Accuracy is the predominant metric used. Additionally, KDD-CUP99, NSL-KDD, CICIDS2017,2018 and UNSW-NB15 are the datasets most commonly used. One article is on DHARPA1999, and two are using KYOTO 2006+, N-BaIOT, and BOT-IOT. These appear in Table 3 and Fig. 1.

Table 4 and Fig. 2 show that PCA is a good choice to reduce the dimension, which leads to improvement in accuracy, detection-rate, reduction in time, etc. Other methods exist for the same purpose[31]. Chattamvelli[29], states that the SVM works in the data-space whereas the PCA works in the attribute-space (ie. SVM identifies boundary-data-points (intrusion attempts) while PCA identifies the most important variables (features)). Thus an autoencoder-based ensemble approach is a good choice for IDS[16,64].

6. CONCLUSIONS

This study reviewed latest IDS that use DR using PCA. Computational complexity is of secondary importance when data are highly sensitive (as in defence and military organisations) because data security cannot be compromised. Computational efforts can be drastically reduced using DRT, resulting in light-weight IDS (Fig. 3).

A simulation of various algorithms revealed that maximum detection-accuracy is achieved by removing irrelevant and redundant features using DRT. This helps to narrow-down the solution-search-space resulting in faster detection (Table 4). Table 5 gives the running time of selected algorithms on an Intel Core i5 processor with 16 GB RAM.

Using AI and DL techniques together gain the benefits of

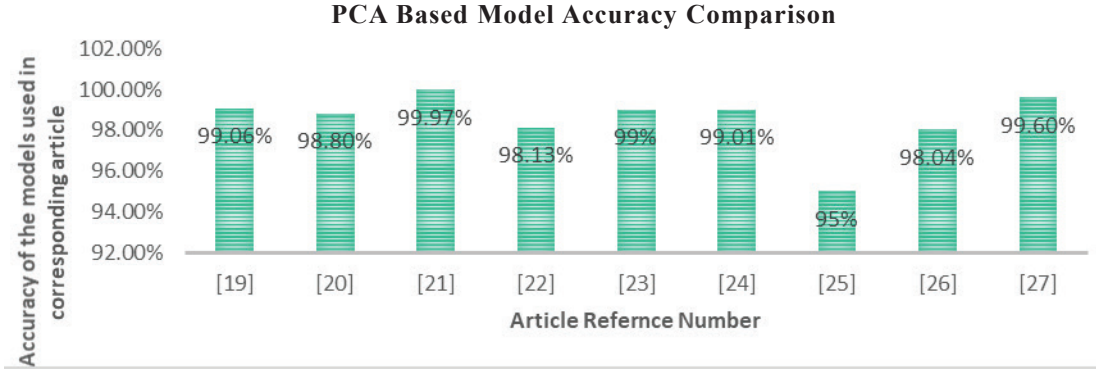

Figure 2. Model Accuracy of PCA algorithms for IDS.

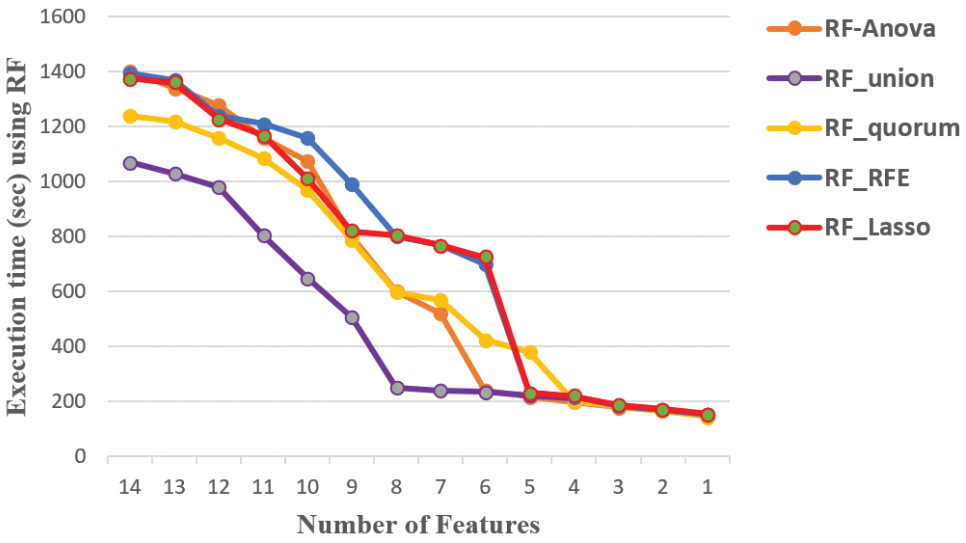

Figure 3. Execution time comparison.

Table 5. Time comparison

Dataset	Taining Time (in sec)	Testing Time (in sec)
KDD99	14.84	0.18
UNSW-NB15	43.36	2.58
TE intrusion dataset	21.29	0.14

both in exploring the solution-space effectively[59-60]. Our study brings a high added-value to next generation IDS where DL algorithms are becoming prominent[60-62]. There have been rapid changes from SB-IDS to AB-IDS[48], and recently into hybrid approaches for effective intrusion detection[47]. *Apollon* is a novel defense system to hinder adversarial machine learning (AML) attacks[67].

REFERENCES

1. Al-Jarrah, O.; Alhussein, Omar.; Yoo, Paul D.; Muhaidat, Sami.; Taha, Kmal. & Kim, Kwanjgo. Data randomization and cluster-based partitioning for Botnet intrusion detection. *IEEE Trans. Cybern.*, 2016, **46**(8), 1796-1806. doi: 10.1109/TCYB.2015.2490802.
2. Santhoshkumar, S.V.N.; Selvi, M. & Kannan, A. A comprehensive survey on machine learning-based intrusion detection systems for secure communication in IoT. *Comput. Intel. Neurosci.*, 2023.
3. Chaganti, R.; Suliman, Wael.; Ravi, Vinayakumar. & Dua, Amit. Deep learning approach for SDN-enabled intrusion detection system in IoT networks. *Inf.*, 2023, **14**(1), 41.
4. Malek, Z.; Trivedi, Bhushan H. & Shah, Axita. User behavior pattern-signature based intrusion detection. Fourth world conference on smart trends in system, security and sustainability. 2020, 549-552. doi: 10.1109/WorldS450073.2020.9210368
5. García,S.; Luengo, J. & Herrera. Feature selection. *Intell. Syst. Ref. Libr.*, 2015, **72**(6), 163-193. doi: 10.1007/978-3-319-10247-4. doi: 10.1007/978-3-319-10247-4_7.
6. Chansarkar, S. & Rai, N. Cyberspace security: An overview for beginners. *Def. Sci. J.*, 2017, **67**(4), 483-484. doi: 10.14429/dsj.67.11542
7. Nicholson, P. Five most famous DDoS attacks and then some, https://www.a10networks.com/blog/5-most-famous-ddos-attacks/, 2021.
8. Thakkar, A. & Lohiya, A.R. Fusion of statistical importance for feature selection in Deep Neural Network-based IDS. *Info. Fusion.*, 2023. doi: 10.1016/j.inffus.2022.09.026
9. Kanimozhi, V. & Jacob, T. Artificial intelligence outflanks all other machine learning classifiers in network IDS on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. *ICT Express.*, 2020, **7(**3), 366-370. doi: 10.1016/j.icte.2020.12.004
10. Nagaraju, S.; Shanmugham, B. & Baskaran, K. High throughput token-driven FSM-based regex pattern matching for network intrusion detection. *Mater. Today: Proc*., 2021, **47,** 139-143. doi: 10.1016/j.matpr.2021.04.028
11. Bertoli, G.; Junior, L.; Saotome, O. & Santos, A. Generalizing intrusion detection for heterogeneous networks: A stacked-unsupervised federated learning approach. *Comps. Secu.*, 2023, **127**, 103106. doi: 10.1016/j.cose.2023.103106
12. Sarkar, A.; Sharma, H. & Singh, M. A supervised machine learning-based solution for efficient network intrusion detection using ensemble learning. *Int. J. Info. Tech.*, 2023, **15**(1), 423-434.
13. Maseno, E.; Wang, Z. & Xing, H. A systematic review on hybrid intrusion detection system. *Secur. Commu. Network*., 2022, 9663052 doi: 10.1155/2022/9663052
14. Shiomoto, K. Network Intrusion Detection System based on an Adversarial Auto-Encoder with few Labeled Training Samples. *J . Network. Syst. Manage.,* 2023, **31**, 5. doi: 10.1007/s10922-022-09698-w
15. Khudhu, A. & Samsudin, K. IoT Intrusion Detection using Autoencoder and Machine Learning Techniques. *J. Comp. Sci.*, 2022, ***18***(10), 904-912. doi: 10.3844/jcssp.2022.904.912
16. Yousefi-Azar, M.; Varadharajan, V.; Hamey, L. & Tupakula, Uday. Autoencoder-based feature learning for cybersecurity applications. *In* Proceedings of 2017 International Joint Conference Neural Networks (IJCNN), IEE., 2017, 3854-3861. doi: 10.1109/IJCNN.2017.7966342
17. Li, X-K.; Chen, W.; Zhang, Q. & Wu, Lifa. Building Auto-Encoder Intrusion Detection System based on random forest feature selection. *Comps. Secu.*, 2020, **95.**
18. Basmati, A. & Faghih M. PDAE: Efficient Network Intrusion Detection in IoT using Parallel Deep Autoencoder. *Inf. Sci.*, 2022, **598**, 57-74. doi: 10.1016/j.ins.2022.03.065
19. Yadav, S. & Kalpana, R. Recurrent nonsymmetric deep autoencoder approach for network intrusion detection system. *Meas.: Sens.*, 2022, **24.** doi: 10.1016/j.measen.2022.100527
20. Narayanarao, K.; Venkaraeao, K. & P.V.G.D., Prasad. A hybrid intrusion detection system based on sparse autoencoder and deep neural network. *Comput. Commu.,* 2021, 180, 77-88. doi: 10.1016/j.comcom.2021.08.026
21. Hou, Y.; Fu, Yusheng.; Guo, Jinhong.; Xu, Jie.; Liu, Renting. & Xiang, X. Hybrid intrusion detection model based on a designed autoencoder. *J. Ambient Intell. Humaniz. Comput*., 2022.
22. Y, Li., & Liu, Q. A comprehensive review study of cyber-attacks and cybersecurity. *Energy Reports*, 2021, **7,** 8176–8186. doi: 10.1016/j.egyr.2021.08.126
23. Yang, Z.; Liu, X.; LI, Tong.; Wu, Di.; Wang, J.; Zhao, Y. & Han, H. A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Comput. Secur.*, 2022, **116,** https://www.sciencedirect.com/science/article/pii/S0167404822000736.
24. Sorzano, A. & Vargas, J. A survey of dimensionality

reduction techniques based on random projection. 12th International Joint Conference, VISIGRAPP 2017, 1–35, doi: 978-3-030-12208-9.

25. Jia,W.; Sun, M.; Lian, J.; Hou, S. Feature dimensionality reduction : a review. *Complex Intell. Syst.*, 2022.
26. Taher, K.; Jisan, B. & Rahman, M. Network intrusion detection using supervised machine learning technique with feature selection. ICREST, 2019, 643-646. IEEE. doi: 10.1109/ICREST.2019.8644161
27. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.*, 2014, **40**(1), 16–28. doi: 10.1016/j.compeleceng.2013.11.024.
28. Ferreira, A. & Figueiredo, M. On the improvement of feature selection techniques: The Fitness filter. *In* Proceeding 10th International Conference on Pattern Recogination Application and Methods., 2021, 365-372.
29. Chattamvelli, R. Data Mining Methods, ed. 2nd, Narosa, New Delhi, 2016.
30. Sarker, I. H.; Kayes, A.S.M.; Badsh, S.; Alqahtani, H.; Paul, W. & Ng, Alex. Cybersecurity data science: an overview from machine learning perspective. *J. Big Data.*, 2020, **7**(1).
31. Suryotrisongko, H. & Musashi, Y. Review of cybersecurity research topics, taxonomy and challenges. Proceeding *IEEE* 12th Conference Service Oriented Computing and Application (SOCA), 2019, 162–167. doi: 10.1109/SOCA.2019.00031
32. Almaiah, M.; Almomani, O.; Alsaaidah, A.; Al-Otaibi, S.; Hani, N.; Hwaitat, A.; Zahrani, A.; Lutfi, A.; Awad, Ali. & Aldhyani, T. Performance Investigation of Principal Component Analysis for IDS using Different SVM kernels. *Electron.*, 2022, **11**(21), 3571. doi: 10.3390/electronics11213571
33. Mishra, A.; Cheng, A. & Zhang, Y. Intrusion Detection Using PCA and SVM. *IEEE* 16th International Conference Control & Automation (ICCA), 2020, 907-912. doi: 10.1109/ICCA51439.2020.9264568
34. Zoppi, T.; Ceccarelli, A.; Puccetti, T.; & Bondavalli, A. Which algorithm can detect unknown attacks. *Comput. Secur.*, 2023, **127,** 103107. doi: 10.1016/j.cose.2023.103107
35. Shieh, C.S.; Nguyen, T.T.; Chen, C.H. & Miu, D. Detection of unknown DDoS attacks with deep learning and Gaussian mixture model. *Appl. Sci.*, 2021, **11**(11), 5213. doi: 10.1109/ICICT52872.2021.00012
36. Ogundokun, R.O.; Awotunde, J.B.; Sadiku, P.; Adeniyi, E.A.; Abiodun, M. & Dauda, O. An enhanced intrusion detection system using particle swarm optimization feature extraction technique. *Procedia Comput. Sci.*, 2021, **193,** 504–512. doi: 10.1016/j.procs.2021.10.052
37. Gavel, S.; Raghuvanshi, A.S. & Tiwari, S. Distributed intrusion detection scheme using dual-axis dimensionality reduction for Internet of things (IoT). *J. Supercomput.*, 2021, **77**(9), 10488–10511. doi: 10.1007/s11227-021-03697-5
38. Alamiedy, T.A.; Anbar, M.; Alqattan, Z. & Alzubi, Q. Anomaly-based intrusion detection system using multiobjective grey wolf optimisation algorithm. *J. Ambient Intell. Humaniz. Comput.*, 2020, **11**(9), 3735–3756. doi: 10.1007/s12652-019-01569-8
39. Ghanem, W. & Jantan, A. Novel multi-objective artificial bee colony optimization for wrapper based feature selection in intruction detection. *Int. J. Adv. Soft Comput. Appl.*, 2016, **8**(1), 70–81.
40. Doreswamy, M. & Gad, I. Feature selection approach using ensemble learning for network anomaly detection. *CAAI Trans. Intell. Technol.*, 2020, **5**(4), 283–293. doi:10.1049/trit.2020.0073
41. Nazir, A. & Ahmed, R. A novel combinatorial optimization based feature selection method for network intrusion. *Comput. Secur.*, 2021, **102**, 102164. doi: 10.1016/j.cose.2020.102164
42. Shende, O.; Pateriya, R.K. & Verma, P. A n-binary Classification and Grouping-based Approach to Improve the Performance of Anomaly Detection. *Arab. J. Sci. Eng.*, 2022, **47**(2), 1275–1287. doi: 10.1007/s13369-021-05871-6
43. Susanto.; Stiawan, D.; Idris, Yazid.; Arifin, M.; Rejito, J. & Budiarto, R. A dimensionality reduction approach for machine learning based IoT Botnet detection. *In* International Conference on Electrical Engineering, Computer Science and Informatics., 2021, 26–30. doi: 10.23919/EECSI53397.2021.9624299
44. Pranto, M. B.; Ratul, H. B.; Rahman, M.; Zahir, Z.B. & Jahan, I. Performance of machine learning techniques in anomaly detection with basic feature selection strategy. *J. Adv. Inf. Technol.*, 2022, **13**(1), 36–44. doi:10.12720/jait.13.1.36-44
45. Kunal,K. & Dua, M. Attribute selection and ensemble classifier based novel approach to Intrusion Detection System. *Procedia Comput. Sci.*, 2020, **167**(9), 2191–2199. doi: 10.1016/j.procs.2020.03.271
46. Song, Y.; Li, H.; Xu, P.; Liu, Dan. & Cheng, S. A Method of Intrusion Detection Based on WOA-XGBoost Algorithm. *Discret. Dyn. Nat. Soc.*, 2022, 5245622. doi: 10.1155/2022/5245622
47. Abdulhammed, R.; Faezipour, M.; Musafer, H. & Abuzneid, A. Efficient network intrusion detection using PCA-based dimensionality reduction of features. International Symposium Networks, Computer & Communication (ISNCC), 2019, 1–6. doi: 10.1109/ISNCC.2019.8909140
48. Desai, M.; Shi, Y. & Suo, K. IoT Bonet and Network Intrusion Detection using Dimensionality Reduction and Supervised Machine Learning. 11th *IEEE* Ubiquitous Computing Electronics & Mobile Communication Conference UEMCON, 2020, 0316–0322. doi: 10.1109/UEMCON51285.2020.9298146
49. Dipon, T.; Hossain, S. & Narman, H.S. Detecting network intrusion through anomalous packet identification. *In* 30th International Telecommunication Networks

Application Conference ITNAC, 2020. doi: 10.1109/ITNAC50341.2020.9315029
50. Lu, L.; Wang, W.; Zhang, Z. & Liu, X. A novel intrusion detection system based on an optimal hybrid kernel extreme learning machine. *Knowl. Based Syst.*, 2020, **195.**
doi: 10.1016/j.knosys.2020.105648
51. Salo, F.; Nassif. A.B. & Essex, A. Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection. *Comput. Networks.*, 2019, **148**, 164–175.
doi:10.1016/j.comnet.2018.11.010
52. Apruzzese, G.; Pajola, L. & Conti, M. The Cross-Evaluation of Machine Learning-Based Network Intrusion Detection Systems. *IEEE Trans. Network Serv. Manage*., 2022, **19**(4), 5152
doi:10.1109/TNSM.2022.3157344
53. Zhang, H. & Li, J. A New Network Intrusion Detection based on Semisupervised Dimensionality Reduction and Tri-LightGBM. Proceeding International Conference Pervasive Artif. Intell. ICPAI 2020, 35–40.
doi: 10.1109/ICPAI51961.2020.00014
54. Jieling, Z.; Li, J.; Liu, Y. & Liu, Z. Semi-supervised machine learning framework for network. 2022, Springer, p.8.
doi: 10.1007/s11227-022-04390-x
55. Abdulhammed, R.; Musafer, H.; Alessa, A.; Faezipour, M. & Abuzneid, A. Features dimensionality reduction approaches for machine learning based network intrusion detection. *Electron.*, 2019, **8**(3).
doi: 10.3390/electronics8030322
56. Varunram, T. N.; Shivaprasad, M.B.; Aishwarya, K.H.; Balraj, Anush.; Savish, S.V. & Ullas, S. Analysis of Different Dimensionality Reduction Techniques and Machine Learning Algorithms for an Intrusion Detection System. *IEEE* 6th International Conference Computing & Communication Automation, ICCCA, 2021, 237–242.
doi: 10.1109/ICCCA52192.2021.9666265
57. Samdekar, R.; Konda, S. & Ghosh, S.M. Efficiency Enhancement of Intrusion Detection in Bioinspire. *In* Proceeding 3rd Internationl Conference Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021) *IEEE,* 2021.
doi: 10.1109/ICICV50876.2021.9388392
58. Aswanandini,R. & Deepa, C. Dimensionality reduction of the attributes using fuzzy optimized independent component analysis. *Sci. Tech. J. Inf. Technol. Mech. Opt.*, 2022, **22(**1), 93–100.
doi: 10.17586/2226-1494-2022-22-1-93-100
59. Sohi, S.; Seifert, J.P. & Ganji, F. RNNIDS: Enhancing network intrusion detection systems through deep learning. *Comp. Secu.*, 2021, **102**, 102151.
doi: 10.1016/j.cose.2020.102151
60. Elsayed, M.S.; Le-Khac, N.; Jurcut, A. & Dev, S. Ddosnet: A deep-learning model for detecting network attacks. *IEEE* 21st International Symposium *Mobile & Multimedia Network.*, 2020, 391–396.
doi: 10.1109/WoWMoM49955.2020.00072
61. Sarker, I.H. Deep learning: a comprehensive overview. *SN Comp. Sci.*, 2021, **2**(6), 1-20.
doi: 10.1007/s42979-021-00815-1
62. Lampe, B. & Meng, W. A survey of deep learning-based intrusion detection in automotive applications. *Expert Sys. Appl.*, 2023, 119771.
doi: 10.1016/j.eswa.2023.119771
63. Keerthivasan, K. & Surendiran, B. Dimensionality reduction using PCA for network intrusion detection. *Perspect. Sci.*, 2016, **8**, 510-512.
doi: 10.1016/j.pisc.2016.05.010
64. Long, C.; Wei, J.; Xiao, J. & Zhao, J. Autoencoder ensembles for network intrusion detection. *In* 24th International Conference ICACT, Korea, 2022, 323-333.
doi: 10.23919/ICACT53585.2022.9728934
65. Lopes, I. O.; Zou, D.; Ruambo, F. A.; Akbar, S. & Yuan, B. Towards effective detection of recent DDoS attacks: A deep learning approach. *Secu. Commu. Network*., 2021, **14**.
doi: 10.1155/2021/5710028
66. Chattamvelli, R. Correlation in engineering and the applied sciences, 2024, Springer.

CONTRIBUTORS

Mr Rohan Gondhalekar is a research scholar in Department of Mathematics at Vellore Institute of Technology, Chennai, Tamil Nadu. He is currently working as a Research Associate on developing a mathematical model for network intrusions under a DRDO funded project (CAIR/CARS-91). His research interests are in network intrusion detection, machine learning, and dimensionality reduction techniques.
In the current study, he conceived the idea of the paper and was involved in the literature review and initial write-up.

Dr Rajan Chattamvelli is an Associate Professor in Department of Computer Science, Amrita Vishwa Vidyapeedam, Amaravati. His primary research interests are in algorithms, machine learning, data mining, cybersecurity, data analytics and engineering statistics.
In the current study, he contributed to the technical analysis of the work, testing the results and participated in the revision and improvement.