DESENVOLVIMENTO DE SOFTWARE PARA DETECÇÃO INTELIGENTE DE INTRUSÕES EM REDES COMPUTACIONAIS UTILIZANDO TÉCNICAS DE APRENDIZADO ADVERSARIAL

Any Vitória Dantas Peviani, Luis Filipe Karatanasov Beloni, Wagner João da Silva Mariano.

Centro Universitário UFBRA, Av. Dr. Nélson d'Ávila, 1202, Centro - 12245-031 - São José dos Campos - SP, Brasil, any.peviani26@gmail.com.

Resumo
Este trabalho desenvolveu um sistema de detecção de intrusões baseado em Random Forest e avaliou sua vulnerabilidade diante de perturbações adversariais transferidas em cenário de caixa preta. Utilizou-se o conjunto CIC-IDS2017 com 200.000 amostras estratificadas em 12 classes e padronização Z-score. Uma rede Perceptron Multicamadas (MLP) balanceada via SMOTE serviu como modelo substituto para gerar ataques FGSM transferíveis. Em condições nominais, o modelo alcançou 99,77% de acurácia e 87,74% de F1-score macro, caindo para 86,27% e 18,49%, respectivamente, sob ataque com $\epsilon = 0{,}05$. O retreinamento adversarial com a incorporação de 140.000 amostras perturbadas restabeleceu a resiliência do modelo (99,90% de acurácia e 96,51% de F1-score), demonstrando mitigação eficaz do ataque, embora a robustez permaneça vinculada à magnitude da perturbação adotada no treinamento. Conclui-se que o treinamento adversarial se apresenta como uma estratégia fundamental para assegurar a resiliência operacional e a confiabilidade de sistemas de segurança cibernética em cenários hostis.
Palavras-chave: Detecção de intrusões. CIC-IDS2017. Aprendizado adversarial. Cibersegurança. Random Forest.

Área do Conhecimento: Engenharia de Computação

Introdução

O crescimento das redes computacionais e a dependência de serviços digitais ampliaram a necessidade de mecanismos capazes de identificar atividades maliciosas de forma automatizada. Nesse contexto, os sistemas de detecção de intrusões (Intrusion Detection Systems — IDS) desempenham papel relevante no monitoramento do tráfego e na identificação de comportamentos que podem indicar comprometimento ou uso indevido de recursos. Abordagens baseadas em aprendizado de máquina têm sido empregadas para reconhecer padrões complexos em grandes conjuntos de características (Ring et al., 2019; Ferrag et al., 2020).

Entre os algoritmos utilizados nessa tarefa, o Random Forest apresenta características adequadas ao tratamento de dados tabulares e de alta dimensionalidade, por combinar múltiplas árvores de decisão construídas a partir de amostras e subconjuntos de características distintos (Breiman, 2001). Trabalhos baseados em aprendizado profundo também demonstram a aplicação de redes neurais à detecção de intrusões, evidenciando a diversidade de modelos utilizados na área (Shone et al., 2018; Vinayakumar et al., 2019).

Apesar do desempenho obtido em condições convencionais, classificadores de aprendizado de máquina podem ser vulneráveis a exemplos adversariais, nos quais pequenas alterações nas entradas são construídas para induzir decisões incorretas. O FGSM utiliza o gradiente da função de perda para determinar a direção da perturbação (Goodfellow, Shlens e Szegedy, 2015). Quando o classificador-alvo não disponibiliza diretamente esse gradiente, a transferência de exemplos adversariais permite gerar perturbações em um modelo substituto e avaliá-las em outro classificador (Papernot et al., 2017). Estratégias de treinamento adversarial incorporam exemplos perturbados ao treinamento com o objetivo de reduzir a degradação causada por ataques (Martins et al., 2020).

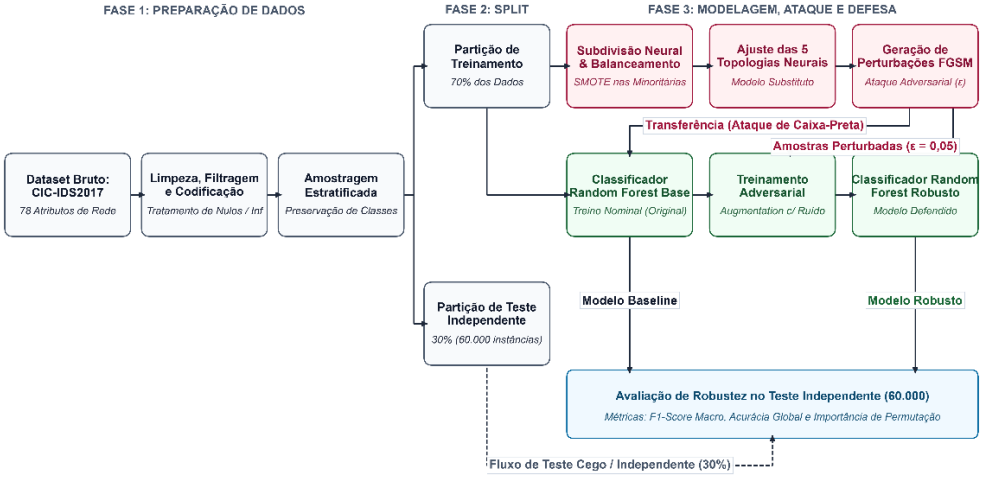
Diante desse contexto, este trabalho teve como objetivo desenvolver e avaliar um sistema de detecção de intrusões utilizando o Random Forest como classificador principal, investigar sua vulnerabilidade diante de ataques FGSM transferidos a partir de uma rede neural auxiliar e analisar o treinamento adversarial como estratégia de mitigação.

Metodologia

O estudo foi conduzido como investigação experimental quantitativa. O dataset CIC-IDS2017 (Sharafaldin, Lashkari e Ghorbani, 2018), originalmente composto por 2.830.743 registros e 79 colunas, foi submetido à remoção das classes com menos de 100 registros — Heartbleed (11), Web Attack Sql Injection (21) e Infiltration (36) — e de 2.867 observações contendo valores NaN ou infinitos. Após a limpeza, permaneceram 2.827.808 fluxos, 78 características numéricas e 12 classes. Em seguida, foi extraída uma amostra estratificada de 200.000 registros, com semente 42.

A amostra foi dividida em 70% para treinamento (140.000 instâncias) e 30% para teste independente (60.000 instâncias). As 78 características foram padronizadas por Z-score, ajustando média e desvio-padrão exclusivamente no treinamento. Para a rede neural auxiliar, os 140.000 registros foram subdivididos em 112.000 para ajuste e 28.000 para validação. O SMOTE (Chawla et al., 2002), com k = 3, foi aplicado somente aos 112.000 registros de treinamento neural, equalizando as 12 classes em 89.959 exemplos por classe e totalizando 1.079.508 registros. A arquitetura metodológica adotada no estudo, desde o pré-processamento do dataset até a avaliação do Random Forest sob perturbações adversariais e o treinamento de um modelo robusto, é apresentada na Figura 1.

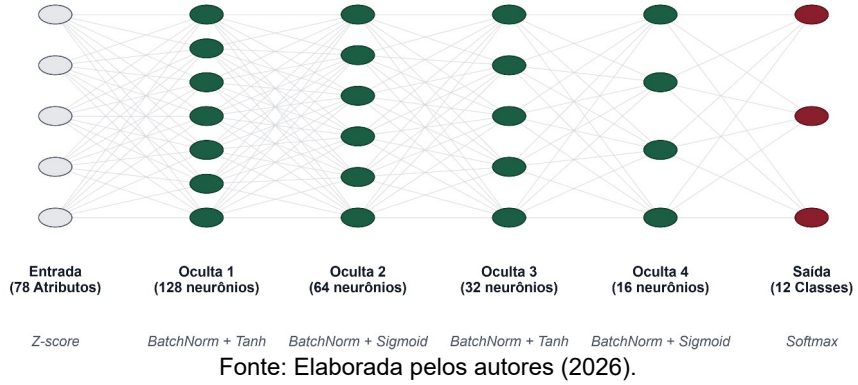
Figura 1 – Fluxograma da arquitetura metodológica e pipeline de defesa adversarial.



Fonte: Elaborada pelos autores (2026).

O Random Forest foi definido como classificador principal, com 100 árvores, critério de impureza de Gini e semente 42, sendo treinado diretamente sobre as 140.000 instâncias originais. Como árvores de decisão não fornecem diretamente um gradiente de entrada aplicável ao FGSM, foi empregada uma MLP como modelo substituto diferenciável. A MLP e o Random Forest foram ajustados para a mesma tarefa e para o mesmo espaço de atributos; a MLP não foi obtida por destilação das decisões do Random Forest nem por consultas iterativas ao classificador-alvo. A arquitetura da rede neural selecionada para atuar como modelo substituto é apresentada na Figura 2.

Figura 2 – Arquitetura da rede neural MLP substituta selecionada (Topologia 04).



Fonte: Elaborada pelos autores (2026).

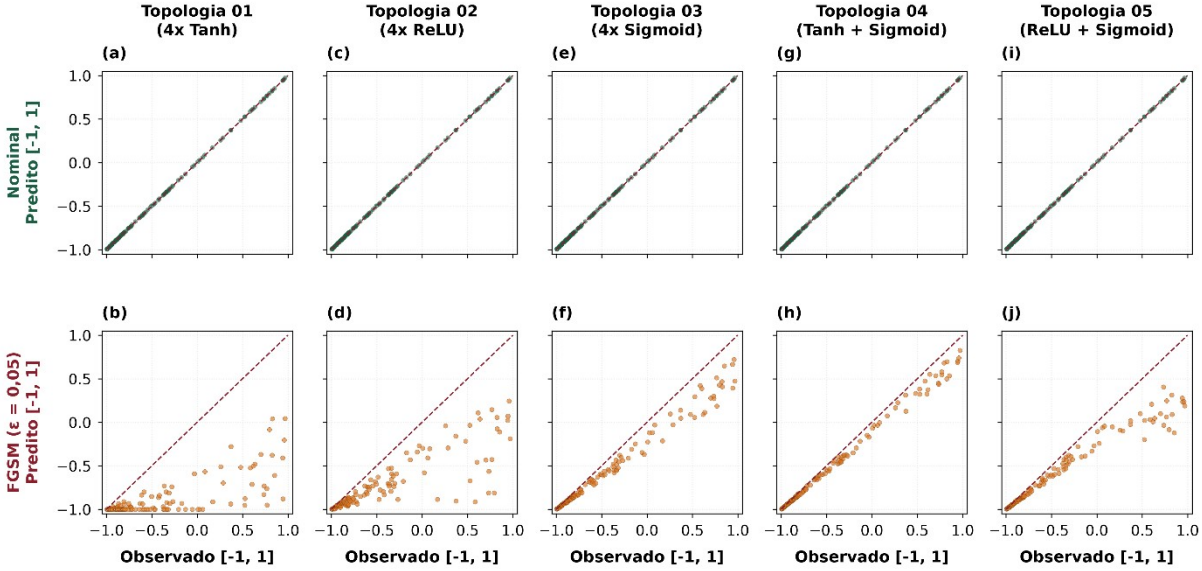
Foram avaliadas cinco topologias MLP com 78 entradas, quatro camadas densas na progressão [128, 64, 32, 16], Batch Normalization e saída softmax com 12 neurônios. O treinamento utilizou 35 épocas, lote de 128, SGD com momento 0,90, taxa de aprendizado 0,01 e perda de entropia cruzada categórica. A Topologia 04 foi selecionada com base no conjunto de métricas de validação, apresentando 98,28% de acurácia, 72,52% de precisão macro, 90,54% de recall macro e 76,77% de F1-score macro.

As perturbações foram geradas pelo FGSM a partir do gradiente da perda da Topologia 04 e aplicadas no espaço padronizado. Foram avaliadas as magnitudes ε = 0,01; 0,03; 0,05; 0,075; 0,10; 0,125 e 0,15. Para o treinamento adversarial, foram incorporadas ao treinamento original 140.000 amostras perturbadas com ε = 0,05, formando 280.000 instâncias no Random Forest robusto. A avaliação foi realizada no mesmo conjunto independente de 60.000 instâncias.

Resultados

A avaliação das cinco topologias neurais indicou diferenças de desempenho no conjunto de validação. A Topologia 04 apresentou o maior F1-score macro (76,77%), a menor perda de validação e o menor erro quadrático médio bipolar (0,922122), sendo selecionada exclusivamente como modelo substituto para a geração das perturbações. A comparação visual das predições obtidas pelas cinco topologias, nos cenários nominal e sob perturbação FGSM, é apresentada na Figura 3. Portanto, os resultados da MLP não representam o classificador final do sistema, mas a etapa auxiliar necessária para viabilizar a geração de ataques transferíveis.

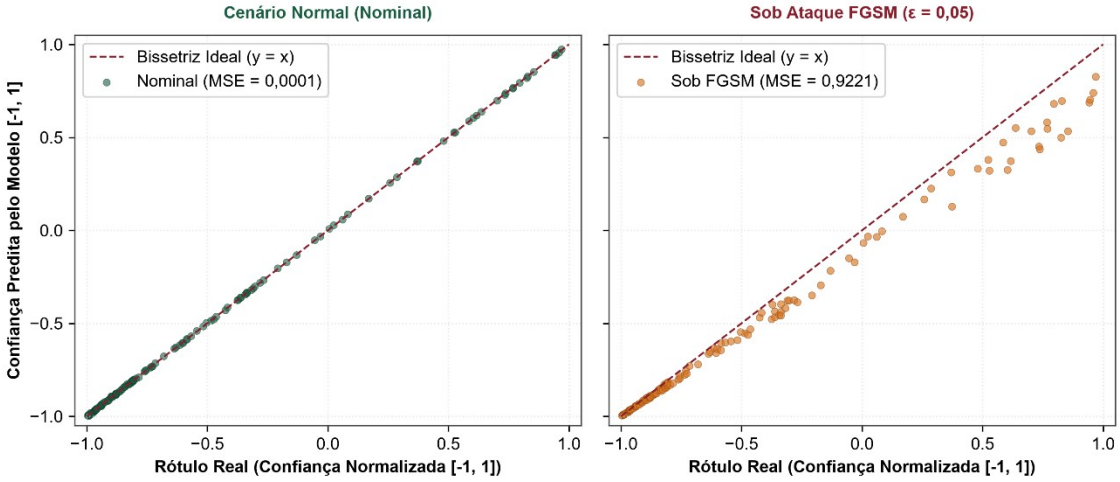
Figura 3 – Dispersão das predições das cinco topologias MLP nos cenários nominal e sob ataque FGSM.



Fonte: Elaborada pelos autores (2026).

A dispersão das predições da Topologia 04 nos cenários nominal e sob ataque FGSM com ε = 0,05 é apresentada na Figura 4, permitindo visualizar o efeito da perturbação sobre as predições do modelo substituto.

Figura 4 – Dispersão da Topologia 04 selecionada nos cenários nominal e sob ataque FGSM (ε = 0,05).

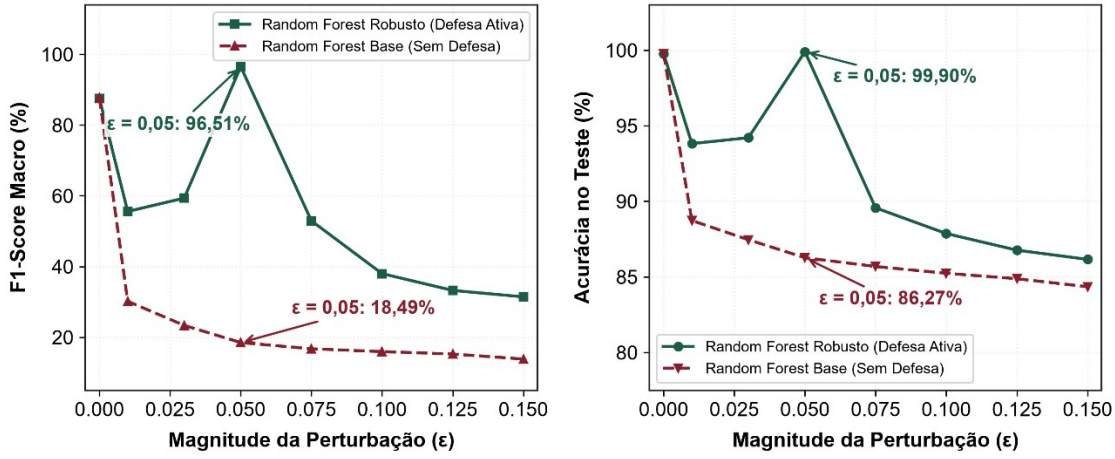


Fonte: Elaborada pelos autores (2026).

No teste independente, o Random Forest base apresentou 99,77% de acurácia e 87,74% de F1-score macro em condições nominais. Sob ataque FGSM transferido com ε = 0,05, a acurácia caiu para 86,27% e o F1-score macro para 18,49%. Após o treinamento adversarial com exemplos gerados nessa mesma magnitude, o Random Forest robusto alcançou 99,90% de acurácia e 96,51% de F1-score macro sob ε = 0,05.

As curvas de degradação da acurácia global e do F1-score macro são apresentadas na Figura 5. A comparação permite analisar a sensibilidade do Random Forest base e do modelo submetido ao treinamento adversarial à variação da magnitude da perturbação FGSM. Observa-se que a recuperação de desempenho do modelo robusto é mais evidente em ε = 0,05, magnitude utilizada na geração dos exemplos adversariais empregados no treinamento, enquanto o aumento de ε evidencia a limitação de generalização da defesa.

Figura 5 – Curvas de degradação da acurácia global e do F1-score macro em função da magnitude da perturbação FGSM.



Fonte: Elaborada pelos autores (2026).

Discussão

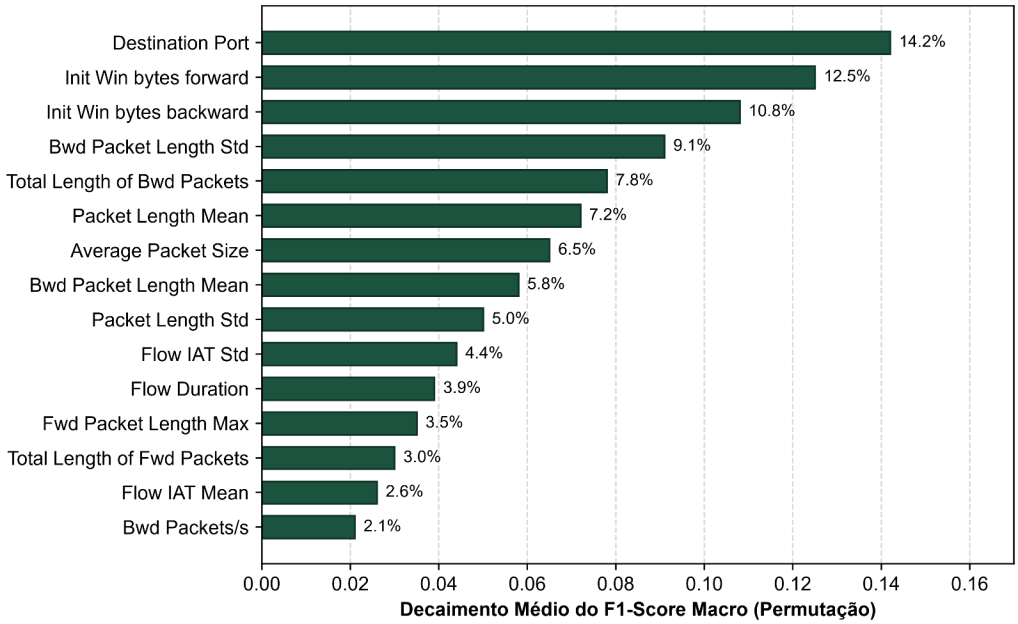
Os resultados mostram que o Random Forest apresentou elevada capacidade de classificação em condições nominais, mas sofreu degradação expressiva quando recebeu perturbações geradas por uma rede neural substituta. A redução de 87,74% para 18,49% no F1-score macro em ε = 0,05

evidencia que a transferência de perturbações pode afetar um classificador baseado em árvores mesmo sem acesso direto aos seus parâmetros internos.

A função da MLP deve ser distinguida da função do classificador principal. A seleção da Topologia 04 não significa que a rede neural tenha substituído o Random Forest; ela foi escolhida apenas como modelo substituto diferenciável para calcular os gradientes utilizados pelo FGSM. O desempenho final de robustez apresentado neste estudo, portanto, refere-se ao Random Forest submetido ao treinamento adversarial.

A análise de importância por permutação apontou Destination Port, Init Win bytes forward, Init Win bytes backward e Bwd Packet Length Std entre os atributos mais relevantes. A distribuição das quinze características com maior importância segundo essa análise é apresentada na Figura 6. Esses resultados indicam associação entre essas variáveis e o desempenho classificatório, mas não estabelecem causalidade individual nem permitem afirmar que tais atributos, isoladamente, determinem a evasão.

Figura 6 – Quinze atributos com maior importância segundo a análise por permutação.



Fonte: Elaborada pelos autores (2026).

Conclusão

Os resultados obtidos atendem ao objetivo de avaliar a vulnerabilidade e a mitigação adversarial de um sistema de detecção de intrusões cujo classificador principal é o Random Forest. Em condições nominais, o modelo atingiu 99,77% de acurácia e 87,74% de F1-score macro, enquanto o ataque FGSM transferido com ε = 0,05 reduziu esses valores para 86,27% e 18,49%, respectivamente.

A Topologia 04 da MLP foi selecionada como modelo substituto auxiliar, e não como classificador final, por apresentar o melhor desempenho entre as cinco arquiteturas avaliadas na etapa de geração das perturbações. Após o treinamento adversarial do Random Forest com exemplos gerados em ε = 0,05, o modelo robusto alcançou 99,90% de acurácia e 96,51% de F1-score macro nessa mesma magnitude. Esses resultados indicam recuperação substancial do desempenho na condição utilizada para o treinamento adversarial, mas não devem ser generalizados para magnitudes superiores de perturbação.

Como limitações, destacam-se o uso exclusivo do FGSM em cenário de transferência, a aplicação das perturbações no espaço estatístico padronizado das características e o fato de o experimento não reproduzir a manipulação de pacotes em tráfego de rede físico. Além disso, o treinamento adversarial exige a expansão do conjunto de treinamento e, consequentemente, maior custo computacional.

Trabalhos futuros podem investigar ataques iterativos, como PGD, restrições funcionais relacionadas aos protocolos de rede e estratégias adaptativas de treinamento para cenários de monitoramento contínuo.

Referências

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002.

FERRAG, M. A.; MAGLARAS, L.; MOSCHOYIANNIS, S.; JANICKE, H. Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. **Journal of Information Security and Applications**, v. 50, p. 102419, 2020.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. In: **INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR)**, 3., 2015, San Diego. Anais [...]. San Diego: ICLR, 2015. p. 1-11.

MARTINS, N.; CRUZ, J. M.; CRUZ, T.; ABREU, P. H. Adversarial machine learning applied to intrusion and malware scenarios: a systematic review. IEEE Access, v. 8, p. 35403-35419, 2020.

PAPERNOT, N.; McDANIEL, P.; GOODFELLOW, I.; JHA, S.; CELIK, Z. B.; SWAMI, A. Practical black-box attacks against machine learning. In: ASIA CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY (ASIA CCS), 2017, Abu Dhabi. Proceedings [...]. New York: ACM, 2017. p. 506–519.

RING, M.; WUNDERLICH, S.; SCHEPERD, D.; LANDES, D.; HOTHORN, A. A survey of network-based intrusion detection data sets. **Computers & Security**, v. 86, p. 147-167, 2019.

SHARAFALDIN, I.; LASHKARI, A. H.; GHORBANI, A. A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: **INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS SECURITY AND PRIVACY (ICISSP)**, 4., 2018, Funchal. Anais [...]. Funchal: SCITEPRESS, 2018. p. 108-116.

SHONE, N.; NGOC, T. N.; PHAI, V. D.; SHI, Q. A deep learning approach to network intrusion detection. **IEEE Transactions on Emerging Topics in Computational Intelligence**, v. 2, n. 1, p. 41-50, 2018.

VINAYAKUMAR, R.; ALAZAB, M.; SOWMYA, V. B.; POORNACHANDRAN, P.; AL-NUMAY, M. Deep learning approach for intelligent intrusion detection system. **IEEE Access**, v. 7, p. 41525-41550, 2019.