Journal of Database Management
Volume 35 • Issue 1

Intrusion Detection System:
A Comparative Study of Machine Learning-Based IDS

Amit Singh, Government of India, India
https://orcid.org/0000-0001-7351-0050

Jay Prakash, Vijay Singh Pathik Government (PG) College, India
https://orcid.org/0000-0002-6167-2412

Gaurav Kumar, Bennett University, India

Praphula Kumar Jain, GLA University, India
https://orcid.org/0000-0001-7651-4444

Loknath Sai Ambati, Oklahoma City University, USA*

ABSTRACT

The use of encrypted data, the diversity of new protocols, and the surge in the number of malicious activities worldwide have posed new challenges for intrusion detection systems (IDS). In this scenario, existing signature-based IDS are not performing well. Various researchers have proposed machine learning-based IDS to detect unknown malicious activities based on behaviour patterns. Results have shown that machine learning-based IDS perform better than signature-based IDS (SIDS) in identifying new malicious activities in the communication network. In this paper, the authors have analyzed the IDS dataset that contains the most current common attacks and evaluated the performance of network intrusion detection systems by adopting two data resampling techniques and 10 machine learning classifiers. It has been observed that the top three IDS models—KNeighbors, XGBoost, and AdaBoost—outperform binary-class classification with 99.49%, 99.14%, and 98.75% accuracy, and XGBoost, KNneighbors, and GaussianNB outperform in multi-class classification with 99.30%, 98.88%, and 96.66% accuracy.

KEYWORDS

Anomaly-Based Intrusion Detection Systems, Cyberattack, Cybersecurity, Intrusion Detection Systems, Machine Learning

1. INTRODUCTION

Because of the Covid-19 pandemic, individuals stayed at home and avoided physical gatherings, and social separation has become the new normal. The usage of new paradigms in corporate transactions, work-from-home culture, and online educational delivery has increased people's reliance on mobile and electronic devices. The use of communication networks and cloud-based processing systems

DOI: 10.4018/JDM.338276

*Corresponding Author

have increased manifold. This change in the pandemic era promotes new threats and lures intruders to exploit vulnerabilities in the data communication network. Organizations usually use diversified protocols to encrypt their data and maintain confidentiality. Volume, heterogeneity of protocols, and encryption have posed several new challenges before the IDS system in detecting malicious activities (Resende & Drummond, 2018; Senthilkumar et al., 2021). An intruder attempts to gain unauthorized access to a system or network with malafide intentions and disrupt the normal execution (Butun et al., 2014; Liao et al., 2013; Low, 2005; Mitchell & Chen, 2014). Several times intruders aim to steal or corrupt sensitive data. In 2020, Emsisoft reported that local governments, universities, and private organizations had spent $144 million in response to the worst ransomware attack (Novinson, 2020). The WHO reported that cyber-attack increased five-fold during the Covid-19 pandemic (WHO, 2020). According to the McAfee quarterly threat report 2020, fraudsters are taking advantage of the pandemic by using Covid-19-themed malicious apps, phishing campaigns, and malware (McAfee, 2020). The report also highlights that in quarter one (Q1), new malware targeting mobile devices surged by 71%, with overall malware increasing by roughly 12% over the previous four quarters (McAfee, 2020).

IDS provides security solutions against malicious attacks or security breaches. It can be a software or hardware device that detects harmful activity to maintain system security (Babu et al., 2023; Liao et al., 2013). It identifies all forms of suspicious network traffic and malicious computer activity that a firewall might miss. Signature-based Intrusion Detection Systems (SIDS) and Anomaly-based Intrusion Detection Systems (AIDS) are two popular categories of IDS that have widely been used to provide security solutions (Axelsson, 2000; Baskerville & Portougal, 2003; Hodo et al., 2017). The SIDS relies on previously known signatures and faces challenges in identifying an unknown and obfuscated malicious attack (Amouri et al., 2020; Atli, 2017; Khraisat et al., 2019; Lin et al., 2015; Low, 2005; Vinayakumar et al., 2019; Wu & Banzhaf, 2010). Therefore, SIDS cannot prevent every intruder based on previously learned indicators of compromises; however, they can detect and prevent similar attacks from happening in the future. As the number of cyber-attacks has increased exponentially and attackers are using evolved techniques to conceal attack patterns, it becomes almost infeasible to identify intruders using SIDS (Amouri et al., 2020; Khraisat et al., 2019; Vimala et al., 2019; Warsi & Dubey, 2019; Wu & Banzhaf, 2010).

Many scholars use AIDS because of its ability to overcome the limitation of SIDS. An AIDS is a typical computer system model created using statistical-based methods, machine learning algorithms, or knowledge-based methods. These methods are designed and developed to detect abnormal behaviour in computer systems. The typical usage pattern is base-lined, and alarms are generated when usage deviates from the expected behaviour. The key benefit of using AIDS is detecting zero-day attacks because it does not rely on a signature database to detect abnormal user behaviour (Alazab et al., 2012; Laughlin et al., 2020). AIDS is further categorized into three main groups: Statistics-based, Knowledge-based, and Machine learning-based. Researchers have investigated many approaches to improve intrusion detection in the last few decades, from data mining and machine learning to time series modelling. The Machine learning-based IDS can learn the attacks' behaviour and pattern, and future attacks can be predicted using trained machine learning models.

Machine Learning is a technique for extracting knowledge from massive amounts of data. It comprises a set of rules, methods, or complex "transfer functions" that can be used to discover intriguing patterns or estimate behaviour in a wide range of applications (Abu Al-Haija et al., 2022; Choudhury et al., 2023; Dua & Du, 2016; Mangal et al., 2023; Prasad Yadav et al., 2023; Sinha & Sharma, 2021). The machine learning techniques use training data to acquire complex pattern-matching capabilities. Researchers (Hamzah & Othman, 2021; Hasan et al., 2016; Mehmood et al., 2021; Niyaz et al., 2015; Shams & Rizaner, 2018) widely use the Support Vector Machine (SVM) for Network Intrusion Detection Systems (NIDS) and different clustering algorithms such as K-means and Exception Maximization (EM) for both NIDS and anomaly detection (Bennett & Demiriz, 1999; Laughlin et al., 2020; Maseer et al., 2021; Syarif et al., 2012; Wazid & Das, 2016). They are mainly concerned with the detection effect and lack practical issues such as detection