Review Article

Journal of Material Sciences & Applied Engineering

Advancing Intrusion Detection Systems: Mitigating Model Bias and Data Imbalance with Machine Learning Approaches

Qazi Waqas Khan*

Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea

***Corresponding author:** Qazi Waqas Khan, Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea.

Submitted: 22 July 2024 **Accepted:** 27 July 2024 **Published:** 30 July 2024

Citation: *Qazi Waqas Khan (2024) Advancing Intrusion Detection Systems: Mitigating Model Bias and Data Imbalance with Machine Learning Approaches J Mat Sci Apl Eng 3(4), 01-03.*

Abstract

Intrusion Detection Systems (IDSs) are essential for securing highly confidential data and protecting network architecture from cyber-attacks. Despite their high accuracy, traditional IDS methods often face significant challenges, such as model bias due to data imbalances and irrelevant features. This study proposes the state-of-the-art machine learning (ML) based IDS that addresses these challenges. By minimizing misclassification errors and correcting model bias, this proposed IDS significantly enhances predictive accuracy and generalizability, thereby offering a promising solution to the current limitations of IDS technologies. This study used the Decision Tree, Xtreme Gradient Boosting, and Adaboost model to classify an attack. The experimental results demonstrate the robustness of a XGB model for the classification of an attack.

Introduction

Over the past 20 years, digital offerings have gained popularity due to technological advancements, particularly during the COVID-19 pandemic. People most commonly use smartphones, tablets, laptops, and other electronic devices to utilize these services from anywhere at any time. Consequently, data that may contain highly confidential information starts flowing through networks between devices and data centers. This situation gives attackers a new chance to breach security barriers and conduct extensive attacks that are dangerous for individuals and organizations. Security flaws in the system are targeted by attackers using innovative strategies, which can lead to client account breaches, illegal access to the system, or the improper use of information. It has become a critical concern for researchers and scientists to defend against these attacks and protect sensitive data and networks from external attacks. IDS has become one of the most well-known and widely used mechanisms in response to these challenges. It examines incoming traffic and classifies it as malicious or legal to detect potential risks in a particular system or network. Nowadays, an IDS is for protecting a network or system against potential threats. In the past, many IDS systems were developed over the last 20 years to detect and protect against potential attacks. These existing methods need more flexibility and scalability to make them vulnerable to threats. This study proposed a method to detect attacks in networks [1-9].

Models

This study classified an attack using the AdaBoost, Decision Tree, and Xtreme Gradient Boosting models. It resampled the attack classes using the SMOTE data resampling method and utilized the Mutual Information Feature selection to select the features.

Ada Boost

AdaBoost or Adaptive Boosting, creates a robust model by merging several weak models into a single model. This model focuses on the errors made by previous models and fixes them in next iteration. This process improves the model's accuracy for classification [10].

Decision Tree (DT)

A Decision Tree is a supervised machine-learning model that creates a tree-like structure of decisions and their possible outcomes by dividing the data into branches according to feature values. It makes the final prediction according to the generated rules [11].

Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting, is an effective machine learning algorithm known for its high performance in classification. This model improves the accuracy by combining weak learners into a powerful predictive model using gradient-boosting techniques. XGB employs regularization, parallel processing, and methods for handling missing data, which makes it effective for complex and large-scale datasets [12].

Table 1: Experimental Results ML Model without Feature Selection

Method	Accuracy	Precision	Recall	F score
XGB	0.820668	0.852144	0.820668	0.802181
DT	0.786302	0.789646	0.786302	0.771764
AdaBoost	0.81486	0.815434	0.81486	0.806822

Table 1 describes the experimental findings of XGB, DT and AdaBoost for the model without feature selection. XGB achieves a higher performance metrics score among all other proposed classifiers, demonstrating its efficiency and reliability; AdaBoost follows XGB, describing good overall performance. The DT has the lowest metrics score, showing its least effectiveness.

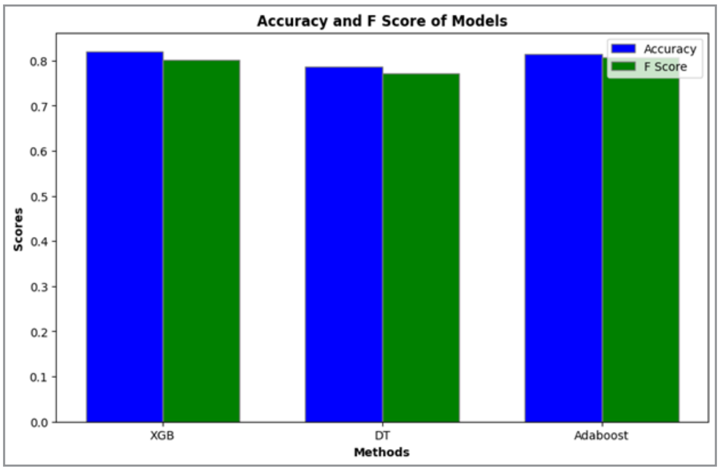

Figure 1: Without feature selection

Figure 1 represents the bar chart showing the Accuracy and F-score for the proposed models. XGB demonstrates high performance, with AdaBoost following behind and DT having the lowest performance score among all proposed classifiers.

Table 2: Experimental Results ML Model with MI Feature Selection

Method	Accuracy	Precision	Recall	F score
XGB	0.890751	0.894767	0.890751	0.887556
DT	0.860925	0.869545	0.860925	0.854433
AdaBoost	0.887735	0.903235	0.887735	0.881867

Table 2 evaluates the proposed models with MI feature selection. Overall, all the models show improved performance rates with feature selection. XGB outperforms with an accuracy of 89.07%, showing its strong performance. AdaBoost also performs well, especially in precision at 90.32% whereas DT has a lower score across all metrics.

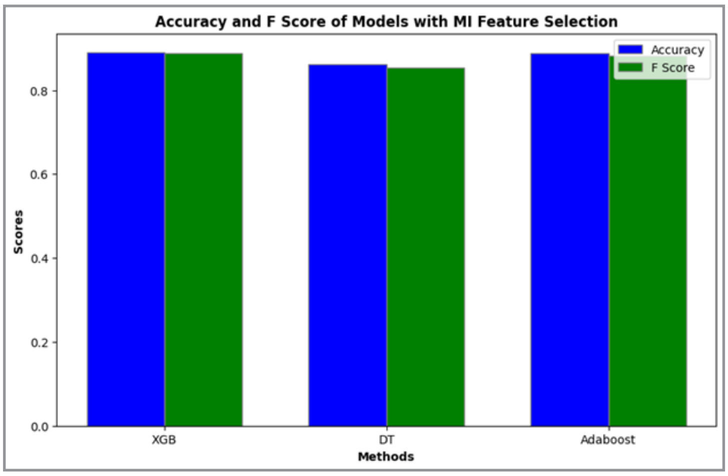

Figure 2: With feature selection

Figure 2 compares three proposed classifiers based on Accuracy and F-score with MI feature selection. As shown, XGB achieves the highest Accuracy and F-score, demonstrating the model's efficiency. AdaBoost also shows excellent performance but is slightly behind XGB, while DT has the lowest score, which shows it is less effective than other classifiers.

Conclusion

IDS plays a vital role in detecting potential security threats within network requirements. This study examines the performance of proposed classifiers with and without feature selection. XGB outshines, demonstrating its efficacy. AdaBoost also shows good performance, while DT shows less effectiveness. Feature selection boosts the model's performance, with XGB consistently showing robustness.

References

1. Mudassar Khan, Nohman Khan, Samina Begum, Muhammad Imran Qureshi (2024) Digital future beyond pandemic outbreak: systematic review of the impact of COVID-19 outbreak on digital psychology. foresight 26: 1-17.
2. McHaney Roger (2011) The new digital shoreline: How Web 2.0 and millennials are revolutionizing higher education. Taylor & Francis 267.
3. Arogundade Oluwasanmi Richard (2023) Network security concepts, dangers, and defense best practical. Computer Engineering and Intelligent Systems 14: 25-38.
4. Ömer Aslan, Semih Serkant Aktuğ, Merve Ozkan-Okay, Abdullah Asim Yilmaz, Erdal Akin (2023) A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. Electronics 12: 1-42.
5. Riggs Hugo, Shahid Tufail, Imtiaz Parvez, Mohd Tariq, Mohammed Aquib Khan, et al. (2023) Impact, vulnerabilities, and mitigation strategies for cyber-secure critical infrastructure. Sensors 23: 4060.
6. Nair Sunil Sukumaran (2024) Securing Against Advanced Cyber Threats: A Comprehensive Guide to Phishing, XSS, and SQL Injection Defense. Journal of Computer Science and Technology Studies 6: 76-93.
7. Heidari, Arash, Mohammad Ali Jabraeil Jamali (2023) Internet of Things intrusion detection systems: a comprehensive review and future directions. Cluster Computing 26: 3753-3780.
8. Trisolino Andrea (2023) Analysis of Security Configuration for IDS/IPS. Diss. Politecnico di Torino.
9. Sood Keshav, Mohammad Reza Nosouhi, Dinh Duc Nha Nguyen, Frank Jiang, Morshed Chowdhury, et al. (2023) Intrusion detection scheme with dimensionality reduction in next generation networks. IEEE Transactions on Information Forensics and Security 18: 965-979.
10. Hornyák Olivér, László Barna Iantovics (2023) AdaBoost algorithm could lead to weak results for data with certain characteristics. Mathematics 11: 1801.
11. Costa Vinícius G, Carlos Pedreira E (2023) Recent advances in decision trees: An updated survey. Artificial Intelligence Review 56: 4765-4800.
12. Zeravan Arif Ali, Ziyad H Abduljabbar, Hanan Tahir, Amira Bibo Sallow, Saman Almufti M (2023) Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review. Academic Journal of Nawroz University 12: 320-334.