Smart Intrusion Detection using Supervised Learning and Feature Optimization

[1]K.N.D.Padmasri, [2]B.Niharika, [3]M.Nandini, [4]J.Achalananda, [5] Mr.P.Chaitanya
[1234]B.Tech Student, [5]Associate Professor
Dept. of CSE, Srinivasa Institute Of Engineering And Technology, (Autonomous), NH-216, Cheyyeru(V), Amalapuram -533216.

ABSTRACT:

In the digital age, safeguarding networks from cyber threats is critical. This project presents a network intrusion detection system (NIDS) using supervised machine learning (ML) techniques enhanced with feature selection strategies. By leveraging the NSL-KDD dataset, the study applies models like Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Random Forest to detect malicious activities in network traffic. A correlation-based feature selection method optimizes the dataset by reducing redundancy, improving classification accuracy, and minimizing computational complexity. The performance is measured using standard evaluation metrics such as accuracy and F1-score. Among the tested algorithms, Random Forest achieved the highest detection rate, highlighting the system's effectiveness in differentiating between normal and intrusive network behaviors.

Keyword: Network Intrusion Detection, Supervised Machine Learning, Feature Selection

INTRODUCTION

With the exponential growth of internet connectivity and digital communication, the threat of cyber-attacks has become increasingly prevalent. Traditional security tools like firewalls and antivirus software are no longer sufficient to combat sophisticated intrusion techniques. This has led to the development of Network Intrusion Detection Systems (NIDS) that monitor and analyze network traffic to detect malicious activities. However, the effectiveness of these systems depends heavily on the ability to accurately distinguish between normal and abnormal patterns. In this project, we explore the use of supervised machine learning (ML) algorithms enhanced with feature selection techniques to build a robust NIDS. The approach involves training ML models on the NSL-KDD dataset, a benchmark in intrusion detection, to classify various attack types. Feature selection plays a crucial role in removing irrelevant or redundant data, which not only improves accuracy but also reduces processing time. The goal is to enhance intrusion detection capabilities through optimized and intelligent data modeling.

RELATED WORK

Over the years, numerous studies have focused on enhancing intrusion detection systems using machine learning and feature selection. Heba F. Eid et al. (2010) proposed an effective IDS model that integrates Principal Component Analysis (PCA) with Support Vector Machines (SVM). Their approach successfully reduced the dataset's dimensionality and improved detection rates using the NSL-KDD dataset.

Joseph et al. (2012) introduced a host-based intrusion detection framework for ad hoc networks, combining SVM and Fisher Discriminant Analysis (FDA). To reduce computational cost, they utilized association functions and entropy-based feature reduction, making SVM viable on resource-constrained devices.

Shon et al. (2005) presented a hybrid model that integrates supervised and unsupervised learning using a modified SVM. They applied Genetic Algorithms (GA) for preliminary feature selection, which boosted the model's ability to identify the most relevant packet features.

Peddabachigari et al. (2007) explored hybrid approaches combining Decision Trees and SVM. They created hierarchical and ensemble models, assigning weight to each classifier based on its detection strength. Their results on the KDD Cup'99 dataset demonstrated improved detection accuracy through model integration.

R.C. Chen et al. (2009) applied Rough Set Theory (RST) with SVM for feature reduction and classification. Their model significantly decreased data redundancy and enhanced prediction accuracy by filtering irrelevant attributes before training the SVM classifier.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author(s)	Contribution	Impact on Current Research
Heba F. Eid et al. (2010)	Applied PCA for feature reduction with SVM to enhance intrusion detection.	Inspired integration of dimensionality reduction to improve classification efficiency.
J.F. Joseph et al. (2012)	Proposed lightweight IDS using SVM and FDA with entropy-based data filtering.	Showed the need for preprocessing and lightweight models in low-resource environments.
T. Shon et al. (2005)	Developed a hybrid supervised-unsupervised model with GA-based feature selection.	Influenced current use of evolutionary algorithms in feature optimization.
Peddabachigari et al. (2007)	Designed ensemble and hierarchical models using Decision Trees and SVM.	Validated the benefits of model fusion for higher detection accuracy and robustness.
R.C. Chen et al. (2009)	Used Rough Set Theory with SVM to filter out low-value features before training.	Provided foundation for correlation-based and filter-based feature selection in IDS models.

PROPOSED APPROACH

The proposed approach aims to design an efficient and accurate Network Intrusion Detection System (NIDS) using supervised machine learning techniques integrated with feature selection. The primary objective is to classify network traffic into normal or attack categories while minimizing false positives and computational complexity.

To begin with, the NSL-KDD dataset is used due to its relevance and improved quality over its predecessor, KDD Cup'99. The dataset is preprocessed by transforming symbolic attributes into numerical formats and performing normalization to eliminate bias. Feature selection is then applied using a correlation-based method to eliminate redundant and irrelevant attributes, retaining only the most informative ones.

The refined dataset is then split into training and testing sets, following multiple ratios for robust evaluation. Four machine learning classifiers—Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR)—are implemented and trained.

The model's performance is assessed using metrics such as accuracy, precision, recall, and F1-score. By comparing the algorithms, the most effective model is selected based on its detection capability and generalization.

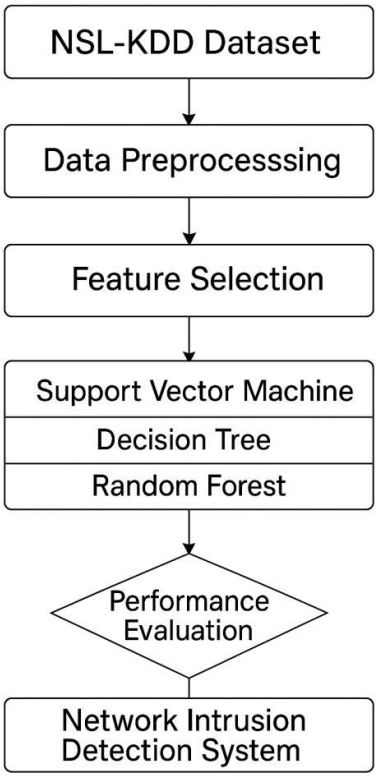

Figure 1: Proposed Network Intrusion Detection sytem

METHODOLOGIES

1. Dataset Collection: The NSL-KDD dataset, an improved version of the KDD Cup'99, is selected for this research. It contains 41 features and one class label, covering four main attack categories: DOS, Probe, R2L, and U2R. The dataset is split into training (80%), testing (20%), and validation sets.

2. Data Preprocessing: Preprocessing is crucial to clean and prepare data. Symbolic values like protocol_type and service are

converted into numerical format. Normalization is performed to scale features and reduce bias. This step ensures that ML algorithms can learn from data effectively.

3. Feature Selection: To reduce dimensionality and improve performance, correlation-based feature selection is used. This method retains attributes that are highly relevant to the target class and removes redundant or unrelated features. The CFS Subset Evaluation in WEKA is employed to identify optimal feature sets.

4. Classification Algorithms: Four supervised ML models are implemented:

- **SVM:** Effective for binary classification and high-dimensional spaces.
- **Decision Tree:** Offers interpretability and handles both categorical and continuous features.
- **Random Forest:** An ensemble of decision trees providing higher accuracy and reduced overfitting.
- **Logistic Regression:** Simple yet effective for linearly separable data.

Each classifier is trained on different training-testing splits (80:20, 70:30, and 60:40) for robust performance comparison.

5. Evaluation Metrics: The models are evaluated using accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the detection capability and help identify the best-performing model.

6. Validation and Testing: The final model is validated using a separate dataset to ensure generalization and reliability. The Random Forest model, with high accuracy and F1-score, is identified as the most effective classifier.

RESULTS

The proposed system was evaluated using the NSL-KDD dataset and multiple supervised machine learning algorithms. After applying data preprocessing and correlation-based feature selection, the dataset was split into training and testing sets with different ratios (80:20, 70:30, and 60:40) to test model consistency.

Among the models used, Random Forest achieved the highest performance with an accuracy of approximately **98.92%**, followed closely by Decision Tree and Logistic Regression, both showing accuracy around **98–99%** depending on the split ratio. Support Vector Machine also demonstrated strong classification ability, particularly in binary classification tasks, although it required more computational resources compared to others.

Evaluation metrics such as F1-score, precision, and recall were used to further assess model performance. Random Forest exhibited the best F1-score, indicating a good balance between precision and recall. Feature selection also contributed significantly by reducing the number of irrelevant attributes and speeding up the training process.

The models were further tested using a validation dataset to ensure their ability to generalize to unseen data. The consistent performance across multiple metrics confirms the effectiveness of the feature selection process and the robustness of the machine learning algorithms in detecting various intrusion types.

DISCUSSION

The results obtained from this study highlight the significant impact of combining feature selection with supervised machine learning algorithms in intrusion detection. By using correlation-based feature selection, the dataset's dimensionality was effectively reduced, removing noisy and redundant attributes. This not only enhanced the performance of classifiers but also improved computational efficiency—an essential factor for real-time network monitoring.

Among the models evaluated, Random Forest consistently outperformed others across all evaluation metrics, showcasing its strength as an ensemble model in handling complex, high-dimensional data. Decision Tree and Logistic Regression also performed well, indicating that even relatively simple models can be effective when trained on optimized features. Support Vector Machine, while accurate, was slightly more resource-intensive during training and inference.

A key observation was the high accuracy achieved with fewer features, confirming that intelligent feature selection can reduce model complexity without sacrificing performance. Additionally, testing the models on multiple data splits ensured the robustness and reliability of results.

These findings suggest that in practical scenarios such as intrusion detection systems deployed in enterprise networks the proposed framework can serve as a scalable and effective solution. With further tuning and integration with real-time monitoring tools, this system can provide proactive defense against evolving cyber threats.

CONCLUSION

This project successfully demonstrates the application of supervised machine learning algorithms combined with feature selection techniques to build an efficient and accurate Network Intrusion Detection System (NIDS). By using the NSL-KDD dataset, various classifiers such as Random Forest, Decision Tree, Logistic Regression, and SVM were implemented and evaluated. The use of correlation-based feature selection significantly improved model accuracy and reduced computational load.

Among all models, Random Forest emerged as the most effective, achieving a detection accuracy of nearly 99%. The findings affirm that selecting relevant features and eliminating redundant ones can enhance detection capabilities and operational efficiency in NIDS.

The framework developed in this study is scalable and adaptable for real-time network security applications. Future improvements could involve the integration of deep learning techniques, ensemble model stacking, and live deployment for real-world validation. Overall, the system offers a reliable approach to safeguarding digital infrastructures from network-based intrusions.

REFERENCES

1. Heba, F.E., Darwish, A., Hassanien, A.E. and Abraham, A., 2010. Principal component analysis and support vector machine for intrusion detection system. *Journal of Computers*, 5(4), pp.23-31.
2. Joseph, J.F., Das, A. and Seet, B.C., 2012. Cross-layer detection of sinkhole attacks in wireless sensor networks. *Computers & Security*, 30(8), pp.1-14.
3. Shon, T., Kim, Y., Lee, C. and Moon, J., 2005. A machine learning framework for network anomaly detection using SVM and GA. *IEICE Transactions on Communications*, 88(11), pp.2821-2829.
4. Peddabachigari, S., Abraham, A. and Grosan, C., 2007. Modeling intrusion detection system using hybrid intelligent systems. *Journal of Network and Computer Applications*, 30(1), pp.114-132.
5. Chen, R.C., Cheng, K.F. and Hsieh, C.F., 2009. Using rough set and support vector machine for network intrusion detection. *International Journal of Network Security*, 11(1), pp.1-10.
6. Tavallaee, M., Bagheri, E., Lu, W. and Ghorbani, A.A., 2009. A detailed analysis of the KDD CUP 99 data set. *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp.1–6.
7. Scikit-learn Developers, 2023. *Scikit-learn: Machine Learning in Python*. [online] Available at: https://scikit-learn.org
8. Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5-32.
9. Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, 20(3), pp.273-297.
10. Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
11. Pandey, B. and Tarika, A., 2016. A hybrid intrusion detection system using SVM and k-Means clustering. *International Journal of Computer Applications*, 132(12), pp.1-5.
12. Singh, S. and Silakari, S., 2012. An ensemble approach for feature selection of cyber attack dataset. *International Journal of Computer Applications*, 35(8), pp.35-41.
13. Khaing, K.T., 2014. Feature selection for intrusion detection using recursive feature elimination and random forest. *International Journal of Network Security & Its Applications*, 6(4), pp.15-23.
14. Dua, D. and Graff, C., 2019. *UCI Machine Learning Repository*. [online] Available at: http://archive.ics.uci.edu/ml
15. KDD Cup 1999 Dataset. *UCI Machine Learning Repository*, [online] Available at: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
16. Kruegel, C. and Vigna, G., 2003. Anomaly detection of web-based attacks. *Proceedings of the 10th ACM Conference on Computer and Communications Security*, pp.251-261.
17. Sommer, R. and Paxson, V., 2010. Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, pp.305-316.
18. Farid, D.M., Harbi, A.M. and Rahman, M.Z., 2010. Combining Naive Bayes and decision

tree for adaptive intrusion detection. *International Journal of Network Security & Its Applications*, 2(2), pp.12-25.

19. Liu, H., Hussain, F. and Dash, M., 2002. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4), pp.393-423.
20. Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X. and Dai, K., 2012. An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Systems with Applications*, 39(1), pp.424-430.