SVM-Based Anomaly Detection for Network Intrusion Systems: An Advanced Approach to Intrusion Prevention

Arihant Singhvi, Abhinav Nambiar, Pranav Menon, Narendra V G

Dept of Computer Science and Engineering , Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, 576104, India

Abstract.

This project develops an SVM-based IDS for anomaly detection, addressing limitations of signature-based methods in identifying new threats. Trained on labeled network traffic, it accurately distinguishes normal from anomalous patterns with minimal false positives. Key challenges like data quality, computational demands, and preprocessing complexity are tackled through careful data selection. The system is scalable, adaptive, and enhances network security by detecting emerging threats.

Keywords. Anomaly-Based Network Intrusion Detection System (A-NIDS), SVM, Naïve-bayes

1. INTRODUCTION

This project develops an SVM-based Anomaly Detection IDS to overcome the limitations of signature-based methods in detecting novel threats. Unlike traditional IDS, it identifies deviations from normal network behavior, enabling detection of zero-day attacks. SVM's robustness in high-dimensional data ensures precise classification of normal and anomalous traffic. Trained on labeled network data, the system adapts to evolving threats. Rigorous preprocessing enhances data quality, and performance evaluation confirms high accuracy with minimal false positives, providing a scalable and effective network security solution.

2. LITERATURE SURVEY

As cyber threats evolve, robust Intrusion Detection Systems (IDS) are crucial. Traditional signature-based methods fail against novel attacks, prompting the use of machine learning. SVM-based IDS map input data into high-dimensional space, outperforming Naïve Bayes [1]. Alternative models like RF, DT, and MLP have varying accuracy on KDD Cup '99 and NSL-KDD datasets [2][3][6]. High computational costs and redundant features challenge SVM-based IDS, mitigated by hybrid feature selection using K-means and IGR [4]. A pipeline-based approach enhances performance by filtering destructive data [5]. Kernel choice impacts accuracy, with RBF performing best (99.11%) on KDD Cup '99 [7]. However, outdated datasets should be replaced by modern alternatives like CICIDS2017, which includes DDoS, XSS, and SQL Injection. Deep learning-based IDS using UNSW-NB15 and CNNs have shown promising accuracy (95.4%-95.6%) [8].

3. METHODOLOGY

3.1. Dataset

We use the CSE-CIC-IDS 2018 dataset, which systematically generates profiles for normal and attack network behaviors. It contains 80 features and over 1 million entries, with no null values. Key features include destination port, protocol, flow duration, and packet counts. The dataset categorizes patterns as:

- **Normal:** Benign
- **Attack:** DoS-GoldenEye, DoS-Slowloris

3.2. Preprocessing

Dropped the ‘Timestamp’ column, converted ‘Label’ to binary, applied one-hot encoding, analyzed feature correlation, split data (33% train, 67% test), handled infinite values by imputing with the mean, and standardized features.

3.3. Hyperparameter Tuning of SVM Model

The SVM model is then developed using the Radial Basis Function (RBF) kernel, specifically Gaussian RBF, selected for its ability to handle the non-linear nature of network traffic data. The model is trained on the labeled dataset with a focus on accurately distinguishing between normal and anomalous traffic. To optimize the model, we employ cross-validation techniques to fine-tune hyperparameters such as the penalty parameter (C) and kernel coefficient (gamma).

The formula for RBF kernels between 2 feature vectors x and x' is given by: The equation should be submitted in the equation editor and should be keyed in as below in editable text:

$$A_0 = \frac{\mu_0 \omega I_0}{4\pi} = 10^{-7} \omega I_0 K(x, x') \; = \; e^{-\gamma \|x - x'\|^2}$$

where:

- $\|x - x'\|^2$ is the squared Euclidean distance between the two feature vectors x and x'.
- γ (gamma) is a parameter that determines the influence of each training example

Name	Abbreviation	Formula
Cubic	CU	$\varphi(r) = r^3$
Thin plate splines	TPS	$\varphi(r) = r^2 \log(r)$
Generalized Thin plate splines	GTPS	$\varphi(r) = r^{2m} \log(r), \quad m \in N$
Inverse quadrics(or Cauchy)	IQ	$\varphi(r) = \frac{1}{c^2 + r^2}$
Multiquadrics	MQ	$\varphi(r) = \sqrt{c^2 + r^2}$
Inverse Multiquadrics	IMQ	$\varphi(r) = \frac{1}{\sqrt{c^2 + r^2}}$
Gaussian RBF	GA	$\varphi(r) = e^{-r^2/c^2}$

Definition 2.2 *θ -method, (0 ≤ θ ≤ 1), is general finite-difference approximation to* $\frac{\partial^2 u(x,t)}{\partial x^2}$ *given by:*

$$\frac{\partial^2 u(x,t)}{\partial x^2} \cong \theta \delta_{2,x} U_{i,j+1} + (1 - \theta) \delta_{2,x} U_{i,j},$$

Figure 1. Different Types of RBF kernel functions

The RBF kernel enables linear separation in a higher dimension without explicit transformation. Regularization (C) balances misclassification and overfitting, while gamma controls the influence range, affecting sensitivity and generalization. GridSearchCV optimizes C and gamma using a parameter grid with an RBF kernel and 5-fold cross-validation, ensuring optimal hyperparameters for maximum accuracy. Efficiency metrics vary by system: spectral efficiency for wireless, joules per bit for electronics, and watts per Gbps for datacenters, with dBε as a universal measure of energy efficiency.

4. RESULTS AND OBSERVATIONS

The SVM-based IDS is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with particular emphasis on minimizing the false positive rate while maintaining high detection accuracy. After obtaining the best model by optimizing the hyperparameters, its performance is estimated against the testing set. The model has been found to achieve exceptional performance with an accuracy of 99.99% on the training set. The model shows similar results in precision, accuracy and recall as provided by the classification report.

4.1. Classification Metrics

The SVM-based IDS is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, prioritizing minimal false positives while maintaining high detection accuracy. The optimized model achieves **99.99% accuracy** on the training set and similar precision and recall.

Classification Metrics:

- **Normal (Class 0)** & **Attack (Class 1):** Precision, Recall, and F1-Score = **1.00**

- **Confusion Matrix:**
  - **True Positives:** 17,205
  - **True Negatives:** 328,787
  - **False Positives:** 29
  - **False Negatives:** 9

The **ROC-AUC** is **1.00**, indicating perfect classification.

Model Comparisons:

- **SVM (Polynomial Kernel):** 99.99% accuracy
- **Random Forest:** 100% accuracy
- **Logistic Regression:** 99.80% accuracy
- **XGBoost:** 99.88% accuracy

The SVM model with the RBF kernel delivers exceptional accuracy, comparable only to the Random Forest model, benefiting from effective data preprocessing.

RBF Kernel Accuracy: 99.99%

	precision	recall	f1-score	support
0	1.00	1.00	1.00	328816
1	1.00	1.00	1.00	17214
accuracy			1.00	346030
macro avg	1.00	1.00	1.00	346030
weighted avg	1.00	1.00	1.00	346030

Figure 2. Classification Report of SVM model

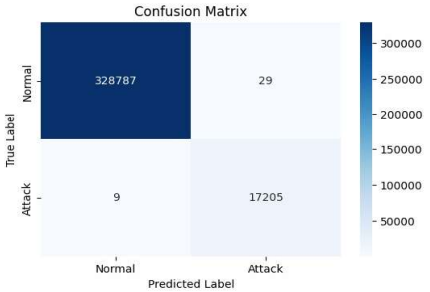

Figure 3. Confusion Matrix

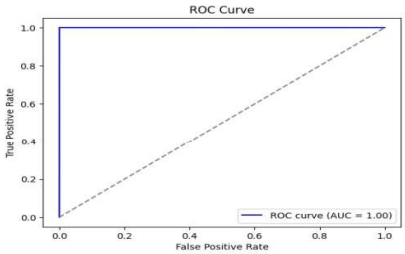

Figure 4. ROC curve of SVM model

5. CONCLUSION

This project implements an SVM-based IDS for real-time network security using an anomaly-based approach to detect unknown threats. Rigorous data preprocessing and model training ensure accurate classification of network traffic. Evaluation metrics confirm high detection accuracy with minimal false alarms. The IDS overcomes traditional limitations by being scalable and adaptive, enhancing threat detection. Challenges like limited resources, data labeling, and evolving threats are addressed through algorithm optimizations, ensuring reliable performance in dynamic environments, unlike signature-based methods.

6. REFERENCES

[1] A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019.

[2] A. S. Ahanger, S. M. Khan and F. Masoodi, "An Effective Intrusion Detection System using Supervised Machine Learning Techniques," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021.

[3] Ngueajio, M.K., Washington, G., Rawat, D.B., Ngueabou, Y. (2023). "Intrusion Detection Systems Using Support Vector Machines on the KDDCUP'99 and NSL-KDD Datasets: A Comprehensive Survey". In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2022.

[4] Jha, Jayshree, and Leena Ragha. "Intrusion detection system using support vector machine." International Journal of Applied Information Systems (IJAIS) 3 (2013): 25-30.

[5] Ajdani, Mahdi, and Hamidreza Ghaffary. "Design network intrusion detection system using support vector machine." International Journal of Communication Systems 34.3 (2021): e4689.

[6] Hasan, M., Nasser, M., Pal, B. and Ahmad, S., "Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS)", Journal of Intelligent Learning Systems and Applications, 2014.

[7] Almaiah, Mohammed Amin, et al. "Performance investigation of principal component analysis for intrusion detection system using different support vector machine kernels." Electronics 11.21 (2022): 3571

[8] Lirim Ashiku, Cihan Dagli, "Network Intrusion Detection System using Deep Learning", Procedia Computer Science Volume 185, 2021.