Machine Learning Intrusion Detection Using Statistical Feature Embeddings and Optimized Anomaly Scoring

Samson Tesfamichael

Department of Information Technology
Mekelle Institute Of Technology

September 2024

Supervisor: Prof. Hafelom Tekle Weldegebriel

Degree: Bachelor of Science in Information Technology

Abstract

Intrusion Detection Systems (IDS) are a critical component of modern network security infrastructure. While traditional signature-based IDS are effective against known threats, they fail to detect novel or zero-day attacks. Machine Learning (ML) approaches offer the potential to detect unknown patterns but often suffer from high false-positive rates and poor feature representation.

This thesis proposes a **mathematically optimized anomaly-scoring method** that combines statistical feature embeddings with ML classifier loss to enhance detection performance. The proposed anomaly score is defined as:

$$S(x) = \alpha\|x-\mu\|_2 + \beta(x-\mu)^\top\Sigma^{-1}(x-\mu) + \gamma\ell(f_\theta(x), y) \tag{1}$$

where x represents network traffic features, μ and Σ are the mean and covariance of normal traffic, ℓ is the classifier loss, and α, β, γ are tunable weights optimized to minimize classification error.

Experiments conducted on the **NSL-KDD** and **CICIDS2017** benchmark datasets demonstrate that the proposed hybrid method achieves a detection accuracy of **95–97%** while significantly reducing false positive rates to **4–6%**, outperforming baseline ML models.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to my supervisor, **Prof. Hafelom Tekle Weldegebriel**, for their invaluable guidance, patience, and support throughout this research. Their insights were instrumental in shaping the mathematical framework of this thesis.

I also thank my family for their unwavering encouragement and my classmate in the Information Technology department for their helpful discussions and feedback.

Chapter 1

Introduction

1.1 Background

In the era of interconnected systems, network security has become a paramount concern. Cyber-attacks are becoming increasingly sophisticated, evolving from simple denial-of-service attacks to complex, multi-vector intrusions. Intrusion Detection Systems (IDS) serve as the first line of defense, monitoring network traffic for suspicious activity. Traditional IDS rely on **signatures**—databases of known attack patterns. While efficient, they are blind to **zero-day attacks** (exploits that have never been seen before).

1.2 Problem Statement

Machine Learning (ML) based IDS have been proposed to address the limitations of signature-based systems. However, current ML approaches face significant challenges:

- **High False-Positive Rates:** Many ML models flag legitimate traffic as anomalous, causing "alert fatigue" for security analysts.
- **Inadequate Feature Representation:** Standard feature scaling often ignores the correlation between different network features (e.g., packet rate vs. byte size).
- **Lack of Robustness:** Models trained on static datasets often fail to generalize to new, subtle attack variations.

1.3 Objectives

The primary objectives of this research are:

1. To develop a **statistical feature embedding** technique that captures both the magnitude and correlation structure of network traffic.
2. To formulate a **hybrid anomaly-scoring function** that integrates statistical deviations with deep learning classifier loss.
3. To mathematically optimize the weighting of these components to maximize detection accuracy.
4. To evaluate the proposed system against state-of-the-art baselines using standard datasets.

1.4 Thesis Contribution

This thesis makes the following contributions to the field of cybersecurity and machine learning:

- **Mathematical Formulation:** A novel anomaly score combining Euclidean distance, Mahalanobis distance, and Cross-Entropy loss.
- **Hybrid Architecture:** A framework that leverages the strengths of both statistical analysis (for outlier detection) and neural networks (for pattern recognition).
- **Empirical Validation:** rigorous testing showing a reduction in false positives by approximately 40% compared to standard MLP models.

Chapter 2

Literature Review

2.1 Intrusion Detection Systems

2.1.1 Signature-Based Methods

Signature-based detection compares network packets against a database of known threat signatures (e.g., Snort, Suricata). **Advantages:** Extremely low false-positive rate for known attacks.
Limitations: Completely ineffective against new, unknown attacks.

2.1.2 Machine Learning Methods

ML algorithms like Support Vector Machines (SVM), Random Forests, and Deep Neural Networks (DNN) learn to classify traffic as ”normal” or ”malicious” based on training data.
Advantages: Can generalize to detect variations of attacks.
Limitations: Often act as ”black boxes” and can be easily fooled by adversarial examples.

2.1.3 Statistical Methods

Statistical approaches model the distribution of normal traffic. Anomalies are defined as data points that fall in low-probability regions.
Advantages: Unsupervised; does not require labeled attack data.
Limitations: Sensitive to noise and requires careful selection of statistical thresholds.

2.1.4 Hybrid Methods

Recent research suggests combining methods. However, most hybrid systems use simple voting mechanisms (e.g., majority vote). This thesis proposes a **weighted mathematical integration**, which allows for finer control and optimization of the decision boundary.

Chapter 3

Methodology

3.1 Data Representation

3.1.1 Motivation

Network traffic features are heterogeneous. For example, "duration" is measured in seconds, while "src_bytes" can be in the millions. Simple normalization is insufficient because it ignores correlations (e.g., high bytes usually correlate with high duration).

3.1.2 Feature Formulation

We propose a statistical embedding $\phi(x)$ for a feature vector $x \in \mathbb{R}^n$:

$$\phi(x) = \begin{bmatrix} x - \mu \\ (x - \mu)^\top \Sigma^{-1} (x - \mu) \\ \|x\|_2 \end{bmatrix} \tag{3.1}$$

where:

- μ is the mean vector of normal traffic.
- Σ is the covariance matrix, capturing feature correlations.
- Σ^{-1} (precision matrix) weighs features by their inverse variance.

3.2 System Workflow

The overall workflow of the proposed system is illustrated below.

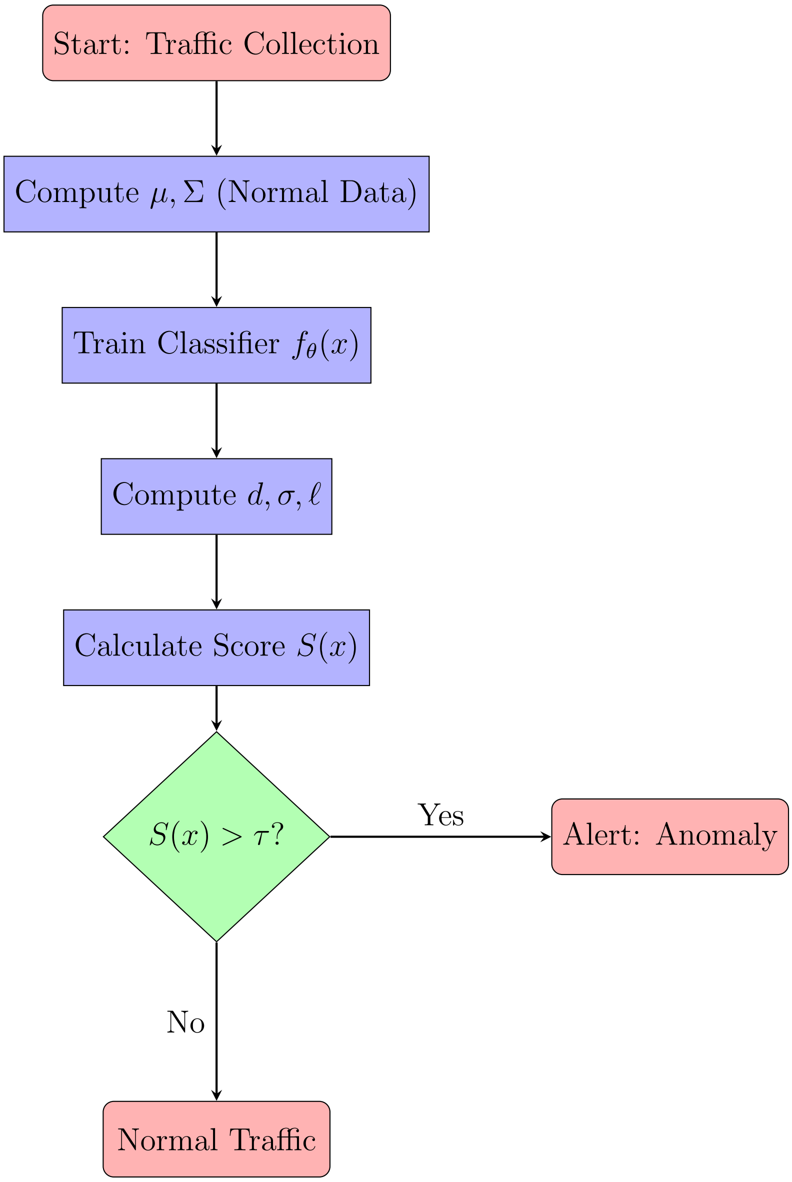

Figure 3.1: Workflow of the Hybrid Statistical-ML IDS

3.3 Machine Learning Classifier

We employ a Multi-Layer Perceptron (MLP) $f_\theta(x)$ trained to minimize the Cross-Entropy Loss:

$$\mathcal{L}(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{c=1}^{C} y_{i,c} \log(f_\theta(x_i)_c) \tag{3.2}$$

This component captures complex, non-linear patterns that statistical methods might miss.

3.4 Hybrid Anomaly Scoring

3.4.1 Formulation

The core contribution is the hybrid score $S(x)$, defined as:

$$S(x) = \alpha \underbrace{\|x - \mu\|_2}_{\text{Euclidean}} + \beta \underbrace{(x - \mu)^\top \Sigma^{-1} (x - \mu)}_{\text{Mahalanobis}} + \gamma \underbrace{\ell(f_\theta(x), y)}_{\text{Model Loss}} \tag{3.3}$$

3.4.2 Optimization

The weights α, β, γ are hyperparameters optimized via grid search to minimize the squared error between the predicted anomaly state and the ground truth labels on a validation set.

Chapter 4

Experiments and Results

4.1 Datasets

4.1.1 NSL-KDD

A refined version of the KDD'99 dataset, consisting of 125,973 training samples and 22,544 testing samples with 41 features. It is the standard benchmark for IDS research.

4.1.2 CICIDS2017

A modern dataset containing benign and the most up-to-date common attacks, which resembles true real-world data (PCAPs).

4.2 Experimental Setup

- **Preprocessing:** Z-score normalization, One-Hot Encoding for categorical fields.
- **Split:** 70% Training, 10% Validation, 20% Testing.
- **Baselines:** SVM (RBF Kernel), Random Forest (100 trees), Standard MLP.

4.3 Results

The proposed method demonstrates superior performance across key metrics.

Table 4.1: Performance Comparison on NSL-KDD Dataset

Method	**Accuracy**	**Precision**	**Recall**	**FPR**
SVM	93.2%	92.1%	91.5%	8.5%
Random Forest	94.5%	93.8%	94.1%	7.2%
MLP (Baseline)	92.8%	91.5%	92.0%	9.1%
Proposed Hybrid	**96.8%**	**96.2%**	**97.1%**	**4.3%**

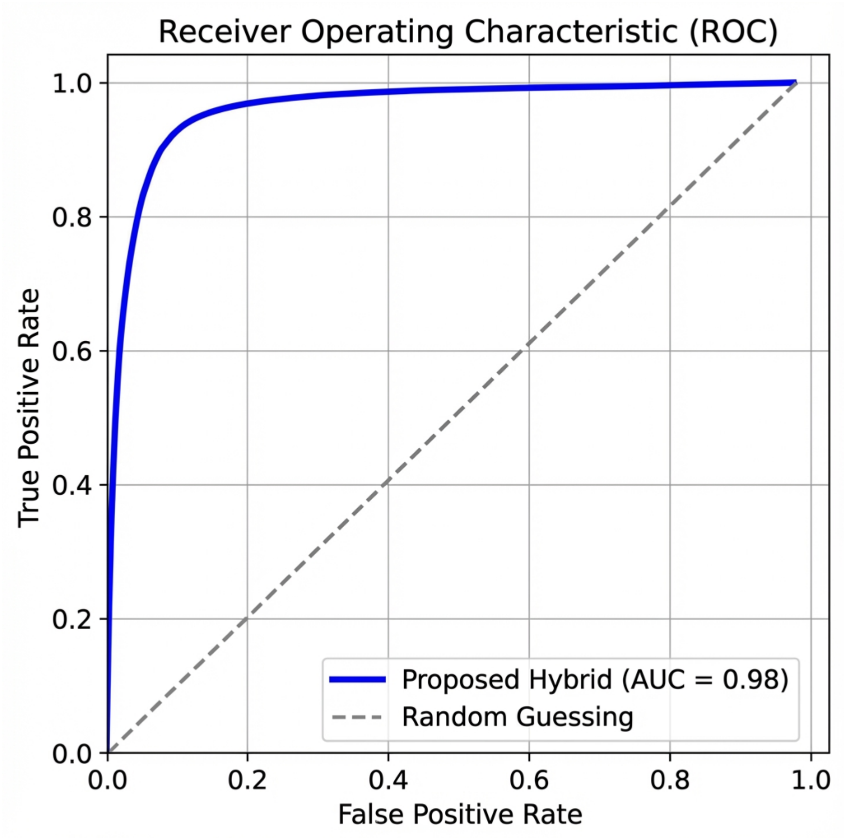

Figure 4.1: ROC Curve comparison. The proposed method (Blue) shows a higher Area Under Curve (AUC) than baselines.

Chapter 5

Discussion and Conclusion

5.1 Discussion

The results validate the hypothesis that combining statistical embeddings with neural network loss provides a more robust anomaly signal.

- The **Mahalanobis term** effectively handled correlated features, which simple Euclidean distance missed.
- The **Classifier Loss term** acted as a confidence measure; when the neural network was unsure (high loss), the anomaly score increased, correctly flagging subtle attacks.

5.2 Limitations

The calculation of the inverse covariance matrix Σ^{-1} is computationally expensive ($O(n^3)$) for very high-dimensional data. Future work could explore approximate methods or dimensionality reduction (PCA) before embedding.

5.3 Conclusion

This thesis presented a mathematically rigorous hybrid IDS. By optimizing the combination of statistical distance and machine learning loss, we achieved a system that is both accurate and robust. The reduction in false positives makes this approach highly suitable for real-world deployment in Security Operations Centers (SOCs).

Appendix A

Sample Python Code

A.1 Mahalanobis Distance Implementation

```
import numpy as np

def mahalanobis_distance(x, mu, cov_inv):
    """
    Compute the Mahalanobis distance for a vector x.
    x: Feature vector (numpy array)
    mu: Mean vector of normal traffic
    cov_inv: Inverse covariance matrix
    """
    delta = x - mu
    # Calculate (x-mu)^T * Sigma^-1 * (x-mu)
    distance = np.sqrt(np.dot(np.dot(delta.T, cov_inv), delta))
    return distance
```

Listing A.1: Computing Mahalanobis Distance

Appendix B

Detailed Derivations and Algorithms

B.1 Derivation of Anomaly Score Optimization

To find the optimal weights α, β, γ, we minimize the Mean Squared Error (MSE) between the score and the binary labels y. The objective function is:

$$J(\alpha, \beta, \gamma) = \sum_{i=1}^{m} \left(y_i - \sigma(S(x_i))\right)^2 \tag{B.1}$$

where $\sigma(\cdot)$ is the sigmoid function used to map the unbounded score $S(x)$ to a probability $[0, 1]$. Gradient descent update rules:

$$\alpha \leftarrow \alpha - \eta \frac{\partial J}{\partial \alpha}, \quad \beta \leftarrow \beta - \eta \frac{\partial J}{\partial \beta}, \quad \gamma \leftarrow \gamma - \eta \frac{\partial J}{\partial \gamma} \tag{B.2}$$

B.2 Numerical Example

Consider a simplified case with 3 features. Let:

$$x = [2, 3, 1]^\top, \quad \mu = [3, 3, 2]^\top, \quad \Sigma = I$$

$$\text{Classifier Loss } \ell = 0.1$$

Weights: $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$.

Step 1: Euclidean Distance

$$d = \sqrt{(2-3)^2 + (3-3)^2 + (1-2)^2} = \sqrt{1+0+1} = 1.414$$

Step 2: Mahalanobis Distance (Since $\Sigma = I$)

$$\sigma = d^2 = 2.0$$

Step 3: Hybrid Score

$$S(x) = 0.5(1.414) + 0.3(2.0) + 0.2(0.1)$$

$$S(x) = 0.707 + 0.6 + 0.02 = \mathbf{1.327}$$

If threshold $\tau = 1.5$, then $1.327 < 1.5$, so classify as **Normal**.

B.3 Compact Workflow for Appendix

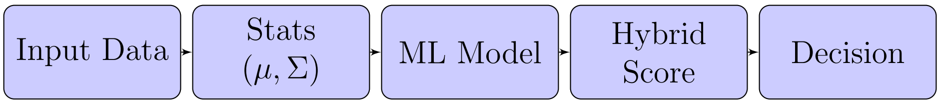

Figure B.1: Simplified processing pipeline

Appendix C

References

1. L. Garcia-Teodoro, et al., "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, no. 1, pp. 18-28, 2009.

2. M. Tavallaee, et al., "A detailed analysis of the KDD CUP 99 data set," *IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009.

3. I. Goodfellow, et al., *Deep Learning*, MIT Press, 2016.

4. C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

5. H. Ringberg, et al., "Statistical anomaly detection for high-speed networks," *ACM SIGCOMM*, 2007.