

Research Proposal

Title: Machine Learning for Network Intrusion Detection

Dr. Mohammed Anber

Student Name: Mohammed Alhawshabi✍

Student number: 261002622

Abstract

This research focuses on applying machine learning techniques to improve network intrusion detection systems. The study aims to enhance detection accuracy using supervised learning algorithms, including Random Forest and Support Vector Machine. The methodology includes data preprocessing, model training, and performance evaluation with standard metrics. The expected outcome is a more accurate and efficient intrusion detection model.

Chapter 1: Introduction

Machine learning plays a critical role in cybersecurity, particularly in detecting network intrusions. Traditional systems fail to detect modern attacks, so ML-based solutions are necessary.

Background

Previous research shows that ML models can improve intrusion detection, but challenges such as scalability and outdated datasets remain.

Problem Statement

Existing models rely on outdated datasets, overfit, and incur high computational costs. This creates a gap between research models and real-world applications, where modern threats are increasingly complex.

Objectives

- Improve detection accuracy
- Compare algorithms
- Reduce false positives
- Evaluate preprocessing impact

Scope

This research focuses on supervised ML techniques for intrusion detection, using selected datasets and algorithms such as Random Forest and SVM.

Chapter 2: Literature Review

Breiman (2001) introduced Random Forest, which improves accuracy and reduces overfitting. Cortes and Vapnik (1995) proposed SVM, which is effective for classification tasks. Tavallaee et al. (2009) highlighted issues in KDD datasets, such as redundancy and outdated attacks. Buczak and Guven (2016) reviewed ML methods in cybersecurity, noting scalability challenges. Ahmad et al. (2021) showed that deep learning improves accuracy but increases cost. Kim et al. (2014) proposed hybrid detection systems that improve performance but increase complexity.

Chapter 3: Methodology

Dataset: NSL-KDD

Data preprocessing: handling missing values, removing outliers using IQR

Models: Random Forest and SVM

Tools: Python, Pandas, Scikit-learn

Evaluation: Accuracy, Precision, Recall

Expected Outcomes

Improved accuracy, fewer false positives, and better model generalization.

Gantt Chart

Week 1: Literature Review

Week 2: Data Preprocessing

Week 3: Model Training

Week 4: Evaluation

References

- **Breiman, L. (2001). Random forests.**
- **Cortes, C., & Vapnik, V. (1995). Support-vector networks.**
- **Tavallaee, M. et al. (2009). KDD dataset analysis.**
- **Buczak, A., & Guven, E. (2016). ML in cybersecurity.**
- **Ahmad, Z. et al. (2021). Intrusion detection systems.**
- **Kim, G. et al. (2014). Hybrid intrusion detection.**