IGRF-RFE's Role in Transformer-Based Intrusion Detection Model Interpretability on UNSW-NB15

Assignee Research

June 11, 2026

Abstract

Network assaults pose significant security concerns to network services; hence, new technical solutions must be used to enhance the efficacy of intrusion detection systems. Existing approaches pay insufficient attention to data preparation and inadequately identify unknown network threats. This paper presents a network intrusion detection model (ID-RDRL) based on RFE feature extraction and deep reinforcement learning. ID-RDRL filters the optimum subset of features using the RFE feature selection technique, feeds them into a neural network to extract feature information and then trains a classi

1 Introduction

This paper examines: ID-RDRL: a deep reinforcement learning-based feature selection intrusion detection model. Research question: What is the impact of IGRF-RFE on model interpretability (e.g., SHAP values, feature importance consistency) when applied to transformer-based network intrusion detection models on UNSW-NB15 compared to standard filter methods?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

8 papers retrieved. 11 claims extracted; 10 independently verified. Quality review score: 8.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Network assaults pose significant security concerns to network services.	✓	0.24
Existing approaches pay insufficient attention to data preparation and inadequately identify unknown network threats.	✓	0.29
ID-RDRL filters the optimum subset of features using the RFE feature selection technique.	✓	0.36
ID-RDRL feeds the selected features into a neural network to extract feature information.	✓	0.24
ID-RDRL trains a classifier using DRL to recognize network intrusions.	✓	0.27
The CSE-CIC-IDS2018 dataset was used for testing the model's performance.	×	0.15
The CSE-CIC-IDS2018 dataset is a comprehensive collection of actual network traffic.	✓	0.23
The ID-RDRL model can select the optimal subset of features.	✓	0.26
The ID-RDRL model can remove approximately 80% of redundant features.	✓	0.24
The ID-RDRL model can learn the selected features through DRL to enhance the IDS performance for network attack identifi	✓	0.33
The ID-RDRL model has promising application potential in IDS in a complicated network environment.	✓	0.26

References

- https://doi.org/10.1038/s41598-022-19366-3
- https://doi.org/10.1038/s41598-025-34790-x
- https://doi.org/10.21203/rs.3.rs-1765453/v1