International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

UNSUPERVISED ANOMALY DETECTION USING NETWORK DATA

Praveen Kumar M[1],Prathap S[2], Shakthivel P [3], C.CHINNIMA[4]

Department of Computer Science, Kingston Engineering College, Vellore
Email: **praveenkumar210704@gmail.com**, **prathapsrinivasan18@gmail.com, shakthivelap.1@gmail.com**
4 Under the guidance of: AP/CSE Kingston Engineering College, Vellore
Email: **chinni16.engineering@kingston.ac.in**, Phone No:8248068911

ABSTRACT :

The growing complexity of cyber threats requires flexible security measures for networks. Like all classical Intrusion Detection Systems (IDS), the one presented in this study operates on the basis of signature sith provided. Latest developments in technology leave little room for the use of signature-based IDS systems because of the bounded efficiency they demonstrate against zero-day attacks. The framework presented here performs unsupervised anomaly detection based on an autoencoder deep learning architecture that detects intrusions by profiling normal traffic. Labeled datasets are not needed because the system works on reconstruction errors. EDA driven preprocessing also features thorough access control by adding borders to structures where change happens. The model developed in this research offers more than 92% accuracy while obtaining less than 5% false positives. In addition to expanding the scope of detection to include hybrid detection systems, future research will integrate edge computing and cloud computing to provide scalable real-time defense systems

Keywords: Network security, Intrusion detection, Autoencoder, Deep learning, Unsupervised learning, Anomaly detection, Cyber threats, Exploratory Data Analysis (EDA), Unsupervised Learning, Intrusion Detection Systems

1. Introduction

With the depiction of nature and networking, it describes how cybersecurity has become a prime concern by being targeted by advanced and ever-evolving threats. Classical IDS are signature-based, which limits their detection capacity against an unknown or zero-day attack. Therein, unsupervised machine learning approaches, mainly autoencoders, are gaining popularity due to their adjustable nature and minimum need for labeled datasets. Autoencoders learn normal network traffic patterns and consider any traffic different from these learned patterns as an anomaly. Thus, they are suitable for live anomaly detection in dynamic network environments. This research proposes an autoencoder-based deep learning intrusion detection system with extensive exploratory data analysis and threshold optimization to enhance performance.

2. Related Work

Since the introduction of SVMs, Random Forests, and Decision Trees, classical machine learning methods have been traditionally used for anomaly detection. Such classical methods have been shown to be less effective because of their reliance on manual feature engineering. Deep learning, conversely, tries to find deep patterns in data organically, but the region of curse still looms upon these models, with the exact problems of overfitting and computational intensity. The latest research has emphasized the advantages of using an autoencoder, especially the variational or hybrid variants, in anomaly detection. They consider reconstruction error as a possible measure of abnormal behavior, which enables effective detection of threats that no one has seen before. This work builds on these existing approaches, incorporating adaptive thresholding, as well as cloud deployment methods, allowing for much greater scalability and precision.

3. Methodology

Data Collection and Preprocessing

Network traffic data is fetched from well-structured datasets, and further processing through the Exploratory Data Analysis (EDA) technique takes place. Feature Engineering includes normalization, categorical encoding, and dimensionality reduction so as to get the data set ready for model input.

Model Architecture

The design of the autoencoder architecture entails:

- **Encoder:** Takes the input and compresses it into latent representation.
- **Bottleneck Layer:** Basically captures the essence of normal traffic
- **Decoder:** Tries to reconstruct the original input.

The reconstruction error is computed using Mean Square Error (MSE). Then, it optimizes the model using Adam.

Anomaly Detection

To separate anomalies and reduce false positives, the threshold is dynamically set by looking at the reconstruction error distribution.

4. System Architecture and Implementation

- Data Acquisition: Gathers network traffic samples.
- **Preprocessing:** Carries out EDA, feature selection, and normalization operations.
- **Model Training:** Autoencoder is given normal traffic for training purposes.
- **Evaluation:** Checked with accuracy, precision, recall, and F1-score.
- **Deployment:** Cloud and edge-based environments are employed to initiate real-time monitoring.

5. Results and Discussion

The new autoencoder model resulted in the success of 92% and traditional IDS was overshadowed. The false positive elimination occurred using the new method of adaptive thresholding, and it was very effective. The robustness of the model was confirmed through the previously mentioned metrics and the confusion matrices. Possible challenges, on the other hand, still exist. For instance, data bias and the susceptibility to changes in traffic flow are the two most severe obstacles in this respect.

6. Conclusion and Future Work

This paper notes the capability of pure deep learning without supervision for network anomaly detection. It is the encoder-based system that provides the most accurate and flexible solution, thus being of great help to zero-day and expanding threats detection. Potential enhancements are the models such as the mixed models combining unsupervised and supervised methods, the use of explainable AI methods (e.g., SHAP, Grad-CAM), and the deployment through cloud-edge integration for real-time protection.

7. Evaluation Metrics

- **Classification Metrics:** Accuracy, Precision, Recall, F1-Score
- **Visualization Tools:** Confusion Matrix, ROC Curve, t-SNE, PCA
- **Error Analysis**: Misclassification trends via threshold analysis

8. Ethical Considerations

- Fair and unbiased model training
- Privacy-preserving data handling
- Transparent decision-making using explainable AI

10. Case Study

A case study evaluates the real-world effectiveness of this approach using live network traffic samples. Comparative analysis with traditional IDS solutions highlights the improved detection capability and adaptability of the autoencoder-based model.

11. Comparative Analysis

Technique	Accuracy	False Positive Rate	Adaptability
SVM	85%	Moderate	Low
CNN	88%	High	Medium
Random Forest	89%	Moderate	Medium
Autoencoder (Proposed)	92%	Low	High

12. Limitations and Recommendations

Constraints:

- Understanding models continues to be difficult.
- Enhancing generalization over various datasets requires improvement.

Suggestions:

- Utilize explainable AI methods (e.g., SHAP) for clear decision-making insights.
- Investigate transfer learning for improved generalization.
- Dynamically adjust the threshold according to network activity