engineering proceedings

MDPI

Proceeding Paper

Attention-Guided Siamese Representation Learning for Cyber Anomaly Detection Using Mahalanobis Distance †

Adnane Touiyate [1,*], Chaïmaâ Kissi [2] and Tarik Boujiha [1]

[1] Engineering Science Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra 14000, Morocco; tarik.boujiha@uit.ac.ma
[2] Advanced Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra 14000, Morocco; chaimaa.kissi@uit.ac.ma
* Correspondence: adnane.touiyate@uit.ac.ma
† Presented at the 8th edition of the International Conference on Advanced Technologies for Humanity (ICATH 2026), Rabat, Morocco, 15–17 July 2026.

Abstract

The identification of malicious actions using traffic analysis is one of the key challenges in current cybersecurity practices. There are two existing approaches for solving this problem; the first utilizes machine learning algorithms on unprocessed data, while the second focuses on the reconstruction of traffic using autoencoders, which ignore subtle cues. In this research work, we present an algorithm where informative representations of network traffic are learned automatically. The main component of our algorithm is a Siamese Bidirectional LSTM network with an attention mechanism. While most solutions learn from raw tabular data, our model learns from sequence data, encoding network traffic in sequences of feature tokens. Therefore, it is possible to discover bidirectional sequential dependencies in structured data. The network learns using a metric learning loss function, where normal traffic samples are tightly packed together, while malicious samples are pushed further away. Anomalies are detected based on their distance to the normal traffic sample cluster using the Mahalanobis distance. We experiment with the NSL-KDD dataset, where our model attains an F1-score of 0.9007 and ROC-AUC of 0.9550, outperforming not only the Mahalanobis distance baseline but also an autoencoder-based anomaly detector. Moreover, compared to existing solutions, the proposed approach provides higher accuracy, improved interpretability, and high applicability in practice.

Keywords: cybersecurity; representation learning; Siamese networks; BiLSTM; attention mechanism; Mahalanobis distance

Check for updates

Academic Editors: Brahim El Bhiri, Ayman S. Mosallam and Younes Karfa Bekali

Published: 20 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Rapid proliferation of interconnected systems, cloud infrastructure, and Internet of Things (IoT) has significantly increased risks associated with cyber threats for organizations, with their frequency becoming higher and even harder to predict [1]. Thus, IDSs become one of the essential components of the modern cybersecurity framework, among which the anomaly-based IDSs are characterized by their ability to recognize both known and new attacks without relying on pre-defined signatures [2].

Early methods for anomaly detection were based on statistical analysis, including clustering or principal component analysis [3,4]; nevertheless, such approaches encounter difficulties dealing with high-dimensional and non-linear nature of traffic. Deep learning offered a breakthrough. For example, autoencoders are able to learn how to reconstruct

normal instances and use reconstruction loss as an indicator of anomalies [5]. However, such approaches have an inherent drawback: reconstruction losses fail to make the latent representations of normal and anomalous instances distinguishable, making it impossible to establish clear separation points [6]. LSTM-based algorithms can efficiently work with temporal data; nevertheless, they have been mostly used to analyze sequential inputs rather than tabular data [7].

There is a direct way to approach the task using metric learning, where siamese neural networks are specifically trained to learn similarity relationships between objects and make similar samples closer to each other and dissimilar samples further away, thus being perfectly suitable for anomaly detection problems [8,9]. Adding the attention mechanism allows the model to focus on the most informative parts of the input while simultaneously providing a degree of interpretability that is particularly valuable in cybersecurity applications [10].

Despite these advances, a few important gaps remain. How can structured tabular network data be effectively adapted for sequence-based deep learning? Can metric learning outperform reconstruction-based methods in terms of class separability? And can interpretability be built in from the start, rather than bolted on afterward? This work addresses all three questions in a single unified framework.

Specifically, we transform raw network records into feature-token sequences and train a Siamese Bidirectional LSTM (BiLSTM) encoder with attention using contrastive loss. The learned embeddings are then used with Mahalanobis distance for statistically grounded anomaly scoring. We evaluate on NSL-KDD [11] and compare against multiple baselines including raw-feature Mahalanobis, an autoencoder, Isolation Forest [12], and One-Class SVM [13].

The main contributions of this work are:

- A feature-token transformation that enables sequential deep learning on tabular cybersecurity data, capturing feature interactions beyond flat representations;
- A Siamese BiLSTM architecture with attention for learning discriminative, separable embeddings via contrastive metric learning;
- Mahalanobis distance-based anomaly detection in the learned latent space, providing statistical grounding and robustness to threshold selection;
- Built-in attention-based interpretability that highlights which features drive anomaly decisions, offering actionable insights for security analysts;
- A comprehensive experimental evaluation including five baselines, threshold robustness analysis, embedding visualizations, and an ablation on temporal modeling.

The rest of the paper is structured as follows. Section 2 walks through the methodology—covering how we preprocessed the data, learned representations, and detected anomalies. Section 3 then presents our experiments and how our approach stacks up against existing methods. In Section 4, we reflect on what the results mean, where the work falls short, and where it could go next. Section 5 wraps things up with our main conclusions.

2. Materials and Methods

2.1. Overview of the Proposed Framework

The proposed framework integrates representation learning, metric learning, and statistical anomaly detection into a unified and interpretable pipeline, as illustrated in Figure 1. Raw network records are transformed into feature-token sequences, enabling sequence modeling over tabular data. A Siamese BiLSTM encoder with attention learns discriminative embeddings using contrastive loss. During inference, anomaly detection is

https://doi.org/10.3390/engproc2026168017

performed using Mahalanobis distance in the learned embedding space, while attention weights provide feature-level interpretability.

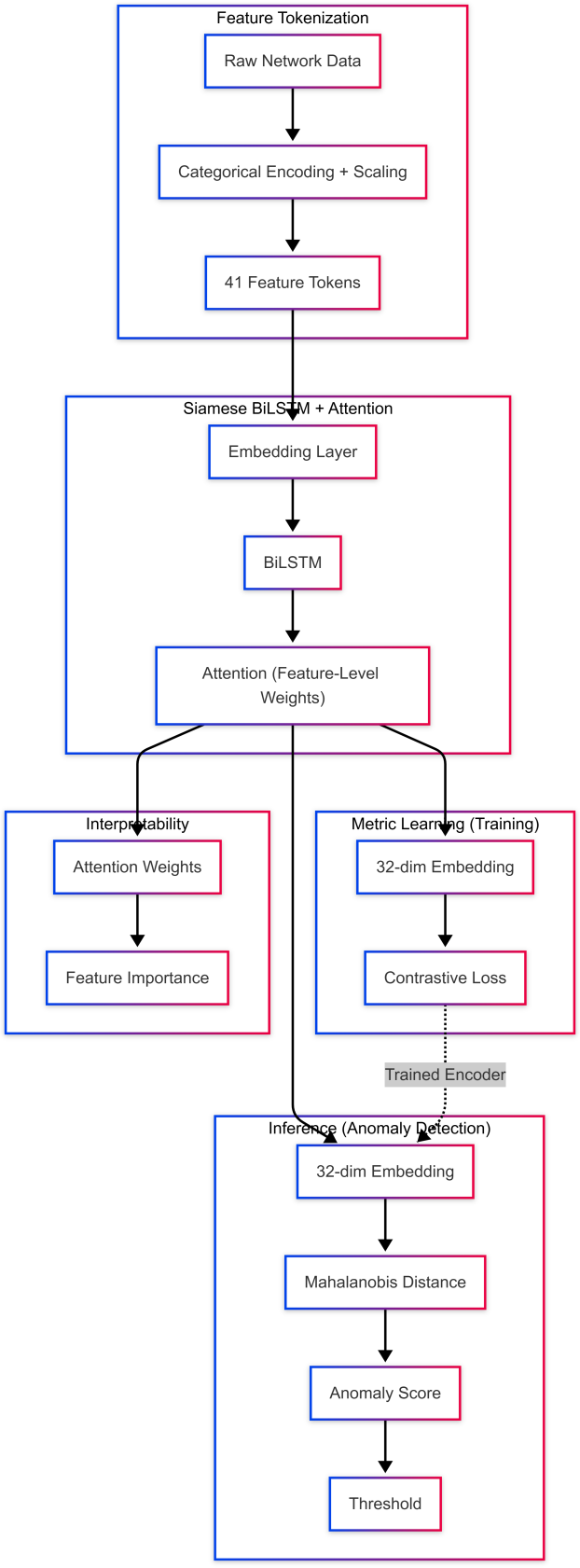


Figure 1. Architecture of the feature-token Siamese anomaly detection pipeline.

https://doi.org/10.3390/engproc2026168017

2.2. Dataset

Experiments are conducted on the NSL-KDD dataset [11], which consists of 41 features describing network connections. The problem is formulated as binary anomaly detection, where normal traffic defines the reference distribution and all attack types are treated as anomalies.

2.3. Data Preprocessing and Feature Tokenization

The categorical variables, which are (protocol_type, service, and flag), are converted using an integer encoding method. The numerical variables are normalized using a logarithm transformation and then standardized. Therefore, each instance is now described by 41 tokens of the input features. Unlike conventional intrusion detection approaches that operate directly on flat feature vectors, the proposed feature-token representation enables sequence-based representation learning while preserving the information contained in each network record.

2.4. Siamese BiLSTM Encoder with Attention

The model employs a Siamese architecture with shared weights. Each branch consists of an embedding layer, a BiLSTM encoder, and an attention mechanism. The BiLSTM captures feature interactions, while attention assigns importance weights to each token.

The attention weights are computed as:

$$\alpha_t = \frac{\exp(w^T h_t)}{\sum_{k=1}^{41} \exp(w^T h_k)} \tag{1}$$

The final embedding vector $z \in \mathbb{R}^{32}$ is obtained as:

$$z = W_z\left(\sum_{t=1}^{41} \alpha_t h_t\right) + b_z \tag{2}$$

2.5. Metric Learning

The model is trained using contrastive loss:

$$\mathcal{L} = y \cdot D^2 + (1 - y) \cdot \max(0, m - D)^2 \tag{3}$$

where $D = \|z_1 - z_2\|_2$ and $m = 1.0$.

2.6. Anomaly Detection

A multivariate Gaussian distribution is fitted on normal embeddings. The Mahalanobis distance is defined as:

$$D_M(z) = \sqrt{(z - \mu)^T \Sigma^{-1} (z - \mu)} \tag{4}$$

Samples exceeding a threshold τ are classified as anomalies.

2.7. Training and Evaluation

The model is trained using Adam (learning rate = 0.001, batch size = 128) for 10 epochs. Performance is evaluated using precision, recall, F1-score, and ROC-AUC. Thresholds are selected based on percentiles of the training distribution.

https://doi.org/10.3390/engproc2026168017

3. Results

All results are obtained on the standard NSL-KDD train–test split without data leakage.

3.1. Overall Performance

The proposed method achieves strong anomaly detection performance:

- Precision: 0.9137.
- Recall: 0.8881.
- F1-score: 0.9007.
- ROC-AUC: 0.9550.

These results indicate a good balance between precision and recall, while the high ROC-AUC confirms reliable discrimination between normal and anomalous samples (Figure 2).

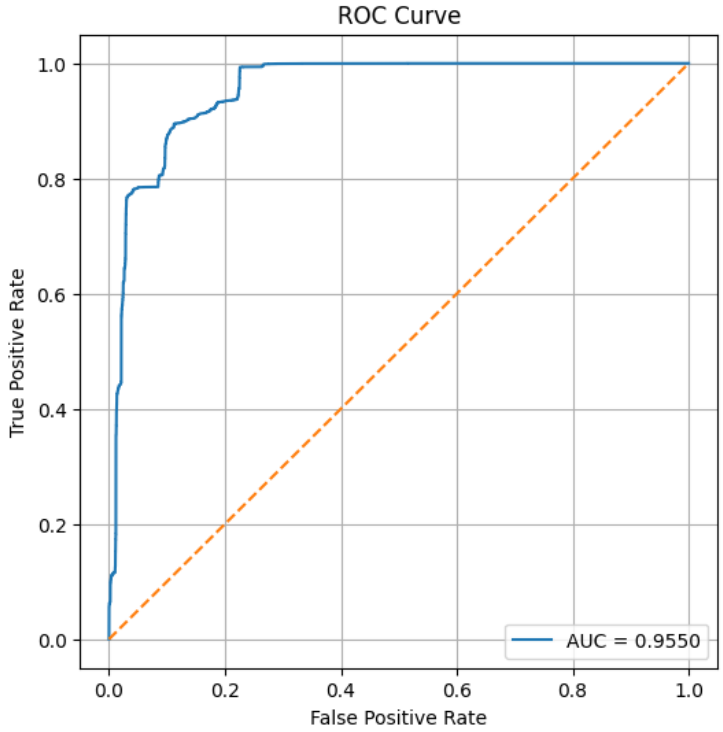


Figure 2. ROC curve of the proposed method. The orange dashed diagonal indicates the chance-level (random-classifier) reference.

3.2. Comparison with Baselines

We compare the proposed method against four baselines: Mahalanobis distance on raw features, an autoencoder-based approach, Isolation Forest, and One-Class SVM (Table 1).

Table 1. Comparison with baseline methods. Bold values indicate the highest score in each column.

Method	F1-Score	ROC-AUC
Autoencoder + Mahalanobis	0.5260	0.8168
Isolation Forest	0.8488	0.9498
Raw Features + Mahalanobis	0.8603	0.9495
One-Class SVM	0.8857	0.9162
Proposed Siamese BiLSTM + Mahalanobis	**0.9007**	**0.9550**

The proposed method achieves the best overall performance, obtaining the highest F1-score and ROC-AUC among all evaluated baselines. While One-Class SVM provides com-

petitive F1 performance, its ROC-AUC remains lower than that of the proposed approach. Isolation Forest and raw-feature Mahalanobis both perform reasonably well, but remain below the proposed method. The weakest performance is observed for the autoencoder-based baseline, suggesting that reconstruction-oriented objectives are less effective than discriminative metric learning for this task.

3.3. Embedding Space Analysis

The learned embedding space is visualized using UMAP (Figure 3). Normal samples form compact clusters, while anomalous samples are clearly separated, supporting the effectiveness of the learned representations.

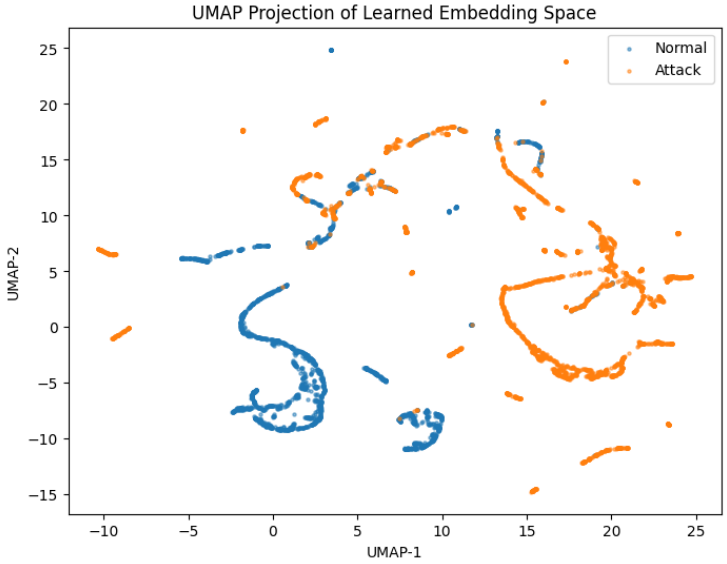


Figure 3. UMAP projection of the learned embedding space.

3.4. Mahalanobis Score Distribution

Figure 4 shows the distribution of Mahalanobis distances. A clear separation between normal and anomalous samples is observed, confirming the suitability of the statistical modeling approach.

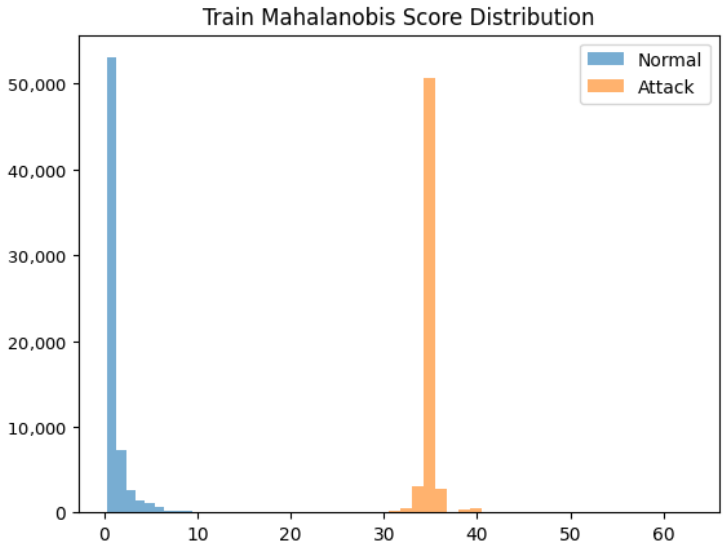


Figure 4. Distribution of Mahalanobis scores.

https://doi.org/10.3390/engproc2026168017

3.5. Threshold Sensitivity

We evaluate performance under different threshold values (Table 2):

Table 2. Threshold sensitivity analysis.

Percentile	Precision	Recall	F1-Score
90%	0.9200	0.8707	0.8947
95%	0.9539	0.7843	0.8608
97.5%	0.9704	0.6207	0.7572

The results show a trade-off between precision and recall, with the 90th percentile providing a balanced operating point.

3.6. Attention-Based Interpretability

In Figure 5, the features which play the most crucial role in discrimination according to the attention mechanism are pointed out, thereby gaining insight into the decision-making process of the model.

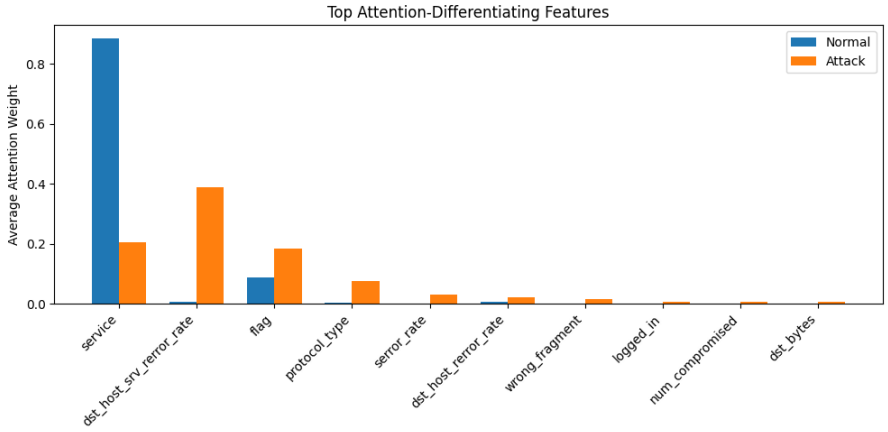


Figure 5. Top attention-differentiating features.

3.7. Impact of Temporal Modeling

Temporal aggregation leads to a performance decrease:

- Without temporal modeling: F1-score = 0.9007.
- With temporal modeling: F1-score = 0.6996.

This suggests that temporal aggregation introduces noise and disrupts the learned discriminative structure of the embedding space, rather than providing additional useful information.

4. Discussion

It is clear from the results that learning an embedding space structured in a certain way is associated with meaningful improvement in anomaly detection compared to working with raw features or applying reconstruction-oriented optimization objectives. For instance, the raw-feature Mahalanobis baseline shows poor performance since it is impossible to separate data using the raw features due to lack of necessary geometry (F1 = 0.8603). The results for autoencoder methods clearly demonstrate this point: reconstruction-oriented algorithms focus on reproducing the original data and not separating classes, as can be concluded from the worst-performing method in terms of F1-score among all considered ones (F1 = 0.5260).

In turn, classical algorithms like One-Class SVM and Isolation Forest show decent results. However, it is worth mentioning that their effectiveness is limited because of their dependency on the raw feature space and inability to perform representation learning that would help to find subtle behavior differences between normal and anomalous data. Such limitations are overcome by the presented approach thanks to training an embedding space suitable for anomaly detection. The results can be illustrated by UMAP plot, which demonstrates that normal samples form a tight cluster, whereas anomalous ones are pushed away into peripheral regions—exactly what is needed for effective application of Mahalanobis distance.

It can also be observed from the training dynamics that positive pairs tend to converge while negative examples become farther apart during optimization—which is expected for a well-formulated contrastive learning task. Speaking of the threshold choice for anomaly prediction, it is obvious that high percentile thresholds lead to better precision but worse recall, which makes the 90th percentile an effective solution.

Furthermore, the attention analysis allows for better interpretability of the framework due to the explicit correlation established between the learned weights of the features and domain-specific knowledge about the corresponding concepts. For example, the weights assigned to the feature `service`, as well as to `dst_host_srv_rerror_rate` and `flag`, always demonstrate a substantial growth that is commensurate with the expected impact on the final decision. In other words, the results obtained are consistent with the domain knowledge of the network security and allow for the interpreting of the results generated by the model as behavior-based outputs, not a consequence of a peculiar distribution of the data used.

One particular point that should be addressed with regard to the methodology is how the BiLSTM is used in the context of the problem. In this case, it is important to note that although the input dataset (NSL-KDD) is tabular in nature, every sample is encoded as a sequence in order to capture the interaction between the features, not between the temporal dimensions. In other words, BiLSTM becomes a feature-space structure encoding technique that allows applying a contrastive metric learning based on Mahalanobis distances.

In other words, the entire architecture can be interpreted as an aggregation of four independent components (according to their order of execution): (i) transformation of a tabular data sample into a sequence, (ii) representation learning at the feature level, i.e., using a Siamese BiLSTM and attention, (iii) learning of the sequence-based Mahalanobis metric using a pairwise approach, and (iv) estimation of the anomalies via Mahalanobis distance calculation. The proposed architecture provides a highly effective solution, whereas its explainable character naturally emerges from the concept of the model. From a computational perspective, the proposed framework remains relatively lightweight. Training converged within 10 epochs using a single GPU, while inference requires only a forward pass through the trained encoder followed by Mahalanobis distance computation in a 32-dimensional embedding space. Consequently, the approach remains suitable for near real-time deployment scenarios and can be integrated into practical intrusion detection systems with limited computational overhead.

An important aspect worth highlighting is that introducing a temporal dimension to the training causes performance deterioration when it comes to anomaly detection. The introduction of the additional feature led to a drop in the F1-score from 0.9007 to 0.6996. In all likelihood, this effect can be explained by temporal pooling that breaks down an already built embedding space of feature sequences.

https://doi.org/10.3390/engproc2026168017

Limitations and Future Work

There are several limitations inherent in this study. First, while the NSL-KDD benchmark dataset is well-established and allows for direct and consistent comparisons to be drawn, it is somewhat dated, exhibiting signs of synthetic design and thus failing to capture the full spectrum of attack signatures seen in today's networks. Testing the performance of the proposed methodology on recent benchmarks, including the CIC-IDS2017 [14] and UNSW-NB15 [15] datasets, is paramount for future work. Since both the feature-token pipeline and the metric learning objective functions are dataset-independent, generalizability should prove feasible.

Aside from the selection of benchmarks, the current implementation considers intrusions as a binary classification problem, distinguishing only between legitimate and malicious activity. Extending the framework to handle multi-class intrusion types would make the algorithm applicable in practice, as understanding which attacks have taken place is crucial beyond detection alone. Lastly, this paper focuses exclusively on offline experimentation. Deploying the model into production poses several problems that must be addressed through further investigation.

5. Conclusions

In this paper, we introduced an attention-guided Siamese BiLSTM framework for anomaly detection in cybersecurity. Rather than feeding raw tabular features directly into a classifier—as most intrusion detection systems do—we first transform each network record into a sequence of feature tokens. This lets the model learn richer, sequence-aware representations while still retaining all the information from the original connection record. Anomaly scoring is then handled through Mahalanobis distance, which gives the detection process a solid statistical footing.

On the NSL-KDD benchmark, our framework outperformed every baseline we tested—One-Class SVM, Isolation Forest, raw-feature Mahalanobis distance, and autoencoder-based detection—reaching an F1-score of 0.9007 and a ROC-AUC of 0.9550. UMAP visualizations backed this up, showing clean separation between normal and attack samples in the learned embedding space.

One aspect we find particularly valuable is interpretability. The attention mechanism highlights which features drive each decision, and the features it flags line up well with what cybersecurity practitioners already know. That alignment makes the model easier to trust and more deployable than black-box alternatives.

That said, NSL-KDD is an older benchmark, and we plan to test the framework on more recent datasets such as CIC-IDS2017 and UNSW-NB15. We are also interested in extending it toward multi-class attack classification, online detection, and eventually real-time deployment.

Author Contributions: Conceptualization, A.T.; methodology, A.T.; software, A.T.; validation, A.T., C.K. and T.B.; writing—original draft preparation, A.T.; writing—review and editing, C.K. and T.B.; supervision, T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study uses a publicly available network intrusion detection benchmark and does not involve human participants or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The NSL-KDD dataset used in this study is publicly available.

Acknowledgments: The authors would like to thank Ibn Tofail University and the National School of Applied Sciences of Kenitra for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cisco. *Cisco Security Outcomes Report, Volume 3*; Cisco Systems: San Jose, CA, USA, 2023.
2. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 1–58. [CrossRef]
3. Lippmann, R.; Fried, D.; Graf, I.; Haines, J.; Kendall, K.; McClung, D.; Weber, D.; Webster, S.; Wyschogrod, D.; Cunningham, R.; et al. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'00), Hilton Head, SC, USA, 25–27 January 2000; Volume 2, pp. 12–26. [CrossRef]
4. Sommer, R.; Paxson, V. Outside the closed world: On using machine learning for network intrusion detection. In *IEEE Symposium on Security and Privacy*; IEEE Computer Society: Washington, DC, USA, 2010; pp. 305–316. [CrossRef]
5. An, J.; Cho, S. *Variational Autoencoder Based Anomaly Detection Using Reconstruction Probability*; Technical Report; Seoul National University: Seoul, Republic of Korea, 2015.
6. Chong, Y.; Tay, Y.H. Abnormal event detection in videos using spatiotemporal autoencoder. In *International Symposium on Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2018.
7. Kim, J.; Kim, J.; Thu, H.L.T.; Kim, H. Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection. In Proceedings of the 2016 International Conference on Platform Technology and Service (PlatCon), Jeju, Republic of Korea, 15–17 February 2016; pp. 1–5. [CrossRef]
8. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 10–11 July 2015.
9. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep learning for anomaly detection: A review. *ACM Comput. Surv.* **2021**, *54*, 1–38. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
11. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), Ottawa, ON, Canada, 8–10 July 2009. [CrossRef]
12. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM'08), Pisa, Italy, 15–19 December 2008; pp. 413–422. [CrossRef]
13. Schölkopf, B.; Platt, J.C.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C. Estimating the support of a high-density region. *Neural Comput.* **2001**, *13*, 1443–1471. [CrossRef] [PubMed]
14. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP), Funchal, Portugal, 22–24 January 2018. [CrossRef]
15. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems. In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 10–12 November 2015. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.