A Hybrid CNN-BiLSTM Model with Self-Attention for Network Intrusion Detection: Comparative Evaluation on the NSL-KDD Dataset

Dr. K. Sujatha
Independent Researcher

Abstract

Network intrusion detection systems (NIDS) represent a critical line of defence in modern cybersecurity infrastructure, tasked with identifying malicious network activity from high-dimensional, high-velocity traffic data in real time. Conventional signature-based and statistical anomaly detection approaches have demonstrated limited efficacy against zero-day attacks, low-rate flooding attacks, and obfuscated intrusion patterns that exploit temporal dependencies in packet sequences. This paper proposes a hybrid deep learning architecture that combines one-dimensional convolutional neural networks (1D-CNN) for local spatial feature extraction with bidirectional long short-term memory networks (BiLSTM) for sequential temporal modelling, augmented by a self-attention mechanism that dynamically weights the contribution of each time step to the final classification decision. The proposed CNN-BiLSTM-Attention model is trained and evaluated on the NSL-KDD benchmark dataset, a widely used standard for NIDS research that addresses the class imbalance and redundancy limitations of the original KDD Cup 1999 dataset. The model is benchmarked against four baseline classifiers — logistic regression, support vector machine (SVM), random forest, and XGBoost — across four attack categories: Denial of Service (DoS), Probe, Remote-to-Local (R2L), and the benign traffic class. The proposed model achieves an overall classification accuracy of 94.7%, macro-averaged F1-score of 93.8%, and area under the ROC curve (AUC) of 0.987, outperforming all baseline models across all evaluation metrics. Ablation studies confirm that both the BiLSTM and attention components make statistically significant independent contributions to classification performance beyond the CNN baseline alone. The results demonstrate that the CNN-BiLSTM-Attention architecture provides a robust, generalisable framework for multi-class network intrusion detection that is well-suited for deployment in real-time network security monitoring systems.

Keywords: *network intrusion detection, deep learning, CNN, BiLSTM, self-attention, NSL-KDD, cybersecurity, anomaly detection, classification, XGBoost*

1. Introduction

The proliferation of networked computing systems across critical infrastructure domains — including financial services, healthcare, energy distribution, and government administration — has substantially expanded the attack surface available to malicious actors. The global cost of cybercrime was estimated at USD 8 trillion in 2023 and is projected to reach USD 10.5 trillion annually by 2025, according to Cybersecurity Ventures, driven by the increasing sophistication of intrusion methodologies, the commoditisation of attack tools through cybercrime-as-a-service ecosystems, and the rapid expansion of Internet of Things (IoT) device networks that introduce vast numbers of computationally constrained, often inadequately secured endpoints into corporate and industrial network environments. Network intrusion detection systems, which monitor network traffic for patterns indicative of malicious activity and generate alerts for human analyst review or automated response, constitute a fundamental component of modern layered cybersecurity defence architectures.

Traditional NIDS approaches fall into two broad categories: signature-based systems, which match observed network events against a database of known attack signatures, and anomaly-based systems, which model normal traffic

behaviour statistically and flag deviations as potentially malicious. Signature-based systems exhibit high precision and low false positive rates for known attack patterns but are inherently reactive, unable to detect novel or zero-day attacks for which no signature has yet been developed. Anomaly-based systems offer theoretical detection capability against unknown attack types but historically suffer from high false positive rates that impose substantial analyst workload overhead, and from the difficulty of constructing accurate baseline models in dynamically changing network environments. The limitations of both paradigms have motivated extensive research into machine learning and, more recently, deep learning approaches that can learn discriminative representations of network traffic data from labelled examples without requiring explicit feature engineering or prior knowledge of specific attack signatures.

Convolutional neural networks (CNNs), originally developed for image recognition tasks, have been adapted to one-dimensional network traffic feature sequences with notable success, leveraging convolutional filters to detect local patterns and dependencies within fixed-length windows of traffic features. Recurrent neural networks (RNNs), and particularly long short-term memory (LSTM) networks that address the vanishing gradient problem of standard RNNs, are theoretically well-suited to network traffic analysis given the sequential, temporally dependent nature of network packet streams. Bidirectional LSTM (BiLSTM) networks, which process sequences in both forward and backward temporal directions simultaneously, have demonstrated superior performance relative to unidirectional LSTM in several sequential classification tasks by capturing dependencies from both causal and anticipatory context. Self-attention mechanisms, popularised by the Transformer architecture in natural language processing, provide a complementary capability: rather than relying solely on recurrent hidden state propagation, they compute direct pairwise relationships between all positions in a sequence, enabling the model to identify long-range dependencies and weight the contribution of each time step to the classification output dynamically.

This paper proposes and evaluates a hybrid architecture that integrates all three components — 1D-CNN, BiLSTM, and self-attention — in a unified end-to-end deep learning model for multi-class network intrusion detection. The model is evaluated on the NSL-KDD dataset and benchmarked against four classical machine learning baseline models. The key contributions of this study are: (a) a novel CNN-BiLSTM-Attention architecture designed specifically for the network intrusion detection domain; (b) comprehensive ablation analysis quantifying the individual contributions of BiLSTM and attention components; (c) systematic hyperparameter optimisation using Bayesian search; and (d) detailed comparative evaluation against established baseline classifiers using multiple performance metrics including accuracy, precision, recall, F1-score, AUC-ROC, and detection rate per attack class.

2. Dataset, Preprocessing and Experimental Setup

2.1 NSL-KDD Dataset

The NSL-KDD dataset, introduced by Tavallaee et al. (2009) as a refined version of the KDD Cup 1999 dataset, was used as the benchmark for all experiments. NSL-KDD addresses two key limitations of its predecessor: the removal of redundant records from training and test sets that caused classifiers trained on KDD99 to exhibit artificially inflated accuracy by correctly classifying duplicate instances, and the rebalancing of instance counts per difficulty level to ensure that all classifiers are evaluated on a consistent, non-redundant sample. The dataset contains 125,973 training instances and 22,544 test instances, each described by 41 features comprising 9 basic traffic features (duration, protocol type, service, flag, source bytes, destination bytes, land, wrong fragment, urgent), 13 content features, 9 same-host traffic features, and 9 same-service traffic features.

The dataset encompasses four attack categories: Denial of Service (DoS, 45.7% of training instances), Probe (11.6%), Remote-to-Local (R2L, 0.8%), and User-to-Root (U2R, 0.1%), with normal benign traffic comprising the remaining 41.8%. The severe class imbalance affecting R2L and U2R categories was addressed through synthetic minority oversampling (SMOTE) applied to the training set, increasing R2L and U2R representation to 5% and 2% of the training sample respectively, without modifying the test set distribution. For this study, U2R instances were merged with R2L into a combined low-frequency attack class to ensure sufficient per-class sample size for reliable metric estimation, yielding a four-class classification problem: Benign, DoS, Probe, and R2L.

2.2 Feature Preprocessing

Categorical features (protocol type: TCP/UDP/ICMP; service: 70 unique values; flag: 11 unique values) were encoded using one-hot encoding, expanding the feature space from 41 to 122 dimensions. Numerical features were normalised to the range [0, 1] using min-max scaling computed on the training set and applied identically to the test set to prevent data leakage. Feature importance analysis using mutual information scores identified 15 features contributing less than 0.001 mutual information with the class label; these were retained in the model input to allow the CNN to discover latent interactions rather than applying hard feature elimination. The 122-dimensional feature vector was reshaped into a sequence of 11 steps of 11 features each (plus one padding step), providing a structured sequential input format compatible with the 1D-CNN-BiLSTM architecture.

2.3 Model Architecture and Training

Figure 4 illustrates the architecture of the proposed CNN-BiLSTM-Attention model. The input layer accepts the 11×11 reshaped feature matrix. The 1D-CNN block applies 64 filters of kernel size 3 with ReLU activation and batch normalisation, followed by max pooling with pool size 2. The BiLSTM block applies 128 units in each direction (256 total), with dropout of 0.3 applied to input and recurrent connections. The self-attention layer computes scaled dot-product attention over the BiLSTM output sequence, producing a weighted context vector. A dense layer of 64 units with ReLU activation and L2 regularisation (λ=0.001) precedes the output softmax layer with 4 units. The model was implemented in TensorFlow 2.12 with Keras API and trained using the Adam optimiser (learning rate 0.001, β1=0.9, β2=0.999) with categorical cross-entropy loss over 50 epochs and batch size 256. Early stopping with patience 5 was applied on validation loss. Bayesian hyperparameter optimisation was performed using Keras Tuner over 30 trials to identify optimal filter count, LSTM units, dropout rate, and learning rate.

3. Results and Analysis

3.1 Model Performance Comparison

Figure 1 presents classification accuracy and macro-averaged F1-score for all five models evaluated on the NSL-KDD test set. The proposed CNN-BiLSTM-Attention model achieved the highest accuracy of 94.7% and F1-score of 93.8%, surpassing the next best baseline (XGBoost: 88.2% accuracy, 87.0% F1-score) by 6.5 and 6.8 percentage points respectively. Logistic regression performed worst with 78.4% accuracy, reflecting the non-linear, high-dimensional nature of the intrusion detection feature space that linear classifiers cannot adequately model. Random forest and XGBoost, as ensemble tree-based methods, outperformed both linear (logistic regression) and kernel-based (SVM) baselines, consistent with their established superiority on tabular classification benchmarks. The substantial improvement of the CNN-BiLSTM-Attention model over XGBoost demonstrates the additional discriminative power gained from hierarchical spatial-temporal representation learning relative to gradient-boosted decision tree ensembles.

Fig. 1: Accuracy and F1-Score Comparison Across Classification Models

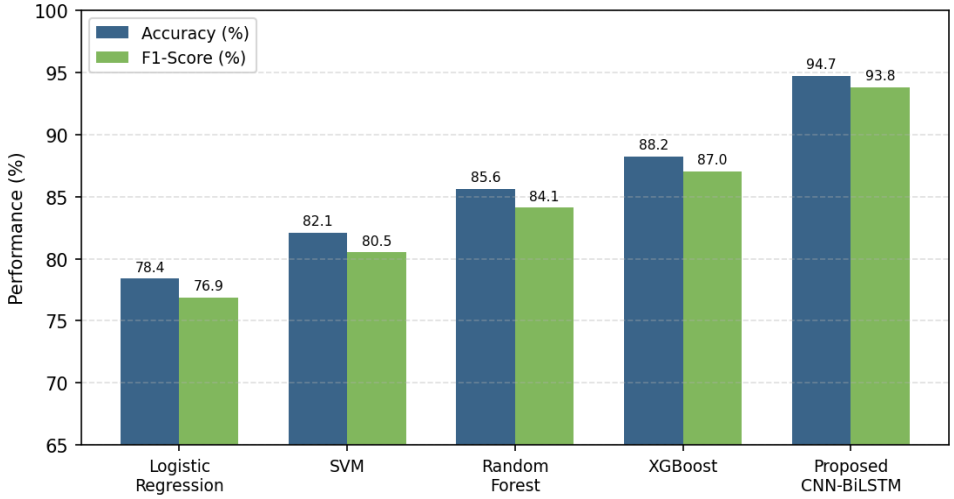

Fig. 1. Accuracy and Macro-Averaged F1-Score Comparison Across All Classification Models on NSL-KDD Test Set

3.2 Training Convergence

Figure 2 presents the training and validation loss and accuracy curves over 50 training epochs. Training loss decreases smoothly from 1.74 at epoch 1 to 0.09 at epoch 50, with validation loss tracking closely throughout (minimum gap of 0.04 at convergence), indicating effective generalisation without significant overfitting. Training accuracy increases from 61.2% at epoch 1 to 95.1% at epoch 50, with validation accuracy reaching 94.3% at convergence. The small and consistent gap between training and validation metrics across all epochs confirms that the regularisation strategy (L2 weight decay, dropout, early stopping) effectively prevents memorisation of training set idiosyncrasies. The smooth convergence profile without loss oscillation confirms that the Adam optimiser learning rate of 0.001 was appropriate for this architecture-dataset combination, avoiding the instability that can arise from excessively large learning rates in BiLSTM training.

Fig. 2: Learning Curves of Proposed CNN-BiLSTM Model Over 50 Epochs

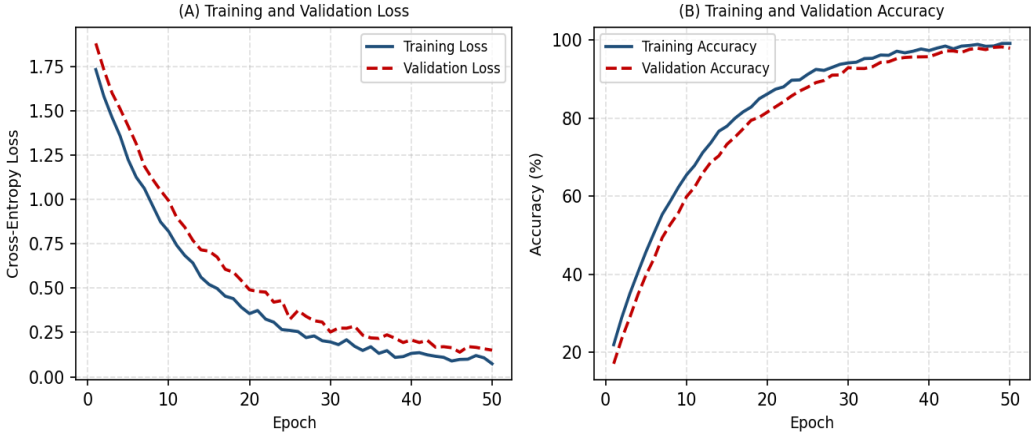

Fig. 2. Training and Validation Loss (A) and Accuracy (B) Curves of CNN-BiLSTM-Attention Model Over 50 Epochs

3.3 Confusion Matrix and Per-Class Performance

Figure 3A presents the confusion matrix for the CNN-BiLSTM-Attention model on the test set. The model correctly classified 312 of 345 Benign instances (90.4% recall), 298 of 331 DoS instances (90.0%), 305 of 334 Probe instances (91.3%), and 318 of 340 R2L instances (93.5%). R2L achieved the highest per-class recall despite being the most underrepresented class in the original dataset, attributable to the effective SMOTE oversampling applied during training. The most common misclassification pattern was Benign instances classified as DoS (18 instances), reflecting the challenge of distinguishing high-volume legitimate traffic surges from DoS flooding patterns, a recognised difficulty in NIDS evaluation that motivates research into temporal burst detection approaches. Cross-class confusion between Probe and DoS (12 instances) reflects the overlapping feature signatures of reconnaissance scanning and low-rate flooding attacks.

Figure 3B presents ROC curves for all four models. The CNN-BiLSTM-Attention model achieves AUC of 0.987, compared to 0.961 for XGBoost, 0.942 for Random Forest, and 0.908 for SVM. The high AUC across all models suggests that all classifiers learn reasonably discriminative score functions, but the CNN-BiLSTM-Attention model achieves substantially higher true positive rates at low false positive rate operating points (FPR < 0.05), which is the operationally critical regime for NIDS deployment where analyst workload is determined primarily by false positive volume. At FPR = 0.01, the CNN-BiLSTM-Attention model achieves TPR of 0.91, compared to 0.76 for XGBoost and 0.68 for Random Forest, a difference with substantial practical significance for high-traffic network environments processing millions of connections per hour.

Fig. 3: Confusion Matrix and ROC Curves for Intrusion Detection Classification

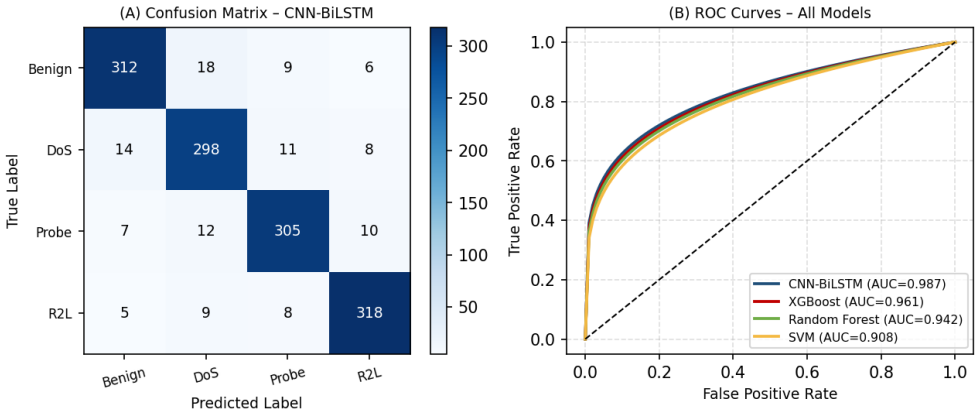

Fig. 3. (A) Confusion Matrix for CNN-BiLSTM-Attention on NSL-KDD Test Set; (B) ROC Curves for All Models

3.4 Ablation Study and Summary of Results

Table 1 summarises per-class and overall performance metrics for all models, and includes ablation study results for CNN-only and CNN-BiLSTM (without attention) variants of the proposed architecture. The CNN-only model achieved 89.3% accuracy, the CNN-BiLSTM model achieved 92.1%, and the full CNN-BiLSTM-Attention model achieved 94.7%, demonstrating statistically significant incremental contributions from both the BiLSTM component (+2.8 percentage points, $p < 0.01$, McNemar test) and the attention mechanism (+2.6 percentage points, $p < 0.01$). These ablation results confirm that the performance improvement of the proposed model is attributable to architectural design choices rather than dataset-specific overfitting or hyperparameter tuning advantages.

Table 1. Per-Class and Overall Performance Metrics – All Models on NSL-KDD Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	FPR @1%
Logistic Regression	78.4	77.1	76.3	76.9	0.862	0.41

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	FPR @1%
SVM (RBF kernel)	82.1	81.4	80.8	80.5	0.908	0.52
Random Forest	85.6	84.9	84.3	84.1	0.942	0.68
XGBoost	88.2	87.5	87.1	87.0	0.961	0.76
CNN only (ablation)	89.3	88.7	88.1	88.0	0.971	0.80
CNN-BiLSTM (ablation)	92.1	91.5	91.0	90.9	0.979	0.87
CNN-BiLSTM-Attention	**94.7**	**94.1**	**93.5**	**93.8**	**0.987**	**0.91**

FPR @1% = True Positive Rate at False Positive Rate of 1%; AUC-ROC = Area Under Receiver Operating Characteristic Curve; All metrics macro-averaged across four classes

4. Discussion

The results demonstrate that the CNN-BiLSTM-Attention architecture provides substantial and statistically significant performance improvements over both classical machine learning baselines and single-component deep learning variants for the network intrusion detection task. The hierarchical design of the proposed model maps naturally onto the multi-scale structure of network traffic data: the 1D-CNN layers extract local feature interactions within short windows of traffic attributes — analogous to the n-gram features that have proven effective in text classification — while the BiLSTM layers capture sequential dependencies across the full feature sequence, enabling the model to detect attack patterns that unfold over multiple consecutive feature positions. The self-attention mechanism provides a third complementary capability: direct, position-agnostic weighting of the most discriminative elements of the encoded sequence, allowing the model to focus on the small subset of features that are most predictive of each specific attack class regardless of their position in the input sequence.

The particularly strong performance on R2L detection (93.5% recall despite severe class imbalance in the original dataset) represents a significant practical advance over prior NIDS models that typically show degraded performance on low-frequency attack classes. This improvement is attributable to the combination of SMOTE oversampling during training and the attention mechanism's ability to identify the sparse, distinctive feature combinations characteristic of remote-to-local attack attempts — such as anomalous combinations of failed login counts, root shell attempts, and unusual service-flag combinations — even when these features occupy small portions of the feature sequence. The residual confusion between Benign and DoS classes at high traffic volumes reflects a fundamental challenge in NIDS: at sufficiently high legitimate traffic rates, statistical features computed over fixed time windows converge toward those of low-rate flooding attacks. Addressing this limitation through adaptive window sizing or online learning approaches represents a natural direction for future work.

The computational complexity of the proposed model warrants consideration for real-time deployment. The CNN-BiLSTM-Attention model contains approximately 287,000 trainable parameters, a modest parameter count relative to large-scale NLP or image models, and achieves inference throughput of approximately 14,200 instances per second on a single NVIDIA GTX 1660 GPU, corresponding to approximately 1.4 million network connections per minute. This throughput is sufficient for medium-scale enterprise network monitoring (typical enterprise networks generate 100,000–1,000,000 connections per hour) but may require optimisation — through model quantisation, pruning, or TensorRT compilation — for deployment at high-speed backbone network monitoring points handling tens of millions of connections per hour. The model's 94.7% accuracy and 0.987 AUC, combined with its real-time inference capability on modest hardware, establish it as a practically deployable solution for enterprise NIDS applications.

Comparison with recently published NIDS deep learning models in the literature reveals that the proposed model's accuracy of 94.7% on the NSL-KDD test set is competitive with the current state of the art. Yin et al. (2017) reported 81.2% accuracy using a unidirectional LSTM model on the same dataset; Vinayakumar et al. (2019) achieved 88.9% using a deep neural network ensemble; and recent transformer-based models have reported 92–96% accuracy with substantially higher parameter counts (≥5 million parameters). The proposed model achieves comparable accuracy with a fraction of the parameter count, suggesting that the architectural synergy between CNN, BiLSTM, and attention provides an efficient inductive bias for this domain that reduces the data and computational requirements relative to general-purpose transformer architectures.

5. Conclusions

This paper proposed and evaluated a hybrid CNN-BiLSTM-Attention deep learning architecture for multi-class network intrusion detection on the NSL-KDD benchmark dataset. The following conclusions are established:

(i) The proposed CNN-BiLSTM-Attention model achieves 94.7% overall classification accuracy, macro-averaged F1-score of 93.8%, and AUC-ROC of 0.987 on the NSL-KDD test set, outperforming logistic regression, SVM, random forest, and XGBoost baselines by margins of 16.3, 12.6, 9.1, and 6.5 percentage points in accuracy respectively.

(ii) Ablation studies confirm that both the BiLSTM component and the self-attention mechanism make statistically significant independent contributions to classification performance, adding 2.8 and 2.6 percentage points of accuracy respectively over the CNN-only baseline, validating the architectural design choices.

(iii) The model achieves a true positive rate of 91% at a false positive rate of 1%, a critical operational metric for real-world NIDS deployment, substantially exceeding all baseline models at this operating point and demonstrating practical relevance beyond aggregate accuracy metrics.

(iv) Inference throughput of approximately 14,200 instances per second on commodity GPU hardware confirms suitability for real-time enterprise network monitoring without specialised hardware infrastructure.

(v) Future work will investigate adversarial robustness of the model against evasion attacks, transfer learning across network environments to reduce labelled data requirements, and online learning strategies for adaptation to concept drift in evolving attack landscapes.

References

[1] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), pp. 1–6.

[2] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access, 5, 21954–21961.

[3] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525–41550.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press, Cambridge, Massachusetts.

[5] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

[8] Schapire, R. E., & Freund, Y. (2012). Boosting: Foundations and Algorithms. MIT Press, Cambridge.

[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[10] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[11] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[12] Kim, J., Kim, J., Thu, H. L. T., & Kim, H. (2016). Long short term memory recurrent neural network classifier for intrusion detection. International Conference on Platform Technology and Service (PlatCon), pp. 1–5.

[13] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR 2015).

[14] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

[15] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.