Tomek Links and Random Undersampling for Infiltration Attack Detection in Imbalanced CSE-CIC-IDS2018 Dataset

Sean William Fernandes
School of Computing
Telkom University
Bandung, Indonesia
seanwillf@student.telkomuniversity.ac.id

Aji Gautama Putrada
School of Computing
Telkom University
Bandung, Indonesia
ajigps@telkomuniversity.ac.id

Ryan Lingga Wicaksono
School of Computing
Telkom University
Bandung, Indonesia
wicaksonoryan@telkomuniversity.ac.id

Feddy Dea Reskyadita
School of Computing
Telkom University
Bandung, Indonesia
feddydr@telkomuniversity.ac.id

Abstract*—Several previous studies have used the CSE-CIC-IDS2018 dataset to develop intrusion detection systems (IDSs) with advanced models, such as long short-term memory (LSTM). However, these studies underscore the imbalance in the dataset, where the `'Benign'` label is much more common than some attack labels, such as `'Infiltration'`. This study aims to build and improve a model for detecting infiltration attacks using the CSE-CIC-IDS2018 dataset. The first step after obtaining the dataset is preprocessing, which involves applying a label encoder, performing principal component analysis (PCA) for dimension reduction, and normalizing the features. The next step is to apply undersampling and use the LSTM as a prediction model. We mitigate imbalance in the dataset by combining several undersampling methods, including Tomek links and random undersampling (RUS). Finally, receiver operating characteristic (ROC), area under the curve (AUC), and g-mean methods optimize the threshold for each prediction method. The test results show that PCA can improve the efficiency of the LSTM model, with the LSTM + PCA model being $1.7\times$ smaller and $1.1\times$ faster. Then Tomek links + RUS is proven to provide the best sensitivity and G-mean compared to the base model, namely 0.73 and 0.80, respectively. Finally, threshold optimization using the ROC curve further enhances the model's predictive performance, increasing its sensitivity to 0.82.

***Index Terms*—Tomek links, infiltration attacks, long short-term memory, principal component analysis, imbalanced dataset**

I. Introduction

An infiltration attack is a type of attack in which an attacker secretly gains persistent access to a system. After gaining access, the attacker can spy on, steal, or modify data in the system, often through several stages [1]. This type of cyberattack is one of the most difficult to detect. The impact of an infiltration attack is similar to that of other cyberattacks, including financial losses, data leaks, operational disruptions, and reputational damage. To counter this, an intrusion detection system (IDS) was created. IDS is an application or tool that is used to detect malicious cyber activity in a system [2]. Yonis [3] reviewed various security and classification techniques applicable to IDS, highlighting the effectiveness of IDS in identifying diverse attack behaviors.

The CSE-CIC-IDS2018 dataset is a benchmark dataset for IDS. Göcs *et al.* [4] used the dataset to conduct IDS, where the focus of this study was to find which features were most important for detection. This method uses the average of feature scores from several different approaches to determine the most important features. Although many studies have used it, the results are still suboptimal for detecting infiltration attacks. Bulavas *et al.* [5] said that the difficulty in detecting infiltration attacks in the dataset is due to the imbalance condition of the label, where the infiltration label is only 1% of the dataset. The study conducted an overlap analysis to improve the decision boundary.

Several studies have used undersampling methods, such as Tomek links and random undersampling (RUS), to address dataset imbalance. Thumpati *et al.* [6] combined Tomek links and RUS for a case study of an imbalanced bank marketing dataset. The study then compared the performance of models such as k-nearest neighbors (KNN), support vector machines (SVM), and logistic regression (LR). Elhassan *et al.* [7] combined Tomek links and RUS for a case study of imbalance in blood pressure and E. coli datasets. The machine learning models used in the study were SVM, LR, an artificial neural network (ANN), and a random forest (RF). There is a research opportunity to use Tomek links and RUS to address the imbalance problem in the `'Infiltration'` label in the CSE-CIC-IDS2018 dataset.

This research aims to build and improve a model for detecting infiltration attacks using the CSE-CIC-IDS2018 dataset. The first step in this testing, after obtaining the dataset, is to preprocess it, such as applying a label encoder, performing principal component analysis (PCA) dimension reduction, and normalizing the features. We then use a long short-term memory (LSTM) autoencoder as a predictive model. We mitigate imbalance in the dataset by combining several undersampling methods, including Tomek links and RUS. Finally, we use receiver operating characteristic (ROC), area under the curve (AUC), and G-mean to optimize the thresholds for each prediction method.

To the best of our knowledge, no research has yet applied Tomek Links + RUS to improve infiltration detection capabilities on CSE-CIC-IDS2018. The following are the contributions this research can make:

1) An efficient model for infiltration attacks on the CSE-CIC-IDS2018 dataset using PCA, with model effi-

ciency that is 1.7× smaller and 1.1× faster.
2) An infiltration attack detection using ROC curve-based threshold optimization to address imbalanced conditions in CSE-CIC-IDS2018 with optimal sensitivity or recall.
3) A hybrid method of Tomek Links + RUS that can address imbalance in the IDS dataset.

The remainder of this paper follows a systematic approach: Section II contains recent papers related to this research topic. Section III follows the steps of this research. Section IV tests and reports the proposed research methods. Finally, Section V contains a summary of the answers to our research objectives.

II. Related Works

Several studies have implemented IDS using CSE-CIC-IDS2018. Research by Andrecut *et al.* [8] shows that the use of a simple KNN classification method can produce comparable performance to more complex classification methods, such as CNN and LSTM, in building an IDS model. This research is relevant to the study to be conducted because it uses the same dataset, namely CSE-CIC-IDS2018. Research by Antunes *et al.* [9] tests and compares several deep learning methods for behavior-based IDS system models on the CSE-CIC-IDS2018 dataset. The feature reduction methods used are autoencoder and PCA to evaluate the model's performance in detecting attacks.

Several studies have used the hybrid Tomek links + RUS method for different case studies. Ruisen *et al.* [10] stated that Tomek links is a method to overcome overlapping labels. They utilized Tomek links + RUS to overcome the challenges of noisy data in Mandarin keyword detection. Bach *et al.* [11] used Tomek links and RUS on several imbalanced datasets, including the abalone dataset. This study also used several classifiers, including KNN, RF, and SVM. There is a research opportunity to use Tomek links + RUS on the imbalanced CSE-CIC-IDS2018 dataset.

Several studies have optimized thresholds using G-mean and ROC curves. Hancock *et al.* [12] stated that finding the best threshold in classification is a crucial step when using imbalanced datasets. The study demonstrated that selecting a threshold using G-mean achieves better performance than using the default threshold of 0.5. Johnson *et al.* [13] stated that threshold optimization can utilize the probability of positive labels. This study also noted that if the data is noisy, determining a non-optimal threshold will result in overfitting. There is a research opportunity to apply threshold optimization to imbalanced IDS datasets. Table I shows a comparison of the papers discussed in this section with our proposed approach.

TABLE I
Related Works Comparison on Imbalanced IDS Datasets

Reference	CSE-CIC-IDS2018 Dataset	Tomek Links + RUS	Threshold Optimization
Andrecut *et al.* [8]	✓	✗	✗
Antunes *et al.* [9]	✓	✗	✗
Ruisen *et al.* [10]	✗	✓	✗
Bach *et al.* [11]	✗	✓	✗
Hancock *et al.* [12]	✗	✗	✓
Johnson *et al.* [13]	✗	✗	✓
Proposed Method	✓	✓	✓

III. Research Method

To accomplish the goals of this study, we developed a research procedure. First, we obtain the CSE-CIC-IDS2018 dataset. Pre-processing the dataset, such as using a label encoder, MDI feature selection, and feature normalization, is the initial stage in this testing procedure after acquiring the dataset. The dataset is then divided into training and testing subsets using 80% training and 20% testing split. The next step is building and tuning a better LSTM model than state-of-the-art models. Next, we evaluate the PCA method for improving the LSTM model's efficiency. Tomek links + RUS undersampling are the next phases. We optimized each prediction method's threshold using ROC, AUC, and G-mean. Fig. 1 illustrates the steps involved in this study.

A. Infiltration Attack in The CSE-CIC-IDS2018 Dataset

The CSE-CIC-IDS2018 dataset is a benchmark IDS dataset produced by the Canadian Institute of Cybersecurity (CIC) at the University of New Brunswick (UNB) [14]. This dataset is not real network traffic data, but rather data generated in a controlled CIC laboratory environment [15]. The dataset creation process includes generating normal traffic data and various types of cyberattacks. The data is then captured in Packet Capture (PCAP) format and extracted using CICFlowMeter. The features extracted from this PCAP are stored in CSV format. The documentation for this dataset does not explicitly state why it is made unbalanced, but if we look at real-world network traffic data from Yang *et al.*'s research [16], it's showed that only 0.01% of attacks were successful, 27.37% were attempted, 48.91% were benign, and 23.72% were unknown, this implies that the CSE-CIC-IDS2018 imbalance reflects the real-world traffic imbalance.

An infiltration attack is a type of cyberattack in which the attacker unauthorizedly infiltrates a system, intending to steal classified information, in a way that remains undetected

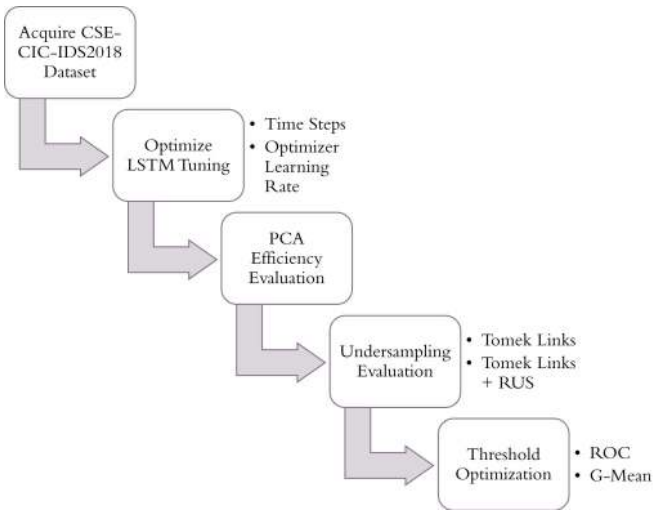


Fig. 1. The research workflow.

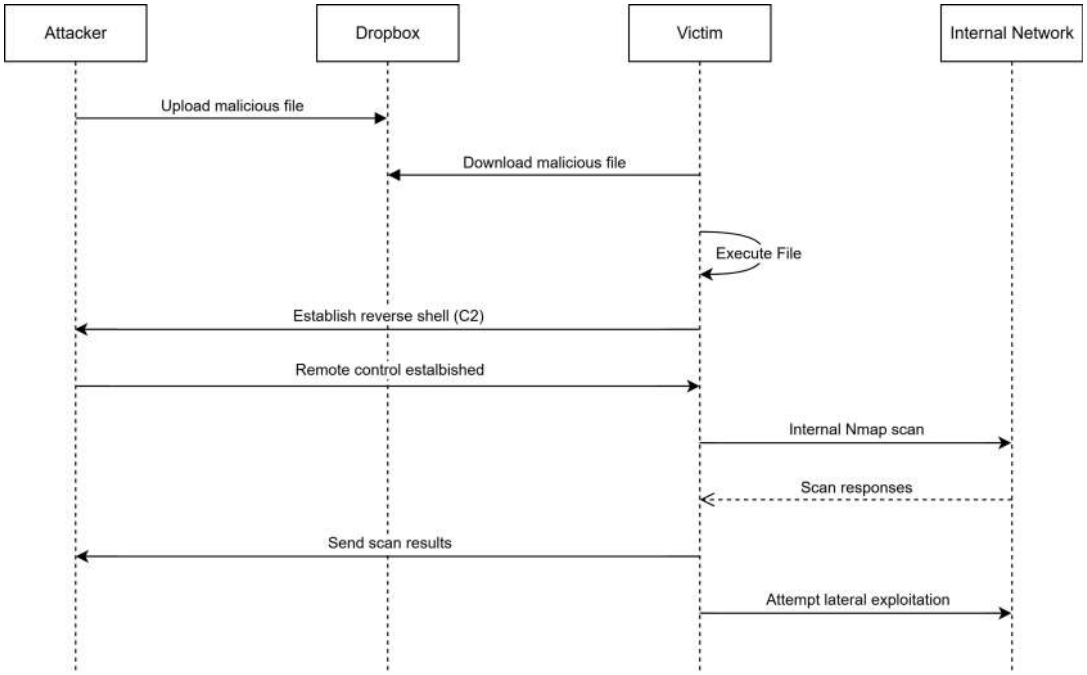


Fig. 2. The infiltration attack sequence diagram.

[17]. This method triggers several alerts. Fig. 2 shows a sequence diagram of how an infiltration attack works. The attacker uploads a malicious file to Dropbox. When the file is downloaded and executed by the victim, the malware within it activates and establishes a connection to the attacker, granting remote access to the device's shell. Through this access, the attacker can access the victim's device and explore its internal server. This can lead to lateral movement attacks and data exfiltration of sensitive information. This explains why infiltration attacks are so dangerous, as attackers can link persistent network compromises and large-scale data breaches [18].

B. LSTM with PCA Dimension Reduction

The curse of dimensionality occurs when a machine learning model is trained on a dataset with many features, making the dataset high-dimensional. PCA statistically reduces dimensionality while preserving as much variability as possible. Each principal component in PCA is a linear combination of several features, so that each principal component has no interrelationships ([19]). PCA utilizes a covariance matrix, eigenvalues, and eigenvectors to reduce dimension, with the formula as follows:

$$Z = XW_k \tag{1}$$

where Z is the new dataset resulting from dimension reduction, X is the original dataset, and W_k is a matrix containing the top k eigenvectors.

We use LSTM, a deep learning model that can capture sequential patterns in data, as a predictive model [20]. LSTM is an extension of deep learning that involves numerical operations. Therefore, before LSTM operations, one important preprocessing method is to normalize the dataset [21]. The next important step for LSTM to capture sequential patterns in the dataset is to transform it using a time window, which changes the dataset's shape from 2D to 3D [22]. The size of the new dataset $\text{Shape}_{\text{New}}$ is as follows:

$$\text{Shape}_{\text{New}} = N \times T \times d\times \tag{2}$$

where N is the number of windows, T is the timestep size, and d is the number of features. Previously, N was the result of the following calculation:

$$N = \left(\frac{L-T}{S}\right) + 1 \tag{3}$$

where L is the number of data items and S is the stride. Stride describes the shift in data with each new window formation.

C. Undersampling with Tomek Links + RUS

Another important characteristic of the CSE-CIC-IDS2018 dataset is its class imbalance, with the benign class accounting for 83% of the dataset. The rest is a mix of various cyberattacks [23]. This condition necessitates special handling of the dataset before it can be used for model training. Tomek links is an undersampling method that pairs data points from the minority and majority classes that are close to each other. This pair is called Tomek links. The next step is to remove the majority-class data in Tomek links to obtain a more balanced dataset [6]. With this concept, Tomek links clarify decision boundaries and reduce noise, thereby minimizing information loss that can degrade the accuracy of machine learning predictions.

Eliminating the dominance of majority labels in undersampling is important, while the label balance in Tomek links is closely related to the size of the decision boundary [24]. To further balance the data, we propose Tomek Links + RUS, which combines Tomek Links with RUS to randomly remove data from the majority class [25]. Fig. 3 illustrates

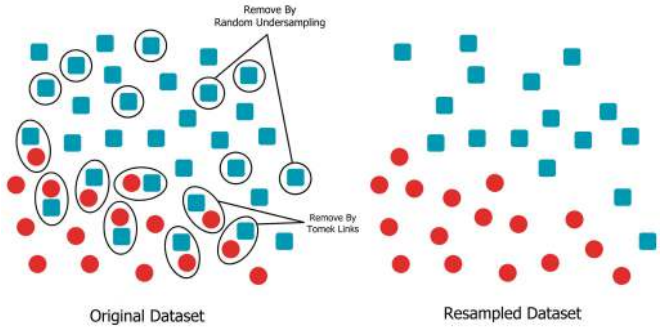


Fig. 3. The concept of Tomek links and RUS

the concept of Tomek links undersampling. The reason why we used Tomek Links + RUS and not SMOTE or ADASYN is because these oversampling techniques produce synthetic data that can be noisy and overfitting on large and extremely imbalanced datasets [26].

Decision making in machine learning is based on a threshold of the probability of what is considered true and what is considered false by the machine learning model [27]. In the case of an imbalanced dataset, the true positive rate (TPR) and false positive rate (FPR) of a prediction model become more sensitive to changes in the threshold [28]. Therefore, we use ROC curves to monitor these changes. We then use G-mean to determine the optimal threshold along the curve for each machine learning method. Algorithm 1 shows the threshold optimization method, where M is the LSTM model, then X_{test} and y_{test} are the test features and test labels. The softmax activation function of the LSTM outputs soft labels $y_{\text{pred_prob}}$. The rest of the algorithm calculates the TPR, FPR, and G-mean G of each threshold $\mathcal{T}$. The output of this algorithm is the optimal threshold t^*, which corresponds to the maximum G-mean G_{max}.

Algorithm 1 Threshold Optimization for LSTM using G-mean

Require: M, $(X_{\text{test}}, y_{\text{test}})$
Ensure: t^*
$y_{\text{pred_prob}} \leftarrow M.\text{predict}(X_{\text{test}})$
Initialize $G_{\text{max}} \leftarrow 0$, $t^* \leftarrow 0$
Define candidate $\mathcal{T} \leftarrow \{0.0, 0.01, \ldots, 1.0\}$
for each $t \in \mathcal{T}$ **do**
  $y_{\text{pred}} \leftarrow (y_{\text{pred_prob}} \geq t)$
  Compute TPR and FPR from $(y_{\text{test}}, y_{\text{pred}})$
  $G \leftarrow \sqrt{\text{TPR} \times (1 - \text{FPR})}$ {G-mean formula}
  if $G > G_{\text{max}}$ **then**
    $G_{\text{max}} \leftarrow G$
    $t^* \leftarrow t$
  end if
end for
return t^*, G_{max}

IV. Results and Discussion

A. Results

The dataset preparation for the IDS model begins with data loading. The dataset is downloaded from Kaggle, and we specifically read and analyze the 03-01-2018.csv file, which contains the infiltration attack data. After the data is loaded, rows containing missing data are removed. The data will also be cleaned of infinite values to prevent errors during training. Additionally, feature columns containing non-numeric data are converted to numeric, as the computational model can only accept numeric inputs.

The first step was to use and compare two baseline models from previous research: CNN and LSTM. LSTM performed better, with accuracy, sensitivity, specificity, and G-mean values of 0.86, 0.65, 0.94, and 0.77, respectively. It is important to highlight that the primary focus in prediction on an imbalanced dataset is on the minority label, where here it is the "`Infiltration`" label. Therefore, sensitivity, which predicts the positive label, becomes our primary metric. Based on these results, we adopted LSTM and extended it with methods to enhance infiltration attack performance using an imbalanced dataset.

Furthermore, we applied PCA to LSTM for dimension reduction. We chose PCA because it is more effective than autoencoders and feature selection. The performance of PCA on LSTM is 0.88, 0.64, 0.94, and 0.78 for accuracy, sensitivity, specificity, and G-mean, respectively. LSTM+PCA has lower sensitivity but the same G-mean as the original LSTM model, and PCA-based dimension reduction makes the model more efficient. Fig. 4 proves this efficiency claim, where we use the LSTM model size and LSTM prediction time as the testing metrics. The test results show that the model size and prediction time of the LSTM increase linearly with the number of PCA principal components. We use principal components $= 18$, determined based on empirical testing, resulting in a model size of 196.2 KB, compared to 340.2 KB for the original LSTM. Consequently, the prediction time is $64.2\ \mu s$ versus $70.2 \mu s$.

The next step involves applying the undersampling method to the infiltration attack detection model, specifically using two methods: Tomek links and Tomek links + RUS. We applied it to the PCA-based model because it is more efficient. Fig. 5 shows the performance results of the two undersampling methods. The Tomek links method can increase the model's sensitivity from 0.64 to 0.69. Then the G-mean also increased from 0.78 to 0.79. Furthermore, Tomek links

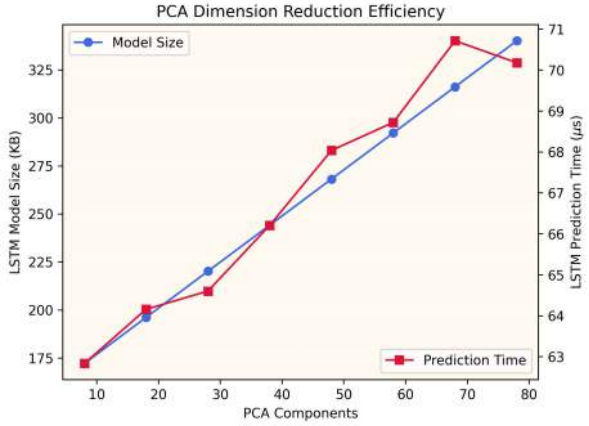


Fig. 4. The PCA efficiency towards the LSTM model.

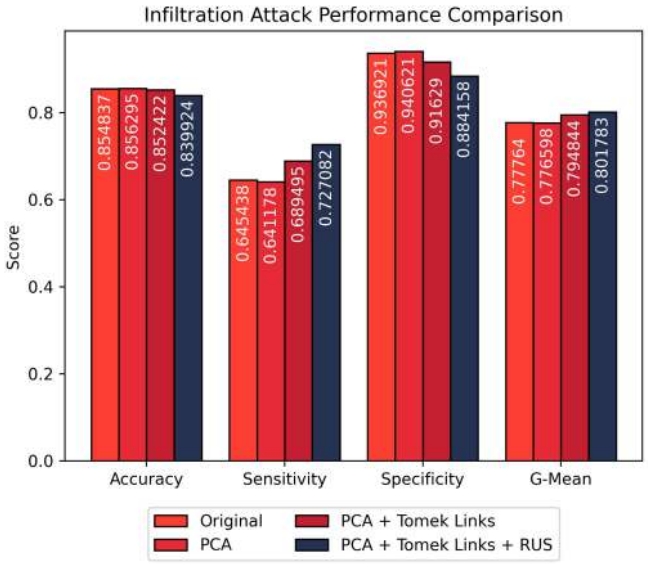


Fig. 5. The comparison of extended LSTM models on infiltration attack detection.

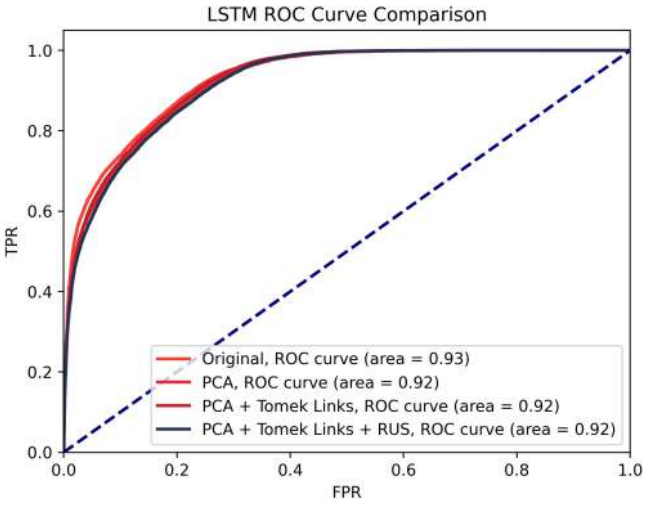


Fig. 6. The ROC comparison of extended LSTM models.

and RUS can improve the performance of the Tomek links method, namely from 0.69 to 0.73 for sensitivity and from 0.79 to 0.80 for G-mean. The conclusion is that Tomek links and RUS can improve the performance of infiltration attack detection on the imbalanced CSE-CIC-IDS2018.

In the next step, we optimized the model's performance by selecting a better threshold from the ROC curve rather than using the default 0.5. Fig. 6 shows a comparison of the ROC curves of the four methods we compared previously. Based on these curves, we found the optimal point between TPR and FPR using the G-mean. Of the four previous metrics we used, this threshold-optimization method improved two metrics —sensitivity and G-mean— for PCA + Tomek links + RUS. The sensitivity increased from 0.73 to 0.78, while the G-mean rose from 0.80 to 0.82. This indicates that threshold optimization using this ROC curve can balance sensitivity and specificity to achieve fairer performance.

TABLE II
COMPARISON WITH PREVIOUS RESEARCH RESULTS

Reference	Method	Threshold	Recall	F1-Score
Andrecut *et al.* [8]	CNN + PCA	Default	0.22	0.35
Andrecut *et al.* [8]	CNN + Autoencoder	Default	0.22	0.34
Andrecut *et al.* [8]	LSTM + PCA	Default	0.35	0.49
Andrecut *et al.* [8]	LSTM + Autoencoder	Default	0.31	0.44
Proposed Method	LSTM	ROC Curve Optimized	0.78	**0.81**
Proposed Method	LSTM + PCA	ROC Curve Optimized	0.79	**0.81**
Proposed Method	LSTM + PCA + Tomek Links	ROC Curve Optimized	0.80	**0.81**
Proposed Method	LSTM + PCA + Tomek Links + RUS	ROC Curve Optimized	**0.82**	0.80

B. Discussion

Several studies, such as by Archana *et al.* [29], have used LSTM+PCA for IDS and performed dimension reduction on IDS datasets with an overabundance of features. However, no study has tested the model's efficiency. This study's contribution is an efficient PCA-based model for infiltration attacks on the CSE-CIC-IDS2018 dataset, achieving $1.7times$ lower model efficiency and $1.1times$ faster inference.

Andrecut *et al.* [8] previously conducted a CSE-CIC-IDS2018 study with particular attention to "`Infiltration`" labels by adding PCA and Autoencoder to CNN and LSTM methods. The study showed that LSTM overall outperformed CNN, and PCA overall outperformed autoencoder. Meanwhile, Obiora *et al.* [30] stated that LSTM under-performance could be due to a lack of hyperparameter tuning. Our study shows that hyperparameter settings such as a softmax activation function, a 20-timestep window, an Adam optimizer with a learning rate of 0.001, and a binary cross-entropy loss metric can achieve higher LSTM performance than those used in that study. Table II compares the results. Then, imbalance treatment with Tomek links and Tomek links + RU on an LSTM with PCA yields precision, recall, and F1-score of 0.84, 0.80, and 0.81 for Tomek links, and 0.79, 0.82, and 0.80 for Tomek links + RU. The contribution of our research is an infiltration attack using ROC curve-based threshold optimization to overcome imbalanced conditions in CSE-CIC-IDS2018, achieving optimal sensitivity (recall).

Several studies, such as those by Thumpati *et al.* and [6], have combined Tomek links and RUS for their respective case studies, namely on an imbalanced bank marketing dataset. However, no study has addressed the imbalance problem in IDS using a hybrid undersampling method, Tomek links + RUS. Our research contribution is a hybrid Tomek links + RUS method that can address imbalance in the IDS dataset.

V. CONCLUSION AND FUTURE WORK

This study aims to build an anomaly-based IDS to detect Infiltration Attacks. This study explores the effectiveness of several methods to enhance LSTM-based infiltration attack detection: PCA for efficiency, Tomek links and RUS for

imbalanced datasets, and the ROC curve for threshold optimization. The test results show that PCA can improve the efficiency of the LSTM model, with the LSTM + PCA model being $1.7\times$ smaller and $1.1\times$ faster. Then Tomek links + RUS is shown to provide the best sensitivity and G-mean compared to the base model, with values of 0.73 and 0.80, respectively. Finally, threshold optimization using the ROC curve further enhances the model's predictive performance, increasing its sensitivity to 0.82. For future work, we can improve the model's efficiency, especially in edge computing environments, by applying pruning and quantization to the LSTM. We will also focus on testing with a real-world environment and more attack variations for the proposed model.

ACKNOWLEDGMENT

We want to thank the Communications Security Establishment (CSE) and the Canadian Institute of Cybersecurity (CIC) for creating the CSE-CIC-IDS2018 dataset. We would also like to thank the Directorate of Research and Community Service (PPM) of Telkom University for providing funding to conduct this research.

REFERENCES

[1] L. Hou, J. Han, B. Yang, Y. Guo, X. Wang, and G. Yang, "Research on attack trapping methods for multi-stage infiltration attacks," in *2023 International Conference on Computer Simulation and Modeling, Information Security (CSMIS)*. IEEE, 2023, pp. 556–560.

[2] F. Laghrissi, S. Douzi, K. Douzi, and B. Hssina, "Intrusion detection systems using long short-term memory (LSTM)," *Journal of Big Data*, vol. 8, no. 1, p. 65, dec 2021. [Online]. Available: https://doi.org/10.1186/s40537-021-00448-4

[3] A. Z. Yonis, "Network communication intrusion detection and classification security techniques," in *2022 IEEE Integrated STEM Education Conference (ISEC)*, 2022, pp. 455–459.

[4] L. Göcs and Z. C. Johanyák, "Identifying relevant features of cse-cic-ids2018 dataset for the development of an intrusion detection system," *Intelligent Data Analysis*, vol. 28, no. 6, pp. 1527–1553, 2024.

[5] V. Bulavas, "Improving machine learning model performance on detection of network infiltration," in *2021 14th International Conference on Human System Interaction (HSI)*. IEEE, 2021, pp. 1–6.

[6] A. Thumpati and Y. Zhang, "Towards optimizing performance of machine learning algorithms on unbalanced dataset," in *CS & IT Conference Proceedings*, vol. 13, no. 19. CS & IT Conference Proceedings, 2023.

[7] T. Elhassan, M. Aljurf *et al.*, "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method," *Global J Technol Optim S*, vol. 1, no. S1, 2016.

[8] M. Andrecut, "Attack vs benign network intrusion traffic classification," *arXiv preprint arXiv:2205.07323*, 2022.

[9] M. Antunes, L. Oliveira, A. Seguro, J. Veríssimo, R. Salgado, and T. Murteira, "Benchmarking deep learning methods for behaviour-based network intrusion detection," in *Informatics*, vol. 9, no. 1. MDPI, 2022, p. 29.

[10] L. Ruisen, D. Songyi, W. Chen, C. Peng, T. Zuodong, Y. YanMei, and W. Shixiong, "Bagging of xgboost classifiers with random under-sampling and tomek link for noisy label-imbalanced data," in *IOP Conference series: Materials science and engineering*, vol. 428, no. 1. IOP Publishing, 2018, p. 012004.

[11] M. Bach and A. Werner, "Improvement of random undersampling to avoid excessive removal of points from a given area of the majority class," in *International Conference on Computational Science*. Springer, 2021, pp. 172–186.

[12] J. Hancock, J. M. Johnson, and T. M. Khoshgoftaar, "A comparative approach to threshold optimization for classifying imbalanced data," in *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2022, pp. 135–142.

[13] J. M. Johnson and T. M. Khoshgoftaar, "Robust thresholding strategies for highly imbalanced and noisy data," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2021, pp. 1182–1188.

[14] M. Ghurab, G. Gaphari, F. Alshami, R. Alshamy, and S. Othman, "A detailed analysis of benchmark datasets for network intrusion detection system," *Asian Journal of Research in Computer Science*, vol. 7, no. 4, pp. 14–33, 2021.

[15] D. Stiawan, D. Wahyudi, T. W. Septian, M. Y. Idris, and R. Budiarto, "The development of an internet of things (iot) network traffic dataset with simulated attack data," *Journal of Internet Technology*, vol. 24, no. 2, pp. 345–356, 2023.

[16] L. Yang, Z. Chen, C. Wang, Z. Zhang, S. Booma, P. Cao, C. Adam, A. Withers, Z. Kalbarczyk, R. K. Iyer, and G. Wang, "True attacks, attack attempts, or benign triggers? an empirical measurement of network alerts in a security operations center," in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 1525–1542. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/yang-limin

[17] A. L. Krall, M. E. Kuhl, S. F. Moskal, and S. J. Yang, "Assessing the likelihood of cyber network infiltration using rare-event simulation," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–7.

[18] V. Malik, A. Khanna, N. Sharma, and S. Nalluri, "Advanced persistent threats (apts): Detection techniques and mitigation strategies," *International Journal of Global Innovations and Solutions (IJGIS)*, August 2024.

[19] S. Xie, H. Lin, T. Ma, K. Peng, and Z. Sun, "Prediction of joint roughness coefficient via hybrid machine learning model combined with principal components analysis," *Journal of Rock Mechanics and Geotechnical Engineering*, vol. 17, no. 4, pp. 2291–2306, 2025.

[20] A. G. Putrada, N. Alamsyah, I. D. Oktaviani, and M. N. Fauzan, "Lstm for web visit forecasting with genetic algorithm and predictive bandwidth allocation," in *2024 International Conference on Information Technology Research and Innovation (ICITRI)*. IEEE, 2024, pp. 53–58.

[21] A. G. Putrada, N. Alamsyah, M. N. Fauzan, and S. F. Pane, "Ns-svm: Bolstering chicken egg harvesting prediction with normalization and standardization," *JUITA: Jurnal Informatika*, pp. 11–18, 2023.

[22] A. G. Putrada, I. D. Oktaviani, M. N. Fauzan, and N. Alamsyah, "Cnn-lstm for mfcc-based speech recognition on smart mirrors for edge computing command," *Journal of Dinda: Data Science, Information Technology, and Data Analytics*, vol. 4, no. 2, pp. 63–74, 2024.

[23] J. L. Leevy and T. M. Khoshgoftaar, "A survey and analysis of intrusion detection models based on cse-cic-ids2018 big data," *Journal of Big Data*, vol. 7, no. 1, p. 104, 2020.

[24] S. Munawar, N. Javaid, Z. A. Khan, N. I. Chaudhary, M. A. Z. Raja, A. H. Milyani, and A. Ahmed Azhari, "Electricity theft detection in smart grids using a hybrid bigru–bilstm model with feature engineering-based preprocessing," *Sensors*, vol. 22, no. 20, p. 7818, 2022.

[25] R. Zuech, J. Hancock, and T. M. Khoshgoftaar, "Detecting web attacks using random undersampling and ensemble learners," *Journal of Big Data*, vol. 8, no. 1, p. 75, 2021.

[26] T. Wongvorachan, P. Meesad, and K. Kerdprasop, "A comparison of undersampling, oversampling, and hybrid resampling for imbalanced classification," *Information*, vol. 14, no. 1, p. 54, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/1/54

[27] A. G. Putrada, I. D. Oktaviani, M. N. Fauzan, and N. Alamsyah, "Quantifying student model accuracy via sample size and soft label mse: Insights into label quality," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*. IEEE, 2024, pp. 1–6.

[28] A. G. Putrada, N. Alamsyah, S. F. Pane, M. N. Fauzan, and D. Perdana, "Auc maximization for flood attack detection on mqtt with imbalanced dataset," in *2023 International Conference on Information Technology Research and Innovation (ICITRI)*. IEEE, 2023, pp. 133–138.

[29] J. Archana and A. Aneetha, "Lstm-mi: Revolutionizing intrusion detection through adaptive learning and mutual information analysis," in *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE, 2024, pp. 1–7.

[30] C. N. Obiora, A. Ali, and A. N. Hasan, "Finding the optimum horizon for short-term solar irradiance forecasting using long short-term memory (lstm) network," in *2022 11th International Conference on Power Science and Engineering (ICPSE)*. IEEE, 2022, pp. 148–152.