SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Detection Model for Ambiguous Intrusion using SMOTE and LSTM for Network Security

Al-Ogaidi Ali Hameed Khalaf[1], Raihani Mohamed[1,*], Abdul Rafiez Abdul Raziff[2]

[1] Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[2] Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, 50728 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 27 August 2023
Received in revised form 30 December 2023
Accepted 17 January 2024
Available online 12 February 2023

Keywords:
Intrusion detection; SMOTE; LSTM; imbalance dataset

ABSTRACT

In today's interconnected world, networks play a crucial role. Consequently, network security has become increasingly vital. To ensure network security, various methods are employed, including digital signatures, firewalls, and intrusion detection. Among these methods, intrusion detection systems have gained significant popularity due to their ability to identify new attacks. However, the accuracy of these systems still requires further improvement. One of the challenges is the potential bias introduced by using imbalance datasets that contains more information on normal activities than on attacks. To address it, SMOTE method was proposed and additionally, the study explores the use of Long Short-Term Memory (LSTM) for classification purposes. The experiments are conducted using two datasets: UNSW NB-15 and CICIDS 2017. The results obtained demonstrate that the proposed methods achieve an accuracy of 96% with the UNSW NB-15 dataset and 99% with the CICIDS 2017 dataset. These findings indicate an improvement of 3% and 1% respectively compared to existing literature.

1. Introduction

The widespread use of network-based connectivity to meet users' needs creates the potential for major network assaults. The most efficient method for detecting such an attack is an Intrusion Detection System (IDS). IDS identified as a monitoring system that distinguishes suspicious activity within the network's dataflow and sends out alarms, with the association of Security-Operations-Centre predictor or incident responder can analyze the problem and proceed the required steps to eliminate the danger grounded on these reports [1-3].

The forms of IDS will be discussed in this literature review, particularly machine learning-based IDS. The identification of known and unknown threats is a significant study field in IDS technology. Moreover, the most prevalent machine learning IDS systems are mentioned in this survey to discuss the IDS criteria for flawless detection as well [4-6]. In order to find the perfect solution for intruders, the academic and industrial researchers contributed with many techniques to enhance the

* *Corresponding author.*
E-mail address: raihanimohamed@upm.edu.my

https://doi.org/10.37934/araset.39.2.191203

mechanism that handles the security for the network [7-9]. A plethora of these techniques centric around Artificial Intelligent (AI), Machine Learning (ML), and Deep Learning (DL), to come with a reliable and effective solution in the aspect of network security prevention and surveillance as will be reviewed in the next part of this report [10-13]. In addition, different classification methods are used to identify distinct sorts of malicious intrusion within the network in order to improve the efficiency of IDS. Choosing an appropriate classification algorithm for IDS development is a difficult undertaking [14]. The datasets used to make IDS are usually imbalanced due to the fact that the data available in normal activities exceeds the data available for attacks. Imbalanced data causes a bias to the ML model. Most of the related work addressed the bias issue by creating ensembled learning algorithms [15,16]. This study investigates a resampling technique to balance the dataset, namely, SMOTE [17].

The rest of the paper is organized as follows; section II presents the literature review. Section III details the framework and model of the study. The implementation results are presented in section IV and the conclusion is in section V.

2. Literature Review

Over the past few years, various researchers conducted their research with the aim to enhance the intrusion detection systems (IDSs), in order to detect and prevent malicious agents from unauthorized access to data that is sensitive in nature [3,18]. In this section, a discussion and an examination of various techniques that leverage on machine learning algorithms for classification purposes is presented. This includes the pre-processing of data and the features selection as well, in addition to how many features are selected. Moreover, classification algorithms will be discussed and finally, algorithms that provide evaluation metrics.

For an appropriate feature selection method, a Hybrid-Feature-Selection-Algorithm (HFSA) was developed [4]. HFSA improved a set of the utmost related features that were utilized toward develop multi-sorting classifiers. The Jpcap library is used to collect real-time packets, which are used in this model. The Naive Bayes classification technique is used to distinguish between benign and malicious assaults. There are two steps to the preprocessing phase. To begin, data transformation is the process of converting symbolic data into a numerical value. Second, features are scaled from the largest range to the smallest range between (0,1) during the data normalization step, and each record is normalized. Then, using Nave Bayes, feature selection is used to detect six different types of attacks: standard, (R2L), (U2R), (DoS), (Probe), and (Brute Force). (HFSA) is used to improve the categorization system by updating it. Over-all, the proposed method achieved a correctness percentage of (92), with a precision of 95 percent.

Previously, an IDS based on "Naive Bayes" and a classification technique SVM was introduced [5]. Barely (24 out of 42) features in the (NSL-KDD) data set were chosen using the correlation subset method of feature selection. In addition, the properties are transformed to binary values and data normalization is performed during data preparation. The experimental findings show that the Support Vector Machine algorithm is the top method to use in classification, with a total of (93.95) percent as accuracy ratio, when compared to the (Naive Bayes) classifier.

A novel supervised approach for classifying and analyzing network data in order to detect malicious assaults using Artificial Neural Network (ANN) and (SVM) methods [6]. Filter method-based Chi-Square and wrapper method-based Correlation were employed for feature selection in both types of feature selection. The training model was based on the (NSL-KDD) data set, which had (25,191) entries. Out of (41) characteristics, the technique uses the Correlation-based wrapper method, with (17) of them being more important. In contrast, a chi-square-based filter is used to

choose 35 characteristics that are more informative and important for the training model stage. In comparison to all other strategies, the performance of the proposed ANN and by picking 17 characteristics achieves the maximum of (94.02 percent) for accuracy-wise, according to the testing data.

Another work suggested a unique feature selection and categorization approach Hybrid Anomaly-based Intrusion Detection System (HAIDS) combining Regression Trees with Random Forest (RF) [7]. The proposed model denoted as HAIDS. Rather of using a single method, the hybrid technique is employed to increase the model's performance. Moreover, to reduce the high dimensionality, the method of eliminating redundant characteristics is employed. The suggested approach was used to choose the highest thirteen characteristics from the UNSW-NB15. With a wrong aware proportion of (11.86 percent) and a rate of accuracy as (87.74 percent), the hybrid technique had the best performance in terms of accuracy [8,9].

In particular, IDS, biometrics, and healthcare are frequently challenged by the problem of class imbalance [3]. Sample quantity imbalances between classes can have a negative effect on a model's accuracy. It is critical to appreciate the complexities of this issue and determine if it constitutes a simple or complex instance of class inequality. Noisy data refers to situations when the bigger class crosses with the smaller class samples, decreasing accuracy. The disparity in sample sizes between the majority and minority classes is a key issue related to class inequality [17,19]. When class proportions are significantly skewed and exhibit a significant overflow of samples from some classes relative to others, a dataset is said to be imbalanced. These abnormalities typically show in real-world settings, particularly within biometrics, gene recognition, and medical datasets [3]. In binary classification scenarios involving two classes, the majority class pertains to the negative observations that outnumber the positive (minority) observations, or vice versa. These situations pose difficulties for classification models, as they grapple to attain heightened accuracy unless the input data is suitably handled prior to the classification phase.

3. Methodology

3.1 Dataset

This study uses two datasets: UNSW NB-15 and CICIDS 2017 [12,20].

(i) UNSW NB-15 Dataset

The UNSW-NB15 dataset consists of raw network packets generated by the IXIA PerfectStorm tool in the Cyber Range Lab of UNSW Canberra [12]. Its purpose was to create a combination of genuine modern activities and synthetic attack behaviours. To capture the raw traffic, the tcpdump tool was employed, resulting in a collection of 100 GB of data stored in Pcap files. This dataset encompasses nine distinct types of attacks, namely Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. In order to generate a comprehensive set of features with their corresponding class labels, the Argus and Bro-IDS tools were utilized. These efforts resulted in the development of twelve algorithms, which collectively produced 49 features.

(ii) CICIDS 2017 Dataset

The CICIDS2017 dataset comprises both benign network traffic and a comprehensive collection of common attacks that closely resemble real-world data captured in PCAPs [20]. The dataset includes the outcomes of network traffic analysis conducted using CICFlowMeter, where flows are

labelled based on various attributes such as time stamp, source and destination IPs, source and destination ports, protocols, and attack types. These labelled flows are stored in CSV files. Furthermore, the dataset provides a definition for the extracted features. This definition offers insights into the specific attributes and characteristics that have been extracted and included in the dataset, allowing for a more comprehensive understanding of the data.

3.2 Proposed Framework

Figure 1 shows the framework used with the UNSW NB-15 and CICIDS 2017 dataset. For the UNSW NB-15 dataset, the csv file is read. The data description reveals that there are extreme values in this dataset that need to be pre-processed. Hence, we apply clamping to extreme values. Additionally, we perform a reduction of the cardinality of the features. This operation is performed on the features of categorical type (e.g., string). The cardinality means the number of unique values present in a feature. A very high number of unique values makes the dataset instances have distinct values in that features and makes the classifier used later unable to make sense of the feature value as each instance exhibits a different value. Narrowing down the number of unique features helps with data understanding and classification. Following that, the labels are encoded to numbers. Next, the features are evaluated to examine the importance of each feature. Then the most relevant features are selected. At this point, we come to the classification step where we strain the LSTM model using a portion of the dataset, namely, the training set. Then, the testing set is used to test the model. And the performance is calculated using the evaluation metrics.

With the CICIDS Dataset, the 8 csv files are imported. These files have to be merged into a single file to be ready for further processing. After merging the instances of the dataset into a single file, the data description shows that the dataset is very unbalanced. And needs proper handling for this issue. If not handled, the classifier may end up being biased toward the majority class, that is the class with the greatest number of instances. To handle the data imbalance, we choose to apply synthetic minority oversampling technique (SMOTE) to increase the number of instances in the minority class and make the dataset almost balanced.

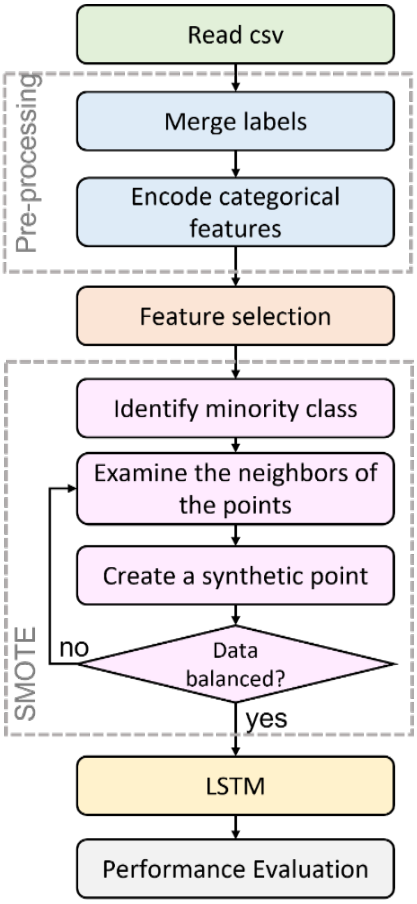


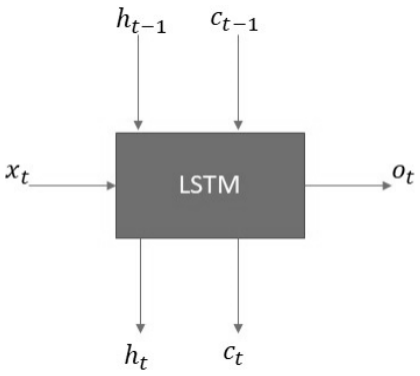
Fig. 1 the proposed framework

SMOTE is though applied after the feature selection step to reduce the number of features first and hence reduce the computational time of the SMOTE as it can be computationally expensive with large datasets [17]. SMOTE works as follows:

(i) Identify the minority class: SMOTE assumes that one class is underrepresented or has fewer instances compared to the other class.
(ii) Select a minority class instance: Randomly pick an instance from the minority class as the starting point for generating synthetic samples.
(iii) Identify the k nearest neighbors: Calculate the Euclidean distance between the selected instance and all other minority class instances. Choose the k nearest neighbors based on this distance metric.
(iv) Generate synthetic samples: For each selected instance, choose one of its k nearest neighbors randomly. Calculate the difference between the feature values of the instance and its selected neighbor. Multiply this difference by a random number between 0 and 1 and add it to the selected instance's feature values. This process creates a new synthetic instance along the line segment connecting the two instances in the feature space.
(v) Repeat steps 2-4: Repeat the process until the desired number of synthetic samples has been generated or until the minority class is balanced with the majority class.

By using SMOTE, the algorithm creates synthetic samples that capture the underlying distribution of the minority class. This helps to overcome the class imbalance problem, enabling better performance of machine learning models by providing a more representative training dataset. At this

point, the data is ready to be passed to the LSTM for training [20]. After that, the dataset is tested and the evaluation metrics are used to evaluate the performance of the model. For the classification model as mentioned earlier, LSTM is used. The LSTM is a recurrent structure. It employs gates to regulate information flow in recurrent computations. The LSTM's gating mechanisms keep the network's long-term dependencies intact. The sketch and the equations below illustrate the working process of the LSTM gates.



$$f_t = \sigma_g\,(W_f \times\ x_t + U_f \times h_{t-1} + b_f)$$

$$i_t = \sigma_g\,(W_i \times\ x_t + U_i \times h_{t-1} + b_i)$$
$$o_t = \sigma_g\,(W_o \times\ x_t + U_o \times h_{t-1} + b_o)$$
$$c'_t = \sigma_c\,(W_c \times\ x_t + U_c \times h_{t-1} + b_c)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t$$
$$h_t = o_t \cdot \sigma_c(c_t)$$

$f_t\ is\ the\ forget\ gate$
$i_t\ is\ the\ input\ gate$
$o_t\ is\ the\ output\ gate$
$c_t\ is\ the\ cell\ state$
$h_t\ is\ the\ hidden\ state$
$\sigma_g\ is\ the\ sigmoid\ function$
$\sigma_c\ is\ the\ tanh\ function$

4. Implementation

(i) UNSW NB-15 Implementation

The implementation starts with Data import and Preprocessing, this includes importing the dataset, understanding the structure of the data, and performing transformation on the data toward a more normal distribution. Next, we perform feature selection, this involves understanding the features importance and selecting the most relevant ones. Figure 2 shows the importance of the top 20 features.

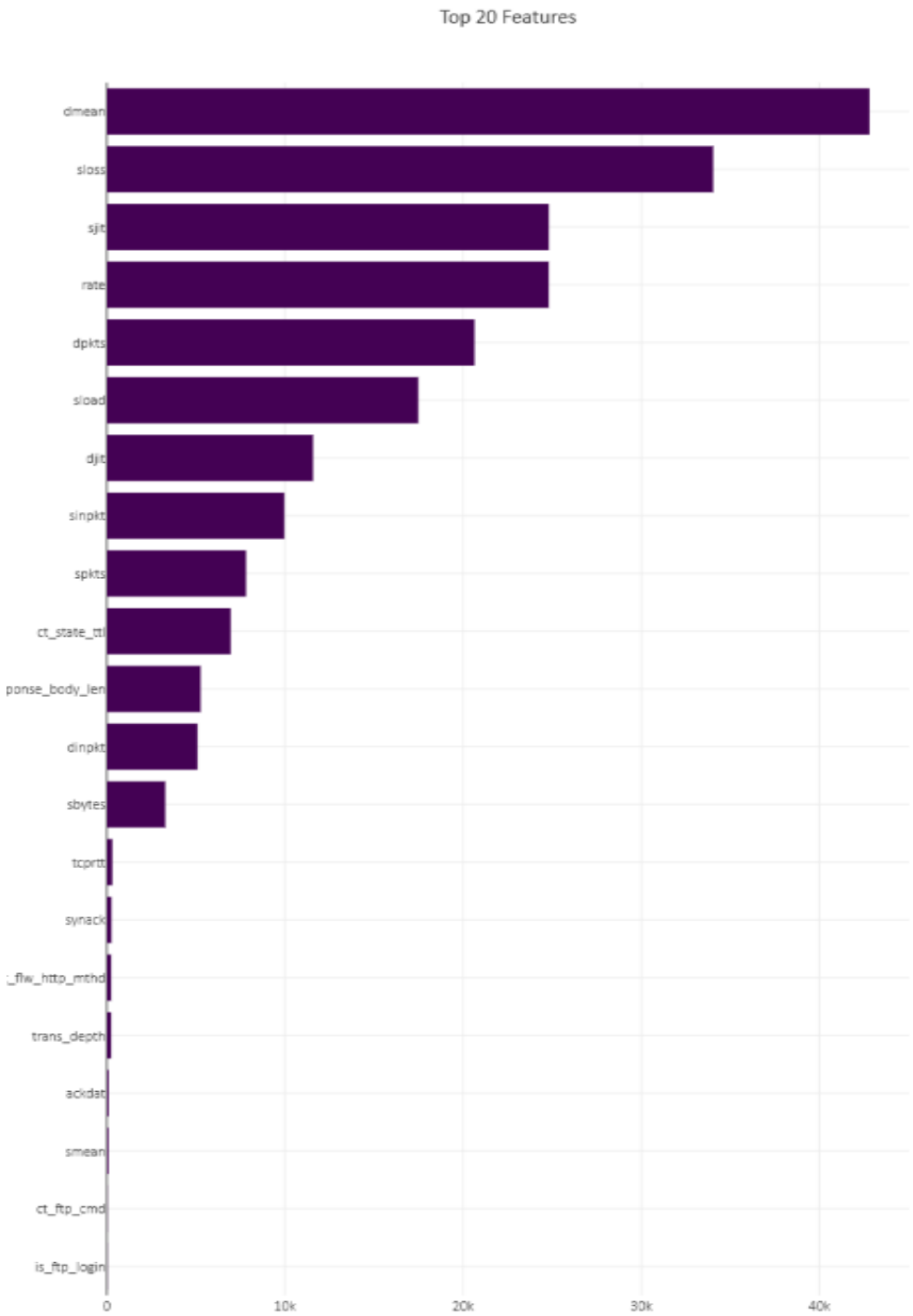


Fig. 2. Importance of top 20 features in UNSW NB-15 dataset

After feature selection, SMOTE resampling is applied, and the dataset instances count before and after are shown in Figure 3. This dataset does not have a major imbalance problem, hence, the before and after does not show a major difference.

The next step is LSTM classification where the LSTM is trained and tested. The results of this phase are shown in the next section.

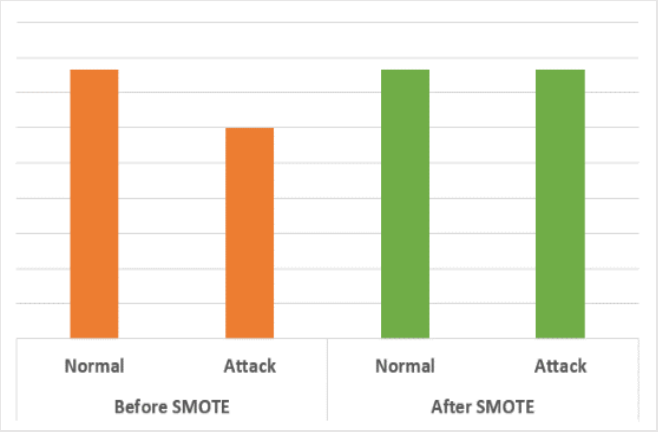


Fig. 3. UNSW NB-15 dataset instances count before and after SMOTE

(ii) CICIDS 2017 Implementation

The implementation starts with data import and Preprocessing, this includes importing the dataset, merging the files and merging the labels to obtain a binary classification. The description of the dataset shows that it has several attack classes, these classes are merged to obtain binary classification (normal and attack). The classes are also imbalances, Figure 4 shows the class distribution.

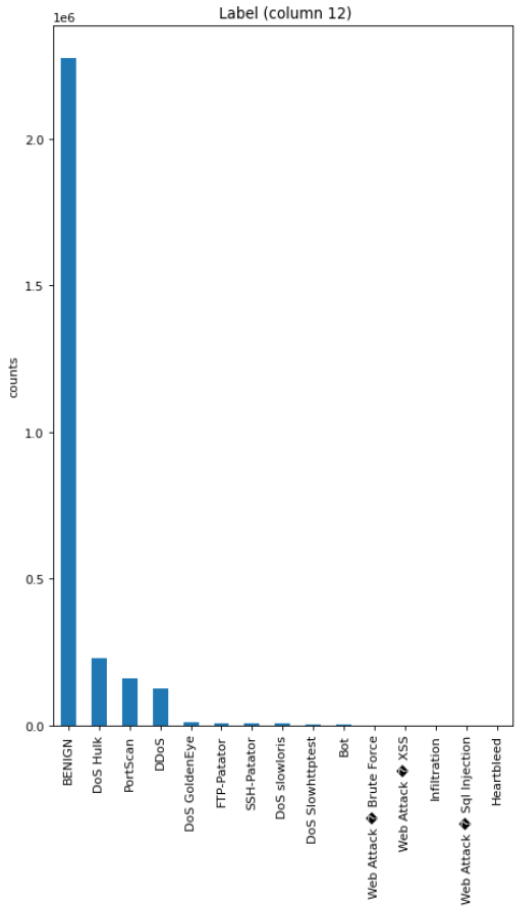


Fig. 4. CICIDS 2017 class distribution

Feature selection is performed before moving to balancing the dataset with SMOTE. this involves understanding the features importance and selecting the most relevant ones. The importance of the features of the dataset is shown in Figure 5.

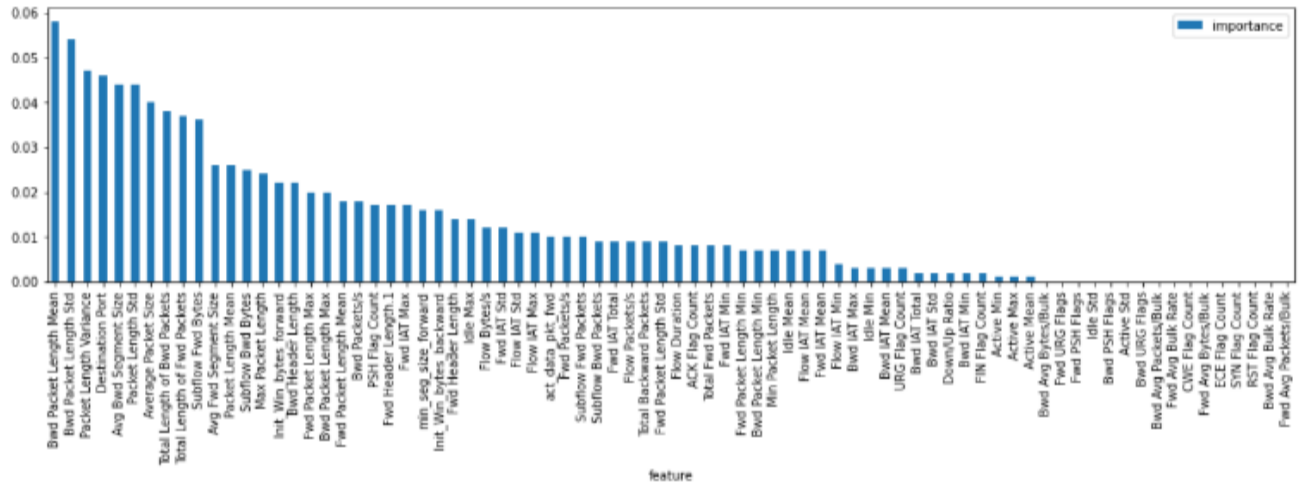


Fig. 5. CICIDS 2017 feature importance

After that, SMOTE resampling is applied, and the dataset instances count before and after are shown in Figure 6.

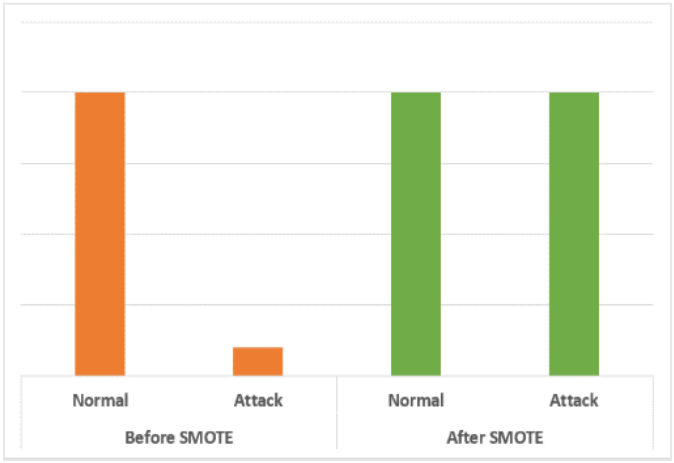


Fig. 6. CICIDS 2017 dataset instances count before and after SMOTE

The figure shows a big difference in the number of instances in the attack class before and after SMOTE. The next step is LSTM classification where the LSTM is trained and tested. The results of this phase are shown in the next section.

(iii) Simulation Parameters

The LSTM model involves some parameters that must be set when building and compiling the model. Using both datasets, the model used is LSTM. Table 1 shows the simulation parameters used with both datasets. For the first dataset, UNSW NB-15, two LSTM layers were added with 20 cells in each and one dense layer with ten neurons in addition to the classification layer that is a dense layer with two neurons that is equivalent to the number of classes. For this dataset, there is no dropout

layer in the model because there were no overfitting problems experienced. The optimization algorithm used is Adam and the network was trained for 200 epochs.

With the CICIDS 2017 dataset, the LSTM model is used with a single LSTM layer with 30 cells and a dropout layer with a dropout rate of 0.2. The dropout serves in elimination random features at each round to introduce more randomness and reduce the effect of overfitting. The model involves a single dense layer that is the classification layer with two neurons. Each neuron represents an output class. The model was compiled using Adam classifier and trained for 200 epochs.

Table 1
Simulation parameters

Network Parameters	UNSW NB-15	CICIDS 2017
Network type	LSTM	LSTM
Number of LSTM layers	2	1
Number of LSTM cells	20	30
Dropout layer	No	Yes
Number of dense layers	1	1
Number of neurons in dense layer	10	2
Optimization algorithm	Adam	Adam
Epochs	200	200

5. Results and Discussion

(i) Results of the UNSW NB-15 Dataset

The results obtained with this dataset have reached an accuracy of 96.47%. Other metrics are used besides the accuracy that are precision, recall, and F-measure. The results obtained with this dataset have shown a similar value for all the accuracy, precision, recall, and F-measure, that is 96.47%. Figure 7 illustrates the four metrics values.

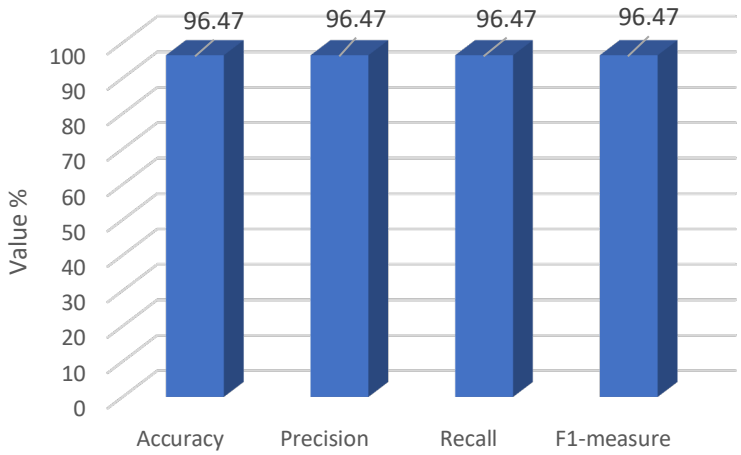


Fig. 7. Results of the UNSW NB-15 dataset

(ii) Results of the CICIDS 2017 Dataset

The CICIDS 2017 dataset is much larger in terms of the number of instances compared to the UNSW NB-15 dataset. A large dataset might contribute to a better training of the LSTM model as DL models generally benefit from the data abundancy. The accuracy obtained with this dataset is 99.5%.

Calculating the other evaluation metrics, the precision value is 98.7%, a value of 99.4% in the recall and 99.2% in the F-measure. Figure 8 illustrates the results of the four metrics.

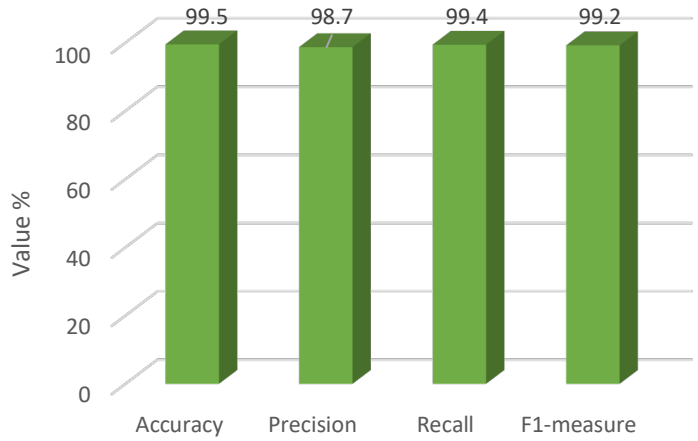


Fig. 8. Results of the CICIDS 2017 dataset

(iii) Comparison and Discussion

First, we compare the results of the two datasets with and without SMOTE to figure out the effect of the SMOTE resampling. Table 2 shows that the accuracy of the UNSW NB-15 has not changed with and without SMOTE. Which indicates that SMOTE does not make any difference with this dataset. this can be interpreted by the fact that this dataset does not have a major imbalance problem in the first place. So, the SMOTE did not have a room to do much in balancing the dataset. For the CICIDS 2017 dataset, the difference in the accuracy is clear. With a value of 93.7% without SMOTE and 99.5% with SMOTE. By subtracting the two values to get the difference in the accuracy we find that there is an increase of almost 6% in the accuracy. This is interpreted by the fact that this dataset has a major imbalance problem. Carrying out with the classification without balancing the dataset affects the classification process negatively. Balancing the dataset gives the classifier the chance to learn equally from both datasets and produce more accurate predictions.

Table 2
Proposed Framework Result with and without SMOTE

	Without SMOTE	With SMOTE
UNSW NB-15	96.47%	96.47%
UNSW NB-15	93.7%	99.5%

We also perform here a comparison between the base work re-implementation and the proposed method results in terms of accuracy, precision, and recall. In the re-implementation of the basework. Table 3 summarizes the obtained results. Comparing the SVM-NB (re-implementation) and the proposed method, the accuracy obtained using UNSW NB-15 dataset is 93.75% [12]. While our proposed method has achieved an accuracy of 96.47%. That is an improvement of almost 3%.

Using the CICIDS2017 dataset, the base work re-implementation has given an accuracy of 98.9%. The proposed method of this study has reached an accuracy of 99.5%. Which is an improvement of 0.6%. A comparison of the proposed method LSTM with LSTM with CNN features shows that the results for the UNSW NB-15 dataset are 93.8% and 96.47% with the LSTM with CNN features and the proposed LSTM with SMOTE respectively [20]. With the CICIDS 2017 dataset the results are 96.1%

and 99.5% with the LSTM with CNN features and the proposed LSTM with SMOTE respectively. This shows that the proposed methods have effectively improved the results over the base work.

Table 3
Summary of the re-implementation and the proposed methods results

		Accuracy	Precision	Recall
SVM-NB (Re-implementation) [12]	UNSW NB-15	93.75%	94.73%	94.48
	CICIDS	98.9%	98.4%	99.3%
LSTM with CNN feature (without SMOTE) [20]	UNSW NB-15	93.8%	94.5%	94.5%
	CICIDS	96.1%	96.3%	96.2%
LSTM with SMOTE (Proposed method)	UNSW NB-15	96.47%	96.47%	96.47%
	CICIDS	99.5%	98.7%	99.4%

For the UNSW NB-15 dataset, the pre-processing might have played a significant role in improving the quality of the data by normalizing the values and dealing with the extreme values and processing the data toward a nearly normal distribution. Another reason might be the use of the LSTM model that is a DL model. DL models are known to be effective with complicated datasets. For the CICIDS dataset, the SMOTE might have played a major role in reducing the bias on the classification model. Especially when using a DL model such as LSTM, these models are prone to be biased with very unbalanced data. Hence, balancing the data using SMOTE and using an LSTM model have both contributed to giving a good performance.

6. Conclusion

This study has proposed an intrusion classification model using two datasets, UNSW NB-15 and CICIDS 2017. The study framework of the UNSW NB-15 dataset involves preprocessing the data and dealing with extreme values to transform the data distribution to a nearly normal distribution. An LSTM model is used after that for classification. For the CICIDS 2017 dataset, due to the fact that this dataset is unbalanced and most of the instances fall in the BENIGN class, the SMOTE is used to balance the dataset and avoid model bias. An LSTM model is then trained using this dataset. The evaluation of the trained models of both dataset and a comparison with the base work shows that there is an improvement of 3% for the UNSW NB-15 dataset over the base work, and an improvement of 0.6% for the CICIDS dataset.

References

[1] Lazarevic, Aleksandar, Vipin Kumar, and Jaideep Srivastava. "Intrusion detection: A survey." *Managing Cyber Threats: Issues, Approaches, and Challenges* (2005): 19-78. https://doi.org/10.1007/0-387-24230-9_2

[2] Taghavinejad, Seyedeh Mahsan, Mehran Taghavinejad, Lida Shahmiri, Mohammad Zavvar, and Mohammad Hossein Zavvar. "Intrusion detection in IoT-based smart grid using hybrid decision tree." In *2020 6th International Conference on Web Research (ICWR)*, pp. 152-156. IEEE, 2020. https://doi.org/10.1109/ICWR49608.2020.9122320

[3] Liang, Xiaomin, Daifeng Li, Min Song, Andrew Madden, Ying Ding, and Yi Bu. "Predicting biomedical relationships using the knowledge and graph embedding cascade model." *PLoS One* 14, no. 6 (2019): e0218264. https://doi.org/10.1371/journal.pone.0218264

[4] Ariffin, Noor Afiza Mohd, and Vanitha Paliah. "An Improved Secure Authentication in Lightweight IoT." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 31, no. 3 (2023): 191-207. https://doi.org/10.37934/araset.31.3.191207

[5] Mohamed, Raihani, Thinagaran Perumal, Md Nasir Sulaiman, Norwati Mustapha, and Mohd Norhisham Razali. "Conflict resolution using enhanced label combination method for complex activity recognition in smart home environment." In *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pp. 1-3. IEEE, 2017. https://doi.org/10.1109/GCCE.2017.8229477

[6] Zainudin, MN Shah, Md Nasir Sulaiman, Norwati Mustapha, Thinagaran Perumal, and Raihani Mohamed. "Recognizing complex human activities using hybrid feature selections based on an accelerometer sensor." *International Journal of Technology* 8, no. 5 (2017): 968-978. https://doi.org/10.14716/ijtech.v8i5.879

[7] Nassar, Mostafa, Nirmeen A. El-Bahnasawy, HossamEl-Din H. Ahmed, Adel A. Saleeb, and Fathi E. Abd El-Samie. "Network intrusion detection, literature review and some techniques comparision." In *2019 15th International Computer Engineering Conference (ICENCO)*, pp. 62-71. IEEE, 2019. https://doi.org/10.1109/ICENCO48310.2019.9027296

[8] Salih, Azar Abid, and Adnan Mohsin Abdulazeez. "Evaluation of classification algorithms for intrusion detection system: A review." *Journal of Soft Computing and Data Mining* 2, no. 1 (2021): 31-40. https://doi.org/10.30880/jscdm.2021.02.01.004

[9] Bhosale, Karuna S., Maria Nenova, and Georgi Iliev. "Data mining based advanced algorithm for intrusion detections in communication networks." In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 297-300. IEEE, 2018. https://doi.org/10.1109/CTEMS.2018.8769173

[10] Gulla, Kishor Kumar, P. Viswanath, Suresh Babu Veluru, and R. Raja Kumar. "Machine learning based intrusion detection techniques." *Handbook of Computer Networks and Cyber Security: Principles and Paradigms* (2020): 873-888. https://doi.org/10.1007/978-3-030-22277-2_35

[11] Pande, Sagar Dhanraj, Govinda Rajulu Lanke, Mukesh Soni, Mukund Anant Kulkarni, Renato R. Maaliw, and Pavitar Parkash Singh. "Deep Learning-Based Intrusion Detection Model for Network Security." In *International Conference on Intelligent Computing and Networking*, pp. 377-386. Singapore: Springer Nature Singapore, 2023. https://doi.org/10.1007/978-981-99-3177-4_27

[12] Gu, Jie, and Shan Lu. "An effective intrusion detection approach using SVM with naïve Bayes feature embedding." *Computers & Security* 103 (2021): 102158. https://doi.org/10.1016/j.cose.2020.102158

[13] Ramasubramanian, Gopalakrishnan, and Singaravelu Rajaprakash. "An Avant-Garde African Vulture Optimization (A^2VO) based Deep RNN-LSTM Model for 5G-IoT Security." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 32, no. 1 (2023): 1-17. https://doi.org/10.37934/araset.32.1.117

[14] Taher, Kazi Abu, Billal Mohammed Yasin Jisan, and Md Mahbubur Rahman. "Network intrusion detection using supervised machine learning technique with feature selection." In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pp. 643-646. IEEE, 2019. https://doi.org/10.1109/ICREST.2019.8644161

[15] Chkirbene, Zina, Sohaila Eltanbouly, May Bashendy, Noora AlNaimi, and Aiman Erbad. "Hybrid machine learning for network anomaly intrusion detection." In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pp. 163-170. IEEE, 2020. https://doi.org/10.1109/ICIoT48696.2020.9089575

[16] Wang, Hong, Qingsong Xu, and Lifeng Zhou. "Seminal quality prediction using clustering-based decision forests." *Algorithms* 7, no. 3 (2014): 405-417. https://doi.org/10.3390/a7030405

[17] Mohamed, Raihani, Abdul Rafiez Abdul Raziff, and Sabri Mohd. Nasir. "A resample-smote balance with random forest for improving seminal quality prediction in healthcare informatics." *ARPN Journal of Engineering and Applied Sciences* 16, no. 21 (2021): 2264- 2274.

[18] Ingle, Darshan, and Divyanka Ingle. "An Enhanced Blockchain Based Security and Attack Detection Using Transformer In IOT-Cloud Network." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 31, no. 2 (2023): 142-156. https://doi.org/10.37934/araset.31.2.142156

[19] Swana, Elsie Fezeka, Wesley Doorsamy, and Pitshou Bokoro. "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset." *Sensors* 22, no. 9 (2022): 3246. https://doi.org/10.3390/s22093246

[20] Halbouni, Asmaa, Teddy Surya Gunawan, Mohamed Hadi Habaebi, Murad Halbouni, Mira Kartiwi, and Robiah Ahmad. "CNN-LSTM: hybrid deep neural network for network intrusion detection system." *IEEE Access* 10 (2022): 99837-99849. https://doi.org/10.1109/ACCESS.2022.3206425