iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** II **Month of publication:** February 2022

DOI: https://doi.org/10.22214/ijraset.2022.40414

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue II Feb 2022- Available at www.ijraset.com

An Efficient ML technique to Descry the DDoS Attacks in Real-time Network Traffic and its Feature Analysis

Dr. S. V. Vasantha
Department of Information Technology, Maturi Venkata Subba Rao (MVSR) Engineering College, Hyderabad

Abstract: ***In this modern Internet era with the advent of sophisticated technology, criminals can more often easily launch various kinds of cyber-attacks. Distributed Denial-of-service (DDoS) is one such attack that can easily bypass firewall and block the service provided by network resources and server machine. It creates a huge traffic from multiple systems to attack a particular server. This paper presents an efficient Machine Learning (ML) model construction process and its deployment to identify the DDoS attack related network traffic in real time. Fourteen different supervised ML algorithms are applied for training the model over the DDoS attack traffic data. The Light Gradient Boosting Machine (LGBM) Classifier has shown outstanding performance i.e., high accuracy of 98.4% in a very less time of 0.4 sec for training among others and expressed equally better performance when analyzed for other metrics such as AUC, Recall, Precision, F1-score, Kappa and MCC. This efficient classifier is further improvised by tuning its hyper parameters resulted in 98.7% accuracy and tested on 30% of unseen data resulted with 98.3% accuracy. When it is tested with real-time network traffic, exhibited 97.6% accuracy. The results show that LGBM Classifier achieves the highest accuracy. In this paper, it is also analyzed that source port is the important feature contributing mainly to the enhanced LGBM classifier accuracy.***

Keywords: Cyber Attacks, DDoS, Attack Detection, Machine Learning, Classification

I. INTRODUCTION

The online services or applications provided by the Internet have become the primary target of several types of attacks. Among these, Denial-of-Service(DoS) and Distributed Denial-of-Service(DDoS) are a special kind of attacks. Both may completely stop or worsen the services provided by the servers to their legitimate customers and even may lead to reputation and financial damage to several online business organizations. They flood too many packets generating massive traffic in the network. [1–7]. DoS attacks, basically aim for consumption of system or network resources by generating large traffic from only one single computer system towards the selected victim[8-10]. Coming to DDoS attacks, they transformed the old style of one-to-one attack scenario into many-to-one attack scenario creating a botnet so, as to flood with huge traffic from many systems towards the victim's machine. Some examples of them are, SYN, Smurf, UDP, and DNS flood attacks [11].

II. RELATED WORK

To descry the DDoS attacks, different Machine Learning(ML) and Deep Learning(DL) solutions are proposed by various researchers. This section presents some of the recent work carried out to provide solution for DDoS attacks.

The author in [12] employed Decision-Tree(DT), Artificial Neural Network(ANN) and Naive Bayes(NB) algorithms for classification of DDoS attack and obtained accuracy of 0.839, 0.843 and 0.765 respectively. [13] presented a hybrid solution, which is a combination of Neural Network(NN) and Support Vector Machine(SVM) in telecommunication networks with accuracy around 0.90 and observed improvement over individual methods (NN and SVM). [14] proposed a novel attack mitigation in Software Defined Networking(SDN) based Internet for ICMP and TCP-SYN flood attacks. They applied KNN and XGBoost techniques and observed mitigation of attack over 98%. [15] designed a semi-supervised ML mechanism for the dataset, which is partially labeled. They used agglomerative along with K-means as clustering approaches and proposed a voting scheme for labeling the data with attack or normal class. Next Random Forest(RF), K-Nearest Neighbors(KNN), and Support Vector Machine(SVM) are applied for supervised learning and obtained 0.96, 0.95 and 0.92 accuracy values respectively.

The authors in [16] used RF for detecting UDP flood, HTTP flood and SSH brute-force attack types. It resulted in accuracy score of 0.893. [17] considered two datasets, one is without any feature selection process and the other is created based on the feature selection process.

These two datasets were applied with SVM, NB, ANN, and KNN methods. They noticed highest score of accuracy i.e., 0.98 for KNN with feature selection. [18] proposed a Deep Learning based DDoSNet model for SDN environment, which is a combination of RNN and auto-encoder, that resulted in 0.99 accuracy. Authors in [19] built a cloud-based student portal to create a new dataset for their study. They employed supervised techniques such as SVM, DT, Logistic Regression (LR) and KNN, which resulted in Jaccard scores of 94.4 %, 94.3 %, 94.2 % and 94.1 % respectively.

Application of ML and DL based approaches have shown its importance in accurate identification of different kinds of DDoS attacks in varied environments but an efficient solution which gives better performance with less processing time is expected when it is deployed for the analysis of real-time network traffic. Hence, this paper presents an efficient ML solution to address these issues.

III.PROPOSED SYSTEM

The proposed system is designed for creating a best model for the considered DDoS attack data, which is depicted in the fig 1. It takes voluminous dataset generated from huge number of DDoS attack related requests as input and applies an unsupervised ML technique to eliminate similar records. Thereby dataset size is reduced, which is then pre-processed to create the required dataset for the model training. Next various popular ML approaches are applied to train the model. The ML method exhibiting top performance along with reduced training time is picked up as an efficient algorithm.

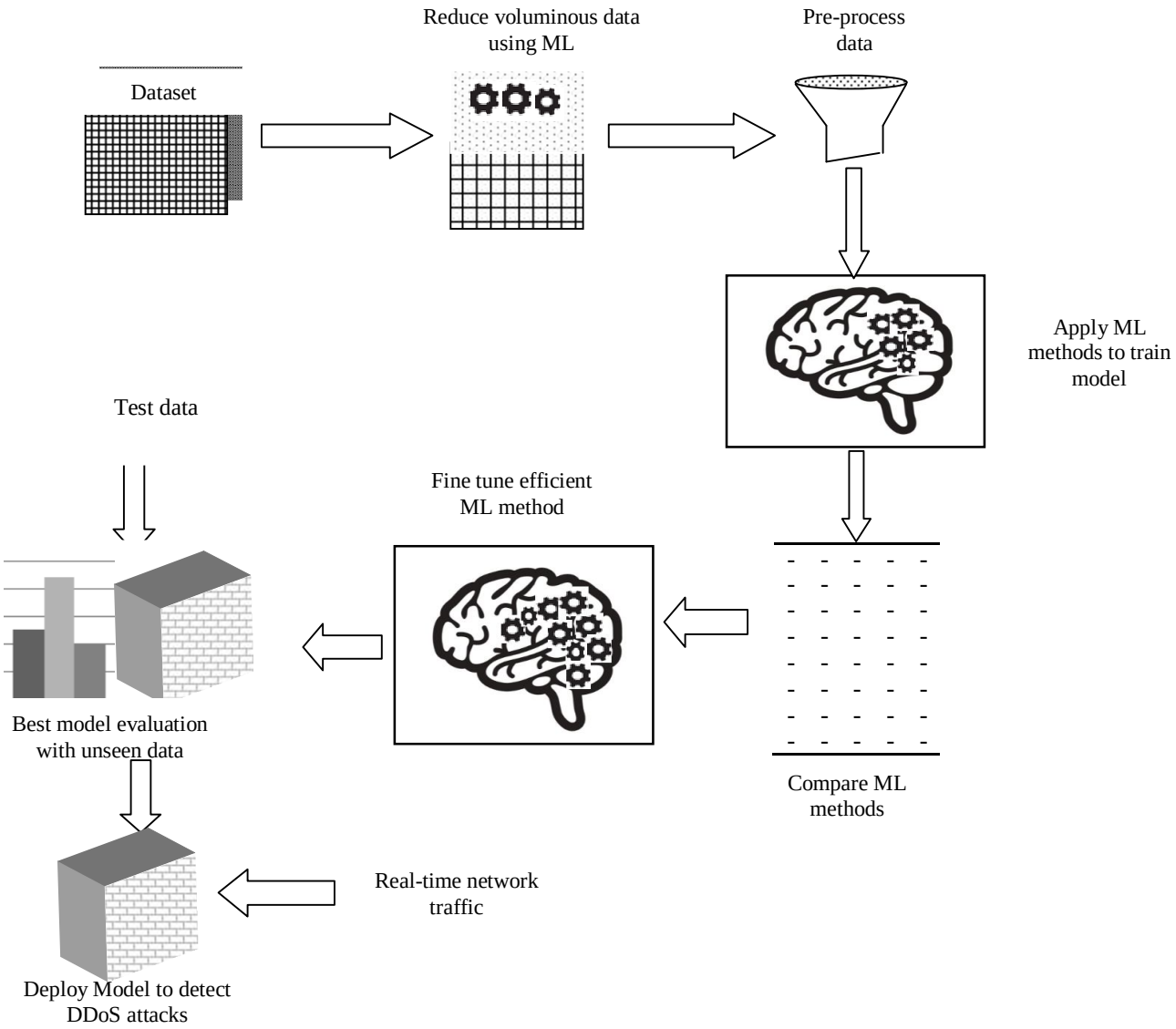


Fig. 1: Proposed System for constructing best deployable model for DDoS attack detection

This efficient algorithm is picked up and then fine tuned its hyper-parameters to produce the best model. It is evaluated with unseen test data and if the performance found satisfied then it is deployed to distinguish the attack traffic from normal traffic in real-time environment.

A. Dataset Preparation

Initially labeled dataset is taken from [1] and k-means algorithm is applied for each class to group similar records under it so, as to drop similar records thereby reducing the use of computational resources needed and time required to process the voluminous dataset. The seven DDoS attack types considered in this paper are LPAD, NetBIOS, MSSQL, UDP, Portmap, Syn, and UDPLag. The seven attack classes and a benign class forms eight attack classes of the dataset. Sample count under each attack class in original version and reduced version of the dataset are expressed in table i.
The new dataset's size is reduced to manageable size i.e., (995, 85). The newly created dataset's characteristics are listed in table ii.

Table I: Sample count under each class of the original dataset

DDoS Attack Class	Sample Count (original dataset)	Sample Count (reduced dataset)
LPAD (Class-0)	200000	189
NetBIOS (Class-1)	200000	320
MSSQL (Class-2)	200000	108
UDP (Class-3)	200000	56
Portmap (Class-4)	186960	76
Benign (Class-5)	56965	87
Syn (Class-6)	2777	52
UDPLag (Class-7)	1873	107
Total	1048575	995

Table II: Dataset Characteristics

Parameter	Value
Target Type:	Multiclass
Original Data:	(995, 85)
Numeric Features:	54
Categorical Features:	30
Train-Test split:	70% - 30%

B. Data Pre-processing

Dataset is applied with normalization method as Zscore for rescaling the numerical feature values. The essential input variables are selected using Classic method as feature selection with 0.6 as its threshold. The StratifiedKFold is used as cross validation approach with number of folds specified as ten and shuffle is set to true value as a part of fair evaluation of the model. Pre-processing parameters and its corresponding values are shown in table iii.

Table III: Pre-processing parameters and its values

Pre-processing parameter	Value
Transformed Train Set:	(696, 249)
Fold Generator:	StratifiedKFold
Fold Number:	10
Shuffle Train-Test:	TRUE
Transformed Test Set:	(299, 249)
Normalize:	TRUE
Normalize Method:	Zscore
Features Selection Threshold:	0.6
Fix Imbalance Method:	SMOTE
Feature Selection:	TRUE
Feature Selection Method:	Classic

C. *ML approaches for Model Training*

The model is trained by applying 14 different popular ML methods over the created and pre-processed dataset. Their performance is analyzed with respect to different metrics along with training time in seconds as shown in the fig 2. It is observed that Gradient-Boosting Classifier gbc obtained the highest accuracy of 0.99 but training time TT is 2.697 sec so, the next highest accuracy obtained method is Light Gradient Boosting Machine LightGBM i.e., 0.984 in 0.420 sec. It is also second top performer in other metric in short TT. Hence LightGBM is marked as the efficient ML method.

D. Fine-Tuning Efficient ML Method

The proposed model is trained by applying LightGBM method. The algorithm is executed several times for different values of its hyper parameters to fine tune the model. At number of iterations = 50 and parameters with values mentioned in fig 3, the model has shown the best performance of 0.987, which is higher than the basic LightGBM model.

When the performance analysis of this model is performed it is noticed that standard deviation (SD) is minimum or almost negligible across all the seven metrics considered, which is expressed in the fig. 4. So, this model is marked as the best fine-tuned model.

The best fine-tuned LightGBM model is then deployed in real-time environment, where it is fed with real-time network packet flow to descry the seven attack classes and normal class. This crucial information is used to filter out unwanted attack traffic and forward only the normal traffic to maintain overall throughput without any major degradation in performance.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.990	1.000	0.992	0.991	0.990	0.988	0.988	2.697
lightgbm	Light Gradient Boosting Machine	0.984	1.000	0.984	0.986	0.984	0.981	0.981	0.420
et	Extra Trees Classifier	0.983	0.998	0.985	0.985	0.983	0.979	0.979	0.471
ridge	Ridge Classifier	0.981	0.000	0.981	0.983	0.981	0.977	0.978	0.022
svm	SVM - Linear Kernel	0.977	0.000	0.974	0.980	0.977	0.972	0.973	0.071
rf	Random Forest Classifier	0.976	0.999	0.975	0.978	0.976	0.970	0.971	0.524
dt	Decision Tree Classifier	0.971	0.983	0.966	0.974	0.971	0.965	0.965	0.030
lr	Logistic Regression	0.968	0.997	0.961	0.972	0.968	0.962	0.962	0.228
nb	Naive Bayes	0.956	0.990	0.950	0.961	0.951	0.946	0.948	0.028
knn	K Neighbors Classifier	0.934	0.989	0.906	0.941	0.933	0.919	0.921	0.133
lda	Linear Discriminant Analysis	0.901	0.943	0.862	0.912	0.892	0.879	0.883	0.061
ada	Ada Boost Classifier	0.648	0.914	0.485	0.522	0.561	0.564	0.601	0.194
dummy	Dummy Classifier	0.300	0.500	0.125	0.090	0.139	0.000	0.000	0.019
qda	Quadratic Discriminant Analysis	0.108	0.522	0.194	0.017	0.029	0.043	0.070	0.046

Fig. 2: Performance Analysis of ML methods

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.986	1.000	0.979	0.987	0.985	0.983	0.983
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	0.986	1.000	0.975	0.988	0.985	0.983	0.983
3	0.971	0.999	0.975	0.971	0.971	0.965	0.965
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5	0.986	0.999	0.991	0.987	0.986	0.983	0.983
6	0.986	1.000	0.984	0.986	0.985	0.982	0.983
7	0.986	1.000	0.969	0.987	0.985	0.982	0.983
8	0.986	1.000	0.992	0.987	0.986	0.982	0.983
9	0.986	0.999	0.982	0.986	0.985	0.982	0.983
Mean	0.987	1.000	0.985	0.988	0.987	0.984	0.984
SD	0.008	0.000	0.010	0.008	0.008	0.009	0.009

Fig. 4: Performance of Fine-tuned Model

```
LGBMClassifier(bagging_fraction=1.0, bagging_freq=1, boosting_type='gbdt',
               class_weight=None, colsample_bytree=1.0, feature_fraction=0.9,
               importance_type='split', learning_rate=0.05, max_depth=-1,
               min_child_samples=16, min_child_weight=0.001, min_split_gain=0.1,
               n_estimators=290, n_jobs=-1, num_leaves=100, objective=None,
               random_state=3380, reg_alpha=1e-06, reg_lambda=0.1,
               silent='warn', subsample=1.0, subsample_for_bin=200000,
               subsample_freq=0)
```

Fig. 3: Fine-Tuned Hyper Parameters

IV. RESULT DISCUSSION AND COMPARATIVE ANALYSIS

This section discusses about best model evaluation over unseen test data, feature analysis and its performance analysis over real-time network traffic. It also includes comparative analysis of the proposed model with other recent solutions for descry the DDoS attacks.

A. Best Model Evaluation

Initially newly created dataset is divided into two parts i.e., 70% for training and 30% for testing. The best marked model's performance is evaluated by giving 30% of unseen test data as input. Performance of model in terms of various metrics is depicted in the table iv, which shows that proposed LightGBM obtained remarkable accuracy of 0.983 at the same time it has shown equally best performance in terms of other metrics.

Table IV: Performace analysis of proposed LightGBM model

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Light Gragient Boosting Machine(LightGBM)	0.983	0.999	0.983	0.984	0.983	0.979	0.979

The fig.5 depicts confusion matrix that clearly shows that the model predicted good number of true-positives with very few errors in the prediction. Thus, the proposed LightGBM is observed as the best classifier for the considered seven attack classes and benign class. Performance of the proposed classifier is measured using AUC-ROC curve to check how accurately the model is able distinguish between considered seven attack classes and a benign class. The AUC-ROC curves of the proposed model is depicted in the fig. 6, which shows that AUC value for all the eight classes is one (i.e. AUC=1). So, it is observed that proposed model is working as perfect classifier.

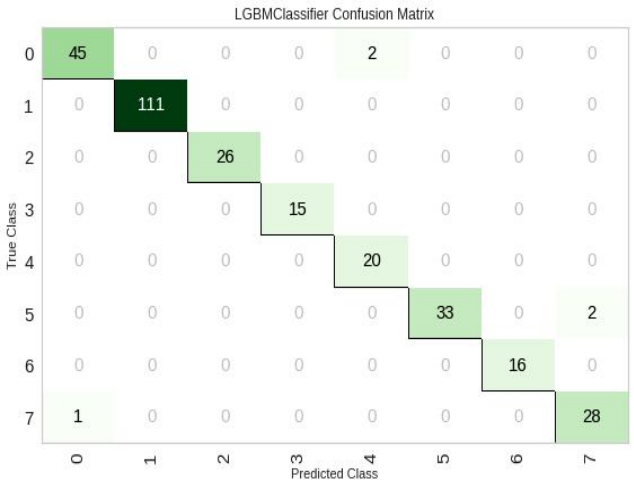


Fig. 5: Proposed LightGBM Classifier Confusion Matrix

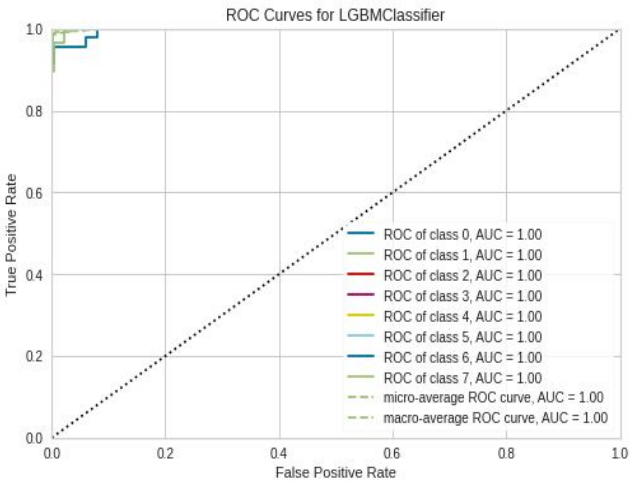


Fig. 6: ROC Curves of proposed LightGBM classifier

The classification report shown in the fig. 7 is used to evaluate the prediction accuracy in terms of classification metrics including

1) Precision, which is the ratio of true-positives to all positives. It is expressed in the equation 1.
Precision(P)=True-Positives/(True-Positives+False-Positives) -(1)

2) Recall, which is the ratio of true-positives to true-positives and false-negatives. It is expressed in the equation 2.
Recall(R)=True-Positives/(True-Positives + False-Negatives) -(2)

3) F1-Score, which is harmonic-mean of both P and R values. It is calculated as expressed in the equation 3.
*F1-Score (F1)= 2 * [(R * P) / (R + P)] -(3)*

It is observed that, the proposed classifier has attained highest precision and recall i.e., P=1 & R=1 for 5 classes and for remaining classes it is considerably high. F1 value is maximum for 4 classes and for other 4 classes it is above 0.94. In fig.8, learning curves of proposed LightGBM model is depicted that includes training and validation curves. It is noticed that training curve constantly and remarkably maintained high accuracy. Coming to validation curve, initially, it is started at around 0.96 and progressed over number of training instances to nearly about 0.99.

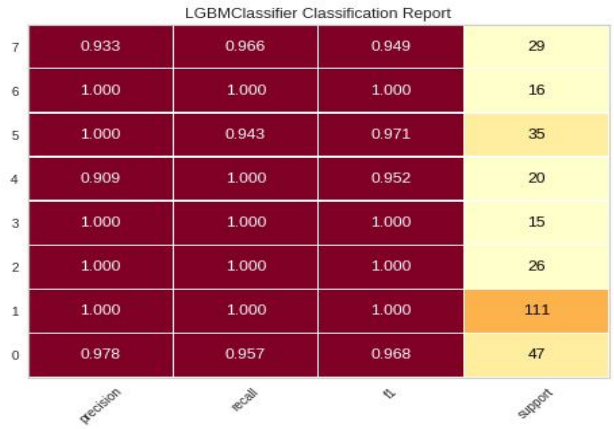


Fig. 7: Classification Report of proposed LightGBM Classifier

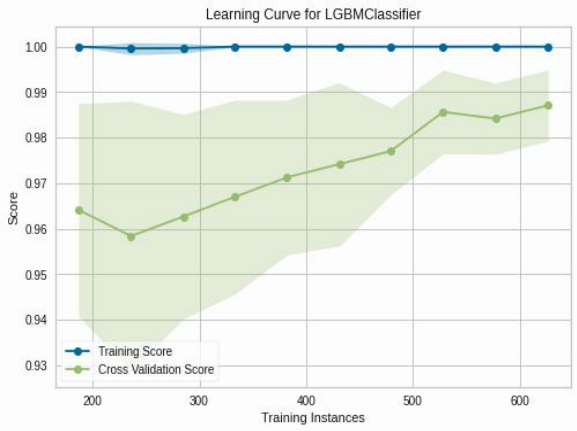


Fig. 8: Training and Validation Curves of proposed LightGBM Classifier

B. Feature Analysis of the proposed Model

The proposed LightGBM model is analyzed to check which features of the feature set are contributing more for the accurate prediction of the classes. It is observed that source_port is the top most contributing feature. Next features that significantly contributing are fwd_packet_length_max,flow_bytes, timestamp_0159.9, destination_port and subflow_fwd_bytes. The feature analysis of the model is shown in the fig. 9.

C. Comparative Analysis

The proposed LightGBM model is compared with the classifiers proposed by [1] as expressed in the table v, since the initial dataset that is considered as input to the proposed system is taken from [1]. It is observed that the proposed LightGBM classifier exhibited outstanding accuracy, recall and F1-Score of 0.983 when compared with XGBoost and CNN presented by the authors in [1]. There is a significant increase of 11.4 percent and 9.1 percent in accuracy measure when proposed LightGBM model is compared with CNN and XGBoost methods.

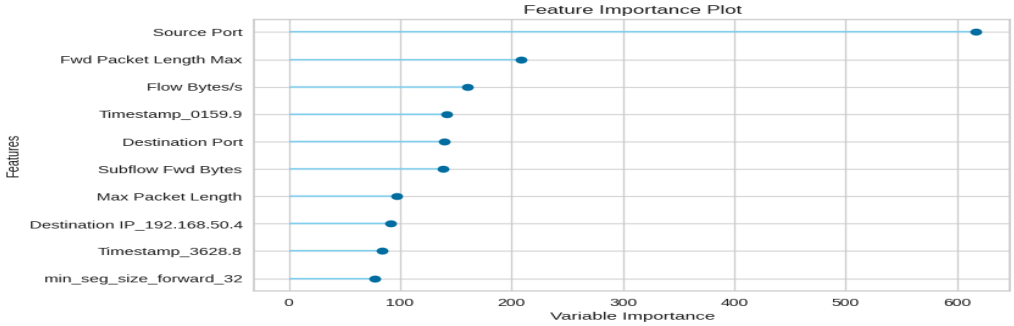


Fig. 9: Feature Importance plot of proposed LightGBM model

Table V: Comparison of proposed LightGBM with XGBoost and CNN classfiers

ML Method	Accuracy	F1-Score	Recall
XGBoost Classifier [1]	0.892	0.877	0.892
CNN [1]	0.839	0.713	0.839
Proposed LightGBM	0.983	0.983	0.983

D. Performance Analysis of the Proposed Model on real-time Network traffic

Network packet-sniffing tool called Wireshark is employed to capture the real-time network traffic under attacking scenario and normal scenario. Data generated using this tool is then pre-processed and fed as input to the best performing LightGBM model. It has attained accuracy of 0.976 and AUC of 0.988. The proposed model is evaluated using other five metrics such as Recall, Precision, F1, Kappa and MCC, which is shown in the table vi, it has exhibited outstanding performance even in case of real-time traffic.

Table VI: Performance analysis of proposed LightGBM in real-time network traffic

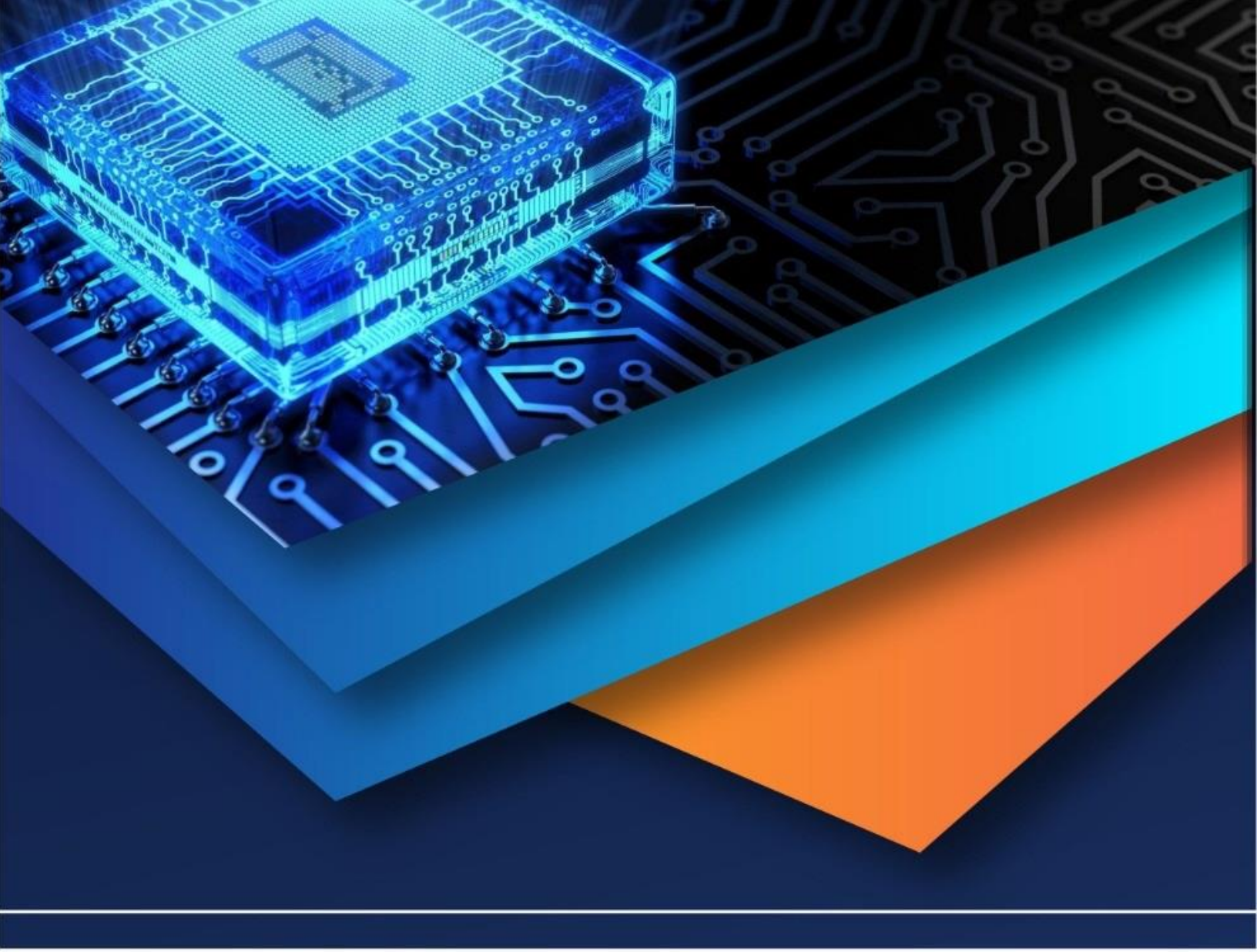
Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Light Gragient Boosting Machine(LightGBM)	0.976	0.988	0.981	0.979	0.98	0.975	0.974

V. CONCLUSION

The proposed LightGBM model is observed as the top-performing ML technique over the newly created condensed DDoS attack dataset in terms of accuracy and training time. This efficient ML Classifier is fine tuned and evaluated using various performance metrics on unseen and real-time network data, which resulted in outstanding accuracy values of 0.983 and 0.976 respectively when compared to the performance of original input dataset that is 0.892. Further, this proposed system can be employed in cloud, SDN and other environments.

REFERENCES

[1] Usha G., Narang M., Kumar A. Detection and Classification of Distributed DoS Attacks Using Machine Learning. In: Smys S., Palanisamy R., Rocha Á., Beligiannis G.N. (eds) Computer Networks and Inventive Communication Technologies. Lecture Notes on Data Engineering and Communications Technologies, vol 58. Springer, Singapore. (2021).
[2] C. Townsley, Are businesses getting complacent when it comes to ddos mitigation?, Network Security 2018 (2018) 6–9.
[3] A. Chadd, Ddos attacks: past, present and future, Network Security 2018 (2018) 13–15.
[4] S. Newman, Under the radar: the danger of stealthy ddos attacks, Network Security 2019 (2019) 18–19.
[5] Akamai, [state of the internet] / security A YEAR IN REVIEW, 2018. https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/2018-state-of-the-internet-security-a-year-in-review.pdf
[6] G. Somani, M. S. Gaur, D. Sanghi, M. Conti, M. Rajarajan, Scale inside-out: Rapid mitigation of cloud ddos attacks, IEEE Transactions on Dependable and Secure Computing 15 (2017) 959–973.
[7] A. Wang, W. Chang, S. Chen, A. Mohaisen, A data-driven study of ddos attacks and their dynamics, IEEE Transactions on Dependable and Secure Computing 14 (2015).
[8] T. Peng, C. Leckie, K. Ramamohanarao, Survey of network-based defense mechanisms countering the dos and ddos problems, ACM Computing Surveys 39 (2007) 3.
[9] A. Saied, R. E. Overill, T. Radzik, Detection of known and unknown ddos attacks using artificial neural networks, Neurocomputing 172 (2016) 385–393.
[10] C. Douligeris, A. Mitrokotsa, DDoS attacks and defense mechanisms: classification and state-of-the-art, Computer Networks 44 (2004) 643–666.
[11] Vinicius de Miranda Rios, Pedro R.M. Inácio, Damien Magoni, Mário M. Freire, Detection of reduction-of-quality DDoS attacks using Fuzzy Logic and machine learning algorithms, Computer Networks, Volume 186, 2021, 107792, ISSN 1389-1286
[12] Ahmad Sanmorino 2019 J. Phys.: Conf. Ser. 1175 012025
[13] Dr. A. Pasumpon pandian, Dr.S. Smys. DOS attack detection in telecommunication network using machine learning. J.Ubiquit. Comput. Commun. Technol. (ucct) 1(01), 33-44 (2019)
[14] N. N. Tuan, P. H. Hung, N. D. Nghia, N. V. Tho, T. V. Phan, and N. H. Thanh, ‘‘A DDoS attack mitigation scheme in ISP networks using machine learning based on SDN,’’ Electronics, vol. 9, no. 3, p. 413, Feb. 2020.
[15] Muhammad Aamir, Syed Mustafa Ali Zaidi, Clustering based semi-supervised machine learning for DDoS attack classification, Journal of King Saud University - Computer and Information Sciences, Volume 33, Issue 4, 2021, Pages 436-446, ISSN 1319-1578,
[16] Radivilova, Tamara & Kirichenko, Lyudmyla & Ageyev, Dmytro & Bulakh, Vitalii. (2019). Classification Methods of Machine Learning to Detect DDoS Attacks. 207-210. 10.1109/IDAACS.2019.8924406.
[17] H. Polat, O. Polat and A. Cetin, "Detecting DDoS attacks in software-defined networks through feature selection methods and machine learning models", Sustainability, vol. 12, no. 3, pp. 1035, Feb. 2020.
[18] M. Elsayed, N. Le-Khac, S. Dev and A. Jurcut, “DDoSNet: A Deep-Learning Model for Detecting Network Attacks.” 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM) (2020): 391-396.
[19] K. K. S. Arpitha, ‘‘DDoS attacks using machine learning,’’ J. Xi’an Univ. Archit. Technol., vol. 2, no. 4, pp. 3380–3384, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY