OPTIMISED DDOS DETECTION USING MACHINE LEARNING

[1] Abdur Rahman, [2] Imtiyaz Khan, [3] Mohammed Waheeduddin Hussain

[1] PG scholar, Shadan College of Engineering and Technology

[2] Professor, Shadan College of Engineering and Technology

[3] Professor, Shadan College of Engineering and Technology

Abstract: The project seeks to create a system for identifying distributed Denial of service (DDOS) the usage of sophisticated techniques, specifically machine learning algorithms including logistics regression, K-Nearest Neighbor and random forest. The proposed technique aims to benefit a high degree of accuracy in figuring out DDOS attacks. Precise detection is essential for proper alleviation of such threats. The aim of the research is to overcome current methodologies and significantly improve DDOS attacks. This means emphasis on innovations and improved solutions. The project selects the NSL KDD data file to evaluate the proposed models. This data file is selected because to reduce redundancy with respect to others, which brings better and more accurate results during testing and evaluation. The project performs a comparative analysis of many machine learning classifiers, including LR, RF, DT and KNN. This comparison emphasizes the advantages and disadvantages of several DDOS attack detection. The project approach includes a systematic procedure, including data collection, function extraction and classification. It uses the characteristics and network behavior as the basic elements for detection procedure. This shows the methodological approach for the improvement of the DDOS detection system. The DDOS detection undertaking includes a comprehensive record method with a voting classifier that integrates RF and Adaboost, in addition to a stacking classifier that combines RF, MLP and LightGBM. This goal is to increase the overall performance of the gadget the usage of the additional properties of several device getting to know techniques.

"Index terms - *DDoS; Deep Learning; Random Forest; Logistic Regression; KNN; NSL KDD Dataset".*

1. INTRODUCTION

The growing number of 5.03 billion Internet users worldwide increases the risks of cyber security [1]. The analysis found a 90% increase in DDOS attacks in the third quarter of 2022 compared to the previous year. [2]. According to [3], DDOS attacks cause significant infrastructure, industrial, government and economic losses. DDOS attacks are disrupted by web server services with a malicious intention.

DDOS attacks use botnet compromised devices to flood the target by malicious data. DOS attack, on the other hand, uses one device to flood the target with operation. DDOS attacks are volume -based, log -based and application layer [16]. quantity -based assaults which includes UDP and ICMP floods are seeking to crush the bandwidth width for bits consistent with second. A protocol -based totally assaults, together with the floods of the son and ping of demise, can exhaust servers, firewalls, and the weight offset by means of shipping of several packets according to second. at some stage in these attacks, communication gear ought to overload and unavailable. software layer attacks together with Get/put up and coffee and low and occasional and low attacks and attacks are predicted in the requirements per 2nd to knock down the internet server [4]. these assaults frequently appear as benign requests.

DDOS attacks disrupted DNS revenues and lost society. more than 65% of DDOS attacks are voluminous and provides a large quantity of unrelated records to overload the sufferer's processing abilties or adjacent network connections. those assaults are successful due to the fact internet routers use popular rules for queues, which includes the first within the first and drug decomposition to handle all operations equally, together with assault and legal visitors. thus, those assaults can prevent the sufferer's records from being manipulated. however, the DDOS Lowvolume attacks [2, 8, 10] use software layers protocols to exhaust target assets without overloading networks. those attacks inside the last mins or much less than an hour, so it is tough to discover the use of ordinary strategies.

Based on our results, we suggest a deep learning technique for DDOS detection [16], which combines data collection, extraction and categorization of functions and binary classification. The proposed technique uses network behavior, packet length, intervals between packets and protocol. We test classifiers Detection of LR, RF, DT and KNN attack. We are testing our access to the NSL KDD data file and the results are encouraging.

2. LITERATURE SURVEY

Precise agriculture or intelligent agriculture gains popularity to satisfy the growing global demand for food. technology and connected devices are utilized in shrewd farms to display vegetation and soil moisture and deployment of drones for spraying insecticides. however, various internet -related devices have created the dangers of sensible ecosystems. These vulnerabilities allow attackers to remotely influence and disturb the data from sensors on fields and autonomous vehicles such as intelligent tractors and drones. This may be disastrous, especially when harvesting when live monitoring is necessary. This research [3] shows a DOS attack that disrupts the sensors in the intelligent field [9, 12, 16]. The Wi-Fi Deauthentication attack is discussed using IEEE 802.11 shortcomings that do not discerning administrative packets. ESP8266 Maker focus Development Board Wifi deauther Monster disconnects Raspberry Pi from the net and prevents cloud storage of sensors. This attack was prolonged to the complete community, which averted all devices from connecting. We advocate professionals to be privy to the existing risks to put in smart agricultural ecosystems and the Cyber security community for domain -specific agriculture.

The Internet is now used virtually everywhere. As IoT technology becomes more popular, billions of devices [5, 7] are associated with the Internet. However, DOS/DDOS attacks are the greatest danger to this expanding technology. The new DDOS attacks are complex and difficult to detect or neutralize by means of detection of disturbances and older approaches. on this research, data mining and machine learning are used to perceive DDOS. This study uses the latest CICDDOS2019 data file to evaluate the most popular machine learning techniques and identify the most interconnected characteristics with projected classes. Adaboost and Xgboost expected network operation of 100% of the time. Increasing the model for multiclassification of DDOS attacks and hybrid algorithms and new data sets can expand future studies.

Vulnerable IoT gadgets [13] allow botnets that cost billions per year. This research examines BASHLITE BOTNETS AND MIRAI. We emphasize the development of malware and behavior of botnet operators [5]. Records from 47 HoneyPots for 11 months are used. Our findings illuminate botnets and show that malware, botnets operators and harmful activities are increasingly sophisticated. Mirai has a more durable hosting and control infrastructure and allows more efficient attacks than its predecessors.

The main problem is the defense of the Internet DDOS. Redirecting all destinations (eg via DNS or BGP) to a 3rd party, DDOS safety-as-a-provider company (eg cloudflare and akamai) with a well-furnished and proprietary filtering that they get before it is within the last business interview with more business interviews than it's far 100 business conversations. The technique itself isn't sufficient, especially for massive companies (eg web hosting organizations, authorities), which can't have enough money to permit the security carriers of the 1/3 authorization to quit their connect community. these groups have to clear out those corporations via their net offerings. In this text, we are dealing with internet services troubles while worrying the net security sector and our sensible contractual answers.

The largest cyber threat is DDOS attacks [2, 4, 6, 7, 8]. Inhibition of server's ability to provide resources to real consumers slows the bandwidth and the size of the buffer. This article shows a mathematical model for DDOS attacks [9]. LR and NB identify attacks and normal circumstances. The Caida 2007 data file is used for experiments. Machine learning algorithms are used, tested and verified using this data file. This research implements Weka data mining and compares findings. Other machine learning methods for attacks on rejection of services are compared.

DDOS is the most serious threat of network security [10, 11, 13]. DDOS attacks are disrupted by key Internet applications. DDOS attacks maintain busy system requirements instead of serving real users. These attacks are increasingly frequent and more sophisticated. Thanks to these attacks it is difficult to identify and protect Internet services. In this article, we found and classified network traffic flows using machine learning [4, 9, 10] [10]. The proposed technique is verified using a fresh data file with “HTTP floods, SID DOS and regular operation. Weka classifies attacks using machine learning”. The results showed that the J48 algorithm surpassed RF and NB algorithms.

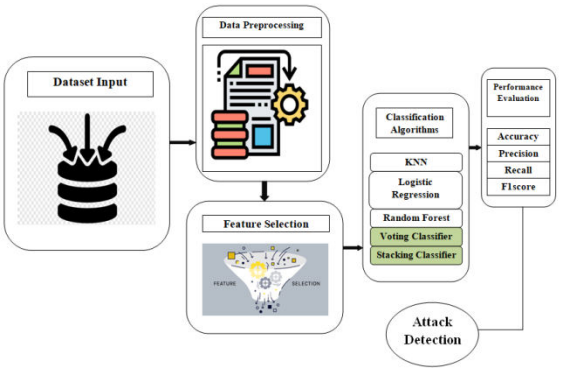
3. METHODOLOGY

i) Proposed Work:

The proposed device makes use of machine learning techniques including LR, KNN and RF, to enhance DDOS attack detection. It includes 3 basic steps: statistics collection, function extraction and classification. This method will increase cyber safety via detecting DDOS attacks in community visitors precisely. To growth the DDoS detection task, a complete file method has been set up using a voting classifier that integrates RF and Adaboost with a stacking classifier that combines RF, MLP and LightGBM. These models seek to increase the performance of the system by additional benefits of several machine learning methods [13]. A user - friendly SQLite flask has been created, including registration and sign features for efficient user testing, improving the availability and practicality of the system in the real world.

ii) System Architecture:

The process begins with aggregation and data file organization. The NSL KDD records file is frequently used, including network traffic facts with annotated examples of normal and offensive operation DDOS. data practise is a necessary phase wherein uncooked data is cleaned, processed and geared up for evaluation. this means solving absent values, removal of duplication and standardization or scaling of features to maintain uniformity. selecting functions approach identifying the maximum relevant functions or features from the information document. This phase strives to reduce dimension and increase the effectiveness of the detection procedure. Prevailing methodologies including correlation analysis, mutual information and evaluation of significance. This phase is the basis of the system and uses several machine learning classification techniques to identify DDOS attacks. In this context, referenced methods are used - KNN, LR, RF, Voting and extension and stacking classifier. Each algorithm analyzes preliminary data for categorizing network traffic as normal or indicative of DDOS attack. The overall performance standards are determined to evaluate the effectiveness of the DDOS detection device. The criteria consist of measurements including “accuracy, precision, recall and F1-score”. these measurements evaluate the efficiency of the device in difference between normal and offensive operation.



“Fig 1 Proposed architecture”

iii) Dataset collection:

NSL-KDD DATASET

This project used the NSL-KDD data file [8] to train machine learning models. It offers a variety of network site visitors facts, which include several offensive sorts and advanced labeling. This data report is an important supply for evaluating the detection of disruption and checking out fashions of machine learning in a balanced and controlled settings.

This is five front rows of NSL-KDD data file. It has 43 columns, of which we present a selection.

```
train_data.head()
```

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot
0	0	tcp	ftp_data	SF	491	0	0	0	0	0
1	0	udp	other	SF	146	0	0	0	0	0
2	0	tcp	private	S0	0	0	0	0	0	0
3	0	tcp	http	SF	232	8153	0	0	0	0
4	0	tcp	http	SF	199	420	0	0	0	0

5 rows × 43 columns

"Fig 2 NSL KDD dataset"

iv) Data Processing:

data processing converts unrefined records to usable data for establishments. facts scientists frequently address records processing, inclusive of series, organization, cleansing, validation, analyzes and facts transformation into interpretable representations along with graphs or articles. facts processing can be done by means of 3 strategies: guide, mechanical and digital. The goal is to growth the cost of statistics and make selections more efficient. This allows companies to strengthen their operations and carry out quick strategic choices. in this context, computerized facts processing technologies, including software program development, are crucial. it is able to transform large facts units, mainly huge facts, to good sized expertise of excellent and selection -making.

v) Feature selection:

the selection of features is the manner of figuring out the most convertible, non -relevant and relevant traits for the development of the version. The systematic minimalization of the dimensions of the records set is essential, at the same time as the quantity and diversity of records sets persist in growth. The number one goal of choosing factors is to increase the performance of the predictive model and on the equal time limit the computing fees of modeling. the selection of features, the basic aspect of useful engineering includes identification of the most vital traits for coming into the gadget mastering algorithms. the choice strategy is used to reduce the quantity of enter variables through the exclusion of redundant or needless features, and consequently improves the set to the ones which might be maximum suitable for the machine learning version. primary benefits of choosing features earlier before allowing the machine learning version to pick the most vital houses [16].

vi) Algorithms:

Logistic Regression- LR is a classification technique used to predict the likelihood of enter belonging to a certain category. The sigmoid feature is used to convert enter information to the probability score, which stages from 0 to 1. The input is divided into one of the many classifications depending on the applied opportunity threshold. at some point of education, the version modifies its coefficients to increase the accuracy of information category and extension of the category.

```
from sklearn.linear_model import LogisticRegression
# instantiate the model
clf = LogisticRegression()

# fit the model
clf.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_hat = clf.predict(X_test)

lr_acc = accuracy_score(y_hat, y_test)
lr_prec = precision_score(y_hat, y_test,average='weighted')
lr_rec = recall_score(y_hat, y_test,average='weighted')
lr_f1 = f1_score(y_hat, y_test,average='weighted')
```

"Fig 3 Logistic regression"

Random Forest: RF is a method of learning a file that makes use of a combined strength of numerous selection bushes to provide predictions. He does this via schooling the gathering of decision trees on

random information subgroups and later via consolidation in their predictions. This record technique will increase predictive accuracy, reduces excess and offers robust performance appropriate for type and regression application [7].

```
from sklearn.ensemble import RandomForestClassifier

# instantiate the model
rf = RandomForestClassifier(random_state=10)

# fit the model
rf.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_pred = rf.predict(X_test)

rf_acc = accuracy_score(y_pred, y_test)
rf_prec = precision_score(y_pred, y_test,average='weighted')
rf_rec = recall_score(y_pred, y_test,average='weighted')
rf_f1 = f1_score(y_pred, y_test,average='weighted')
```

“Fig 4 Random forest”

KNN: k-Nearest neighbors (KNN) is a multilateral approach of machine learning used for type and regression software. all through the training phase, KNN retains statistics points together with their related labels or values. when predicting a class or the cost of a brand new records point, KNN determines the nearest acquaintances from the schooling set the use of a selected distance measurement. For type, it refers back to the essential class label amongst pals, however for regression it calculates the average in their values. KNN is based totally on the concept that analogy records points regularly display shared houses, making it an obvious and direct set of rules for plenty applications [7].

```
from sklearn.neighbors import KNeighborsClassifier
# instantiate the model
clf = KNeighborsClassifier(n_neighbors=3)

# fit the model
clf.fit(X_train, y_train)

#predicting the target value from the model for the samples

y_hat = clf.predict(X_test)

knn_acc = accuracy_score(y_hat, y_test)
knn_prec = precision_score(y_hat, y_test,average='weighted')
knn_rec = recall_score(y_hat, y_test,average='weighted')
knn_f1 = f1_score(y_hat, y_test,average='weighted')
```

“Fig 5 KNN”

Voting Classifier: The voting classifier is a technique of machine learning that integrates predictions of numerous basic classifiers, inclusive of RF and adaboost to increase accuracy. The RF makes use of numerous DT to lessen excessive portions, at the same time as Adaboost gradually trains vulnerable students and emphasizes rectification of incorrectly categorized cases. The voting classifier improves the accuracy of individual predictions by way of most or weighted voting, using the diversity of the model to treatment the mistakes resulting from every classifier, for that reason bringing extra sturdy predictions [7].

```
rfc = RandomForestClassifier()
parameters = {
    "n_estimators":[250],
    "max_depth":[200]

}

from sklearn.model_selection import GridSearchCV
forest = GridSearchCV(rfc,parameters,cv=10)

clf2 = DecisionTreeClassifier(random_state=1000)

eclf1 = VotingClassifier(estimators=[('rf-parameter', forest), ('dt', clf2)], voting='soft')
eclf1.fit(X_train, y_train)
y_pred = eclf1.predict(X_test)
```

“Fig 6 Voting classifier”

Stacking Classifier: The stacking classifier is a report approach that integrates three primary models: RF, MLP and LightGBM. The RF is known for its

set of DT, the MLP is a neural community adept in distinguishing complex samples and LightGBM is a frame for gradient boosting. Stacking uses predictions from those simple fashions as an access to training meta-modern. The intention is to increase the accuracy of prediction by means of the usage of the obvious benefits of every basic model and clarifying complicated facts interconnection and accordingly provide more accurate forecasts [7].

```
from sklearn.neural_network import MLPClassifier
from lightgbm import LGBMClassifier
from sklearn.ensemble import StackingClassifier

estimators = [('rf', forest),('mlp', MLPClassifier(random_state=1, max_iter=3000))]

clf = StackingClassifier(estimators=estimators, final_estimator=LGBMClassifier(n_estimators=1000))


clf.fit(X_train,y_train)

y_pred = clf.predict(X_train)
```

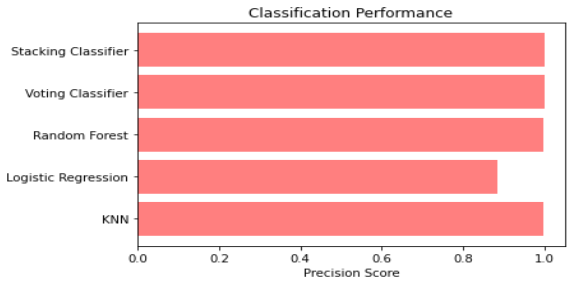
"Fig 7 Stacking classifier"

4. EXPERIMENTAL RESULTS

Precision: Precision quantifies the percentage of efficiently identified positive cases or samples. Precision is decided by using the components:

"Precision = True positives/ (True positives + False positives) = TP/(TP + FP)"

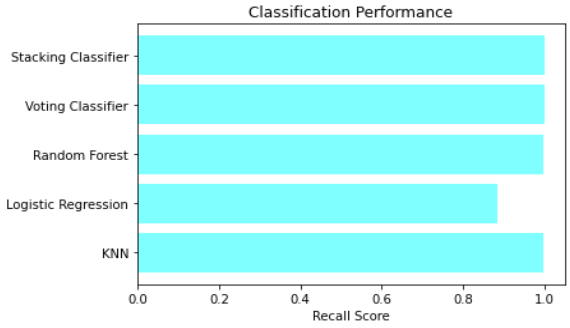
$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$



"Fig 8 Precision comparison graph"

Recall: ML recall assesses a model's potential to choose out all relevant times of a class. It demonstrates a version's efficacy in encapsulating times of a class by using comparing nicely anticipated high satisfactory observations to the general variety of positives.

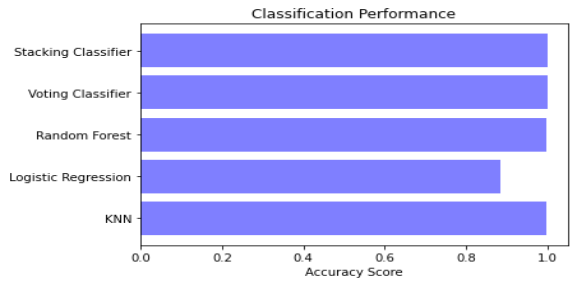
$$Recall = \frac{TP}{TP + FN}$$



"Fig 9 Recall comparison graph"

Accuracy: A test capacity towards create a proper difference between healthy & sick cases is a measure of accuracy. We can determine accuracy of a test through calculating proportion of cases undergoing proper positivity & genuine negative. It is possible towards express this mathematically:

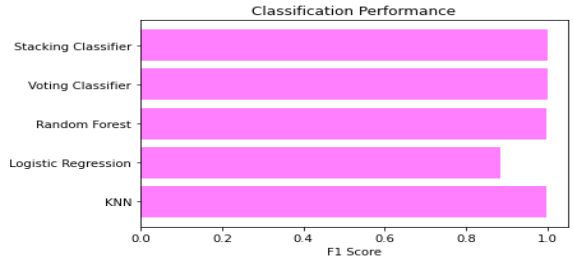
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$



"Fig 10 Accuracy graph"

F1 Score: The accuracy of a system ML of model is classed the usage of the F1 score. Integrating the precision and do not forget metrics of the model. The accuracy metric quantifies the frequency of proper predictions made through a model at some level inside the dataset.

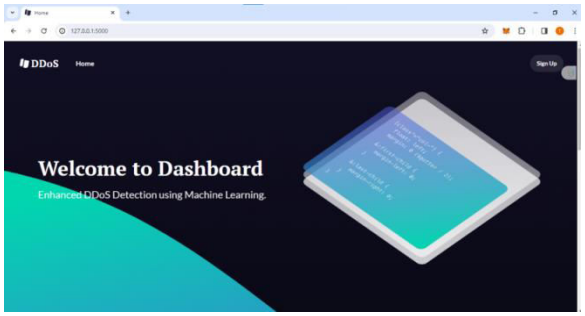
$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$



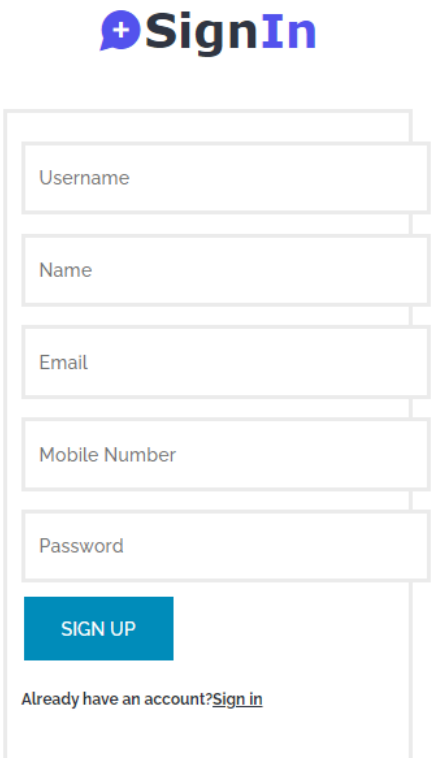
"Fig 11 F1Score"

	ML Model	Accuracy	Precision	Recall	F1-Score
0	KNN	0.997	0.997	0.997	0.997
1	Logistic Regression	0.883	0.885	0.883	0.884
2	Random Forest	0.998	0.998	0.998	0.998
3	Voting Classifier	1.000	1.000	1.000	1.000
4	Stacking Classifier	1.000	1.000	1.000	1.000

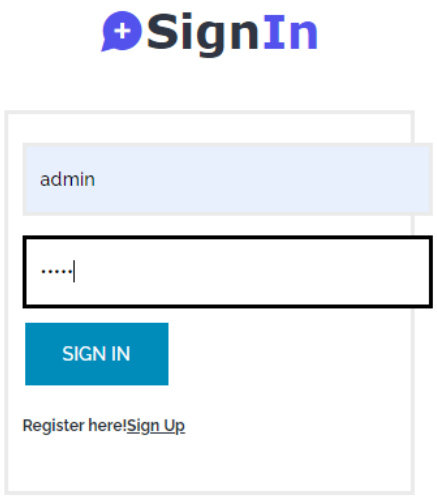
"Fig 12 Performance Evaluation"



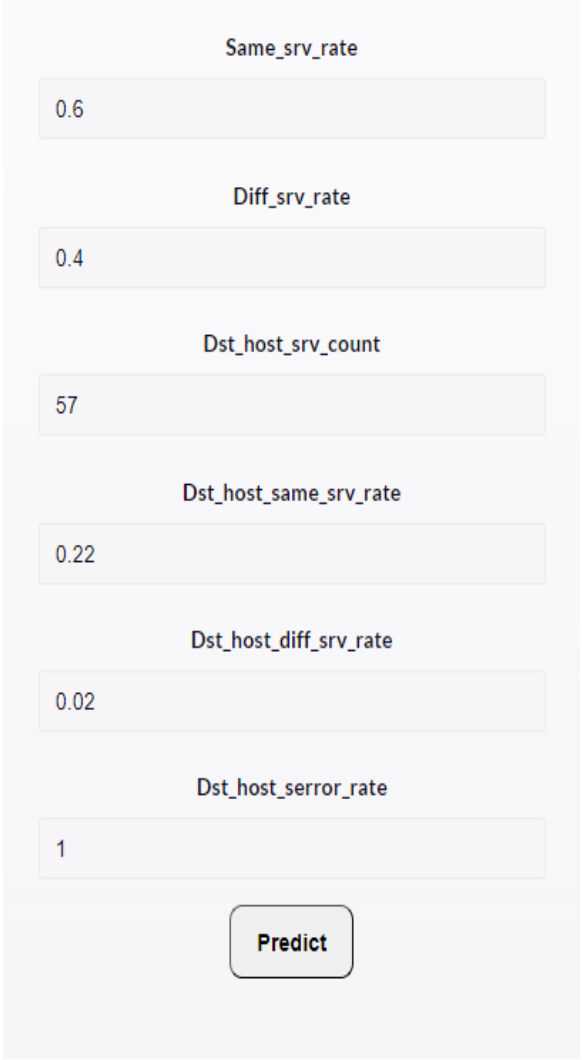
"Fig 13 Home page"



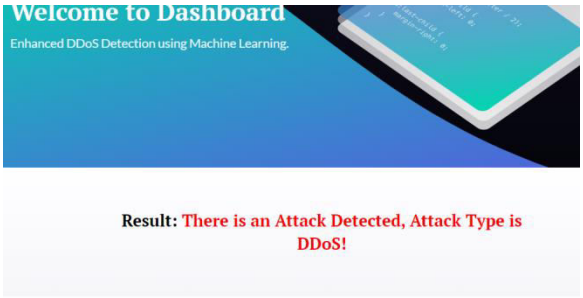
"Fig 14 Signin page"



"Fig 15 Login page"



"Fig 16 User input"



"Fig 17 Predict result for given input"

5. CONCLUSION

The project increases the security of digital systems by improving DDOS attack detection, thereby ensuring the continuity of online services and the protection of digital assets. We use several machine learning models to detect and alleviate DDOS attacks, including KNN, LR and RF [7]. We significantly increase accuracy using specialized approaches, such as the voting classifier and stacking classifier. They integrate the forecasts of many models to increase the robustness and reliability of our system. The project exceeds the development of algorithms by including a user -friendly front end using a flask frame. This practical solution allows users to test, allowing stakeholders to easily participate in the model. Integration of user verification using SQLite guarantees secure registration and logged in procedures, which increases the overall user experience. This experiment shows how machine learning can improve safety by effective DDOS attack detection [2, 3]. File models have shown exceptional accuracy and reliability. This initiative benefits several participating parties, including organizations, network administrators, cyber security and end -user experts. Increases the reliability and security of Internet services for all users.

6. FUTURE SCOPE

In the future, we can improve the system by including DDOS detection techniques in real time, increasing its speed and efficiency [15]. We can explore many methods to identify the most important criteria for a better effective prediction of DDOS attacks. By examining other machine learning techniques, we can increase the capabilities of the system beyond the possibilities provided by existing approaches such as LR, RF and KNN [13]. We need to increase the ability of the system to detect minor DDOS attacks that use sophisticated techniques that are now difficult to identify. Finally, we can increase the accuracy and functionality of the system by means of deep learning and the use of excellent methods to analyze network information.

REFERENCES

[1] Statista Research Department, “Worldwide digital population July 2022”, Available: https://www.statista.com/statistics/617136/digitalpopulation-worldwide/ (Last Accessed on: December 31, 2022)

[2] Ramil Khantimirov, “DDoS Attacks in 2022: Trends and Obstacles Amid Worldwide Political Crisis”, Available: https://www.infosecurity-magazine.com/blogs/ddos-attacks-in-2022- trends/ (Last Accessed on: December 31, 2022)

[3] S. Sontowski et al., "Cyber Attacks on Smart Farming Infrastructure," 2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC), 2020, pp. 135-143, doi: 10.1109/CIC50333.2020.00025.

[4] Seifousadati, Alireza and Ghasemshirazi, Saeid and Fathian, Mohammad, “A Machine Learning Approach for DDoS Detection on IoT Devices”, arXiv, 2021. Doi: 10.48550/ARXIV.2110.14911

[5] A. Marzano, D. Alexander, O. Fonseca et al., “The evolution of bashlite and mirai IoT botnets,” in Proceedings of the 2018 IEEE Symposium on Computers and Communications (ISCC), 2018.

[6] S. Kottler, “February 28th DDoS incident report,” 2018, https://github.blog/2018-03-01-ddos-incident-report/.

[7] Y. Cao, Y. Gao, R. Tan, Q. Han, and Z. Liu, “Understanding internet DDoS mitigation from academic and industrial perspectives,” IEEE Access, vol. 6, pp. 66641–66648, 2018.

[8] S. Newman, “Under the radar: the danger of stealthy DDoS attacks,” Network Security, vol. 2019, no. 2, pp. 18-19, 2019.

[9] Kumari, K., Mrunalini, M., “Detecting Denial of Service attacks using machine learning algorithms”, . J Big Data 9, 56 (2022).

[10] P. S. Saini, S. Behal and S. Bhatia, "Detection of DDoS Attacks using Machine Learning Algorithms," 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), 2020, pp. 16-21, doi: 10.23919/INDIACom49435.2020.9083716.

[11] Jiangtao Pei et al “ A DDoS Detection Method based on Machine Learning“, J. Phys.: Conf. Ser. 1237 032040, 2019.

[12] Abdullah Soliman Alshra’a, Ahmad Farhat, Jochen Seitz, “Deep Learning Algorithms for Detecting Denial of Service Attacks in Software-Defined Networks”, Procedia Computer Science, Volume 191, 2021, Pages 254-263, ISSN 1877-0509.

[13] Seifousadati, Alireza, Saeid Ghasemshirazi, and Mohammad Fathian. "A Machine Learning Approach for DDoS Detection on IoT Devices." arXiv preprintr Xiv:2110.14911 (2021).

[14] Francisco Sales de Lima Filho, Frederico A. F. Silveira, Agostinho de Medeiros Brito Junior, Genoveva Vargas-Solar, Luiz F. Silveira, "Smart Detection: An Online Approach for DoS/DDoS Attack Detection Using Machine Learning", Security and Communication Networks, vol. 2019, Article ID 1574749, 15 pages, 2019.

[15] R. Doshi, N. Apthorpe and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," 2018 IEEE Security and Privacy Workshops (SPW), 2018, pp. 29-35, doi: 10.1109/SPW.2018.00013.

[16] Ebtihal Sameer Alghoson, Onytra Abbass, “Detecting Distributed Denial of Service Attacks using Machine Learning Models”, International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, 2021.