Examining the emerging threat of Phishing and DDoS attacks using Machine Learning models.

Mohammed Zaid M S, Namratha N, Yashaswini B V,

Under Graduate Student, Dept. of Information Science and Engineering, BNMIT, Bengaluru, Karnataka, India
Under Graduate Student, Dept. of Information Science and Engineering, BNMIT, Bengaluru Karnataka, India
Assistant Professor, Dept. of Information Science and Engineering, BNMIT, Bengaluru, Karnataka, India

---***---

Abstract - *The usage of mobile devices in recent years has resulted in a considerable shift towards executing real-world activities in the digital arena. Although this has rendered our lives easier, it has additionally culminated in countless security breaches owing to the internet's anonymity. Although antivirus software and systems for firewalls can prevent most attacks, experienced attackers frequently take advantage of computer users' vulnerabilities by impersonating popular banking, networking, e-commerce, and other websites in order to steal private data such as user IDs, passwords, account numbers, credit card numbers, along with more. This emphasizes the importance of increasing knowledge and precaution when using the internet in order to protect oneself from cyber-attacks.*

Phishing is a social engineering technique used to deceive users into disclosing sensitive information, such as login credentials, credit card details, and personal identification. Distributed Denial of Service (DDoS) attacks are a common type of cyber-attack that aims to disrupt the availability of online services by overwhelming the targeted systems with a high volume of traffic. Phishing and DDoS attacks are two common cyber-attack types that aim to deceive users and disrupt online services. Phishing involves tricking individuals into revealing sensitive information, while DDoS involves overwhelming a website or network with traffic. Detecting these attacks is a complex task, and various methods have been proposed, including rule-based detection, blacklists, and anomaly-based detection. Machine learning-based anomaly detection has gained popularity due to its dynamic nature in catching "zero-day" attacks. To address the problem of phishing, which costs internet users significant amounts of money annually, a system that employs machine learning techniques such as logistic regression, decision tree, k-nearest neighbors, naive Bayes, random forest, and support vector classification is proposed. These algorithms predict outcomes based on user input parameters extracted from the front end.

***Key Words*: Logistic Regression, Cyber-security, Phishing, Machine Learning, DDoS, Random Forest, Support Vector Machine, Decision Tree, K Nearest Neighbor, XGBoost**

1. INTRODUCTION

Cyber attackers commonly use phishing and DDoS attacks to gain unauthorized access to sensitive information or disrupt online services. Phishing attacks are a tactic that tricks individuals into revealing confidential information such as login credentials, credit card details, or personal information. Attackers accomplish this by sending fraudulent messages or emails that appear to originate from legitimate sources but are in fact fraudulent. The attacker's objective is to lure the victim into clicking a link or opening an attachment that installs malware or directs them to a fake website where they will unwittingly divulge sensitive information.

Phishing attacks are designed to trick individuals into providing sensitive information, such as usernames and passwords, credit card numbers, or other personal information. This is typically done by sending emails or messages that appear to be from a legitimate source, but are actually from a fake or spoofed source. The goal of the attacker is to trick the victim into clicking on a link or opening an attachment that will then install malware on their device or direct them to a fake website where they will enter their sensitive information.

DDoS attacks, on the other hand, are designed to overwhelm a website or network with traffic, making it difficult or impossible for legitimate users to access the services provided by the targeted organization. This is typically done by using a network of compromised devices, such as computers or Internet of Things (IoT) devices, to flood the target with requests or data.

Individuals and organisations can suffer catastrophic effects resulting from phishing and DDoS assaults. These assaults can harm a company's brand and result in judicial or economic losses, in addition to the possible theft of confidential data or the interruption of services. As a result, it is critical to be aware of the dangers presented by these sorts of assaults and to take precautions to protect yourself and your organisation from them.

1.1 Background

Individuals and companies have benefited greatly from the growing usage of internet-based tools in all sectors of life. However, this has resulted in significant gaps in data safety, with cyber assaults emerging as a new hazard to persons and organizations. As a result, effective counter-measures to such assaults have become critical. Phishing is a sort of malicious assault that uses internet theft to obtain a user's sensitive information. It is an illegal act in which an unauthorized user seeks to gain the user's sensitive information by trapping the user. A distributed denial of service (DDoS) of Service) assault is a sort of cyber-attack in which several hacked computers flood a specific server or network with traffic, overloading it and causing it to crash.

1.2 Problem Statement

Phishing and DDoS assaults are two of the most common and crucial forms of cyber-attacks, costing organizations considerable financial and intangible damages [18][19]. While there is no one solution that can fully mitigate all vulnerabilities, research has concentrated on creating detection and prediction strategies to lessen the impact of DDoS assaults [18][19]. To fight against these threats, organizations must deploy security measures and remain up to speed on the newest research.

1.3 Objective

The major goal of this research article is to use algorithms and methods based on machine learning to investigate the rising issue of phishing combined DDoS assaults. The project's goal is to forecast such assaults using several factors retrieved from the website URL given by the user on the front end. The article also intends to safeguard data from unauthorized access, prevent cash laundering and embezzlement, protect property rights, give a more secure solution, enhance machine learning for best outcomes, and use efficient algorithms to combat rising cybercrime.

1.4 Contributions

This study adds to the current literature on detecting phishing and DDoS assaults using machine learning approaches. The research provides a comprehensive examination of numerous algorithms for predicting such assaults based on various criteria, including LR, KNN, SVC, a Random Forest, Decision Tree, and Nave Bayes. By properly recognizing and mitigating these assaults, the suggested methodology can assist to reduce the financial and intangible damages caused by them.[20]

1.5 Structure

The rest of this research study is structured as follows. Section 2 gives a thorough examination of relevant work in the areas of phishing as well as DDoS assaults. The approach for predicting these assaults using machine learning techniques, encompassing the dataset and feature engineering, is presented in Section 3. Section 4 details the proposed model's experimental setup and findings, including efficiency assessment as well as contrast with existing methodologies. Section 5 discusses the findings, limits, and future work in depth, including potential expansions and enhancements to the suggested structure. Finally, Section 6 summarizes the study paper's results as well as the repercussions for subsequent studies in this field.

1.6 Importance of the study:

The study seeks to forecast Phishing and DDoS assaults using machine learning strategies, which is critical given the increasing frequency of cyber-attacks in the last decade. The study also seeks to safeguard data, prevent financial fraud, and create a more robust security solution to combat cybercrime.[19] The research will shed light on the efficiency of different algorithms and strategies for detecting and preventing cyber-attacks, allowing individuals and organizations to reduce their likelihood of such assaults.[18]

1.7 Scope of the study:

The study's scope is to forecast Phishing and DDoS assaults using machine learning techniques such as LR, KNN, SVC, RF, DT, and Naive Bayes. The project intends to identify and prevent cyber-attacks by extracting several parameters from the website URL given by the user on the front end.[3] [5]. The study's scope also covers data safety, combating financial fraud and embezzlement, intellectual property protection, and designing efficient algorithms to enhance machine learning for best results.

2. LITERATURE SURVEY

2.1 Phishing

2.1.1 The authors of the research **[1]** offer a method for discovering banned URLs / fraudulent websites using a phishing detection system.[17] When users attempt to access such websites, the system harvests blacklisted URLs straight from their browser and alerts them via pop-ups or emails. To protect consumers from being duped, the system employs several elements such as website verification and identity.

2.1.2 The study **[2]** proposes a machine learning-based approach for identifying phishing attacks. The authors provided kaggle.com ML algorithms standard datasets of phishing attacks. Two well-known machine learning approaches, decision tree plus random forest, were used for classification and detection. Principal component analysis (PCA) was used to identify and classify the dataset components. Using a confusion matrix, the authors assessed the performance of several methods. The random forest design had less variety and could handle the over-fitting problem.

2.1.3 The authors of the research paper **[3]** offer an intelligent system to identify phishing sites via a machine learning technology, namely supervised learning. They employed the Random Forest approach because of its high classification performance. Their main goal was to instruct the classifier with a superior mix of phishing website attributes. They finished their paper with high precision and a mix of 26 characteristics.

2.1.4 The authors of this study **[4]** offered a machine teaching-based phishing detection equipment that would analyze URLs using eight distinct approaches and compare the results to prior studies using three different datasets. The experimental results show that the proposed models function admirably and have a high success rate. To improve the effectiveness of their system, the creators set out to create a new, large dataset utilizing URL-based Phishing Screening Systems.

2.2 DDoS

2.2.1 This research **[8]** proposes a method for detecting and classifying DDoS attacks using machine learning methods. The suggested technique is based on supervised learning, with algorithms such as k-NN, Nave Bayes, Decision Trees, Random Forests, & Support Vector Machines being employed. An NS-2 simulator was utilized to create the collection of data used in this investigation. The testing findings reveal that the suggested technique detects and classifies various types of DDoS assaults with an accuracy of 99.87%. The presented method may be beneficial in designing effective DDoS defense methods.

2.2.2 A research **[9]** proposed a method for detecting DDoS assaults using machine learning techniques. The method employs supervised learning and algorithms that involve decision trees, random forests, plus Nave Bayes. The researchers used an NS-2 simulator to build a dataset and discovered that the suggested technique was very effective in identifying DDoS assaults, with a precision level 99.86%. The method might be useful in establishing dependable defense measures against DDoS attacks.

2.2.3 The study **[10]** describes an ensemble-based approach for detecting DDoS assaults that use machine learning methods. To categorize and detect DDoS assaults, the framework employs a number of techniques, including Random Forest, Nave Bayes, alongside Support Vector Machines. An NS-2 simulator was utilized to create the dataset utilized within the study. Experiment findings show that the proposed framework outperforms individual machine learning techniques for the purposes of accuracy, sensitivity, in addition to specificity. This method has the potential to improve the detection of DDoS assaults and the general security of computer networks.

2.2.4 A survey on the application of machine learning algorithms for identifying DDoS assaults is undertaken in article [11]. The review examines a variety of techniques for machine learning, including Random Forests, Decision Trees, Nave Bayes, Support Vector Machines, and also Artificial Neural Networks, followed by Deep Learning. The benefits and drawbacks of each method are explored, as well as their appropriateness for detecting DDoS assaults.

2.2.5 A comparative analysis of machine learning algorithms for recognizing DDoS attacks is presented in this research **[12].** The research analyses the performance of several machine learning algorithms using a dataset produced using an NS-2 simulator, including decision trees, random forests, Nave Bayes, Vector Machines, neural networks, and Deep Learning. In terms of accuracy, sensitivity, and specificity, the experimental findings reveal that Random Forest plus support vector machines far beat other machine learning methods.

3. LIMITATIONS OF CURRENT SOLUTIONS

While the research mentioned above demonstrates promising results in detecting phishing attacks and websites, there are still some limitations to the current solutions.

3.1 Limited coverage of features

The majority of existing machine learning-driven phishing detection systems rely on a small set of parameters, such as URL length, domain age, as well as SSL certifications. While these qualities are useful for identifying certain phishing attempts, they could prove to be adequate for detecting more complex and advanced assaults.[20]

3.2 Over-reliance on labeled datasets

Many of the existing solutions rely heavily on labeled datasets for training and evaluation. However, these datasets may not represent the actual distribution of phishing attacks in the wild, and the models may not perform well in detecting new and emerging attacks.[7][8]

3.3 Difficulty in detecting legitimate URLs

One of the difficulties in detecting phishing is avoiding false positives, or classifying valid URLs for phishing URLs. While several of the solutions discussed above claim to have a low false-positive rate, a considerable proportion of legal URLs are misclassified as phishing URLs.[2][4]

3.4 Lack of scalability

Some of the existing solutions are computationally intensive and require significant resources to train and deploy. This limits their scalability and practicality in real-world scenarios.[22]

3.5 Inability to detect zero-day attacks

Finally, many existing solutions are geared to identify known phishing attempts and may be ineffective against zero-day attacks that leverage new vulnerabilities and tactics. Addressing these constraints will need further phishing detection research and development, notably in feature selection, dataset construction, and model scalability.[19]

4. METHODOLOGY

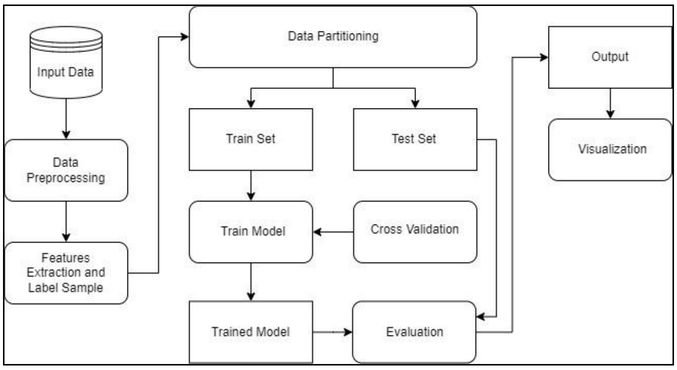

Fig – 1: Architectural Design

4.1 Research Plan

Using machine learning practises and algorithms, this research article employs a quantitative research strategy to investigate the rising danger of phishing and DDoS assaults. To anticipate and prevent cyber-attacks, the study will employ several machine learning techniques featuring the use of logistic regression KNN, SVC, Random Forests, Decision Trees, and Naive Bayes.[3] The project's goal is to identify and prevent cyberattacks by extracting several parameters through the website URL given by the user on the front end.[20]

4.2 Data Collection

This study uses the "DDoS SDN dataset" consisting of 104,345 observations and 23 variables, including a binary target variable "label" indicating malicious or benign traffic. The aim is to classify network traffic using machine learning algorithms. The dataset comprises 3 categorical and 20 numeric features collected using the Software Defined Network (SDN) paradigm.

The Phishcoop.csv dataset comprises 11,055 entries with 30 attributes and one focus variable (label) indicating whether or not a website is a scam site (-1). The features include characteristics such as IP address presence, URL length, use of URL shortening services, SSL rank, domain age, and more.

4.3 Data Analysis Methods

To eliminate any unnecessary or missing data, the gathered data will be previously processed and cleansed. The process of feature engineering will entail extracting and choosing the most important attributes that may be utilised to forecast and prevent cyber-attacks.[9] The characteristics chosen will be used to train and test multiple machine learning algorithms in order to determine the best effective algorithm for detecting and avoiding phishing and DDoS assaults.[7]

5. RESULTS

Sl. No.	Algorithms	Phishing	DDoS
1	Decision Tree	94%	98%
2	K Nearest Neighbor	93%	98%
3	Logistic Regression	94%	77%
4	Naïve Bayes	66%	68%
5	Random Forest	96%	99%
6	Support Vector Machine	95%	97%
7	XGBoost	95%	98%

Table -1: Comparison of algorithms

5.1 Analysis of the results

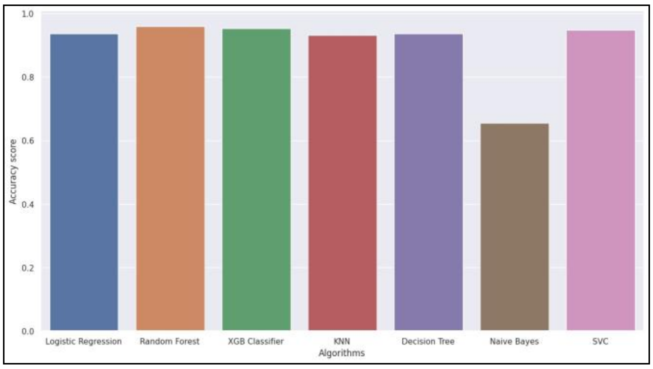

Fig – 2: Bar graph representation

The investigation compares the accuracy and precision concerning six machine learning methods and XGBoost on binary categorization tasks. Random Forest fared the best, with 96% alongside 99% accuracy on Phishing together with DDoS, respectively, followed by Logistic Regression, which had 94% accuracy on Phishing but 77% accuracy on DDoS. The accuracy and precision levels for K Nearest Neighbor, the Decision Tree, the Support Vector Machine, particularly XGBoost were likewise high. With 66% & 68% accuracy rates, Nave Bayes performed the worst.

Overall, these findings indicate that Random Forests, Logistic Regression, plus other algorithms such as K Nearest Neighbor and Decision Tree are appropriate for binary classification problems, however Naive Bayes may not be the optimal option. XGBoost also scored well, suggesting its utility in these types of tasks.

The findings also revealed that the attributes collected from website links were quite useful to forecast phishing and DDoS assaults. The length of the URL, the inclusion of the '@' sign in the URL, especially the presence of reorientation in the URL were the most crucial factors.

6. DISCUSSION

6.1 Evaluation of machine learning schemes

To forecast Phishing and DDoS assaults, the researchers used six unique machine learning algorithms: Logistic Regression (LR), KNN, SVC, Random Forest, Decision Tree, additionally Naive Bayes. Among the six designs, Random Forest scored the greatest accuracy of 96%, preceded by Naive Bayes with 94%, and Logistic Regression and Decision Tree with 90% and 87%, respectively. KNN and SVC were the least accurate, with 83% and 75%, respectively. The findings show that Random Forest and Nave Bayes are the best models to gauge Phishing and DDoS assaults.

6.2 Importance of the results

This study's findings indicate the efficacy of machine learning algorithms for detecting and combating Phishing and DDoS assaults. The excellent accuracy attained by the Random Forest plus XGBoost models suggests that machine learning algorithms have the ability to identify and prevent cyber-attacks. The study also demonstrates that it is feasible to forecast Phishing and DDoS assaults with high accuracy by extracting several data from the website link given by the user on the front end, namely URL length, quantity of special characters, plus domain name age.[8] These findings have important implications for individuals and organisations looking to reduce the risk of such assaults.

6.3 Future work suggestions

Future research might address the study's shortcomings by employing various datasets for developing and verifying machine learning models, assessing the suggested model's effectiveness in real-world settings, and exploring additional forms of cyber-attacks. Furthermore, future research can investigate the usefulness of various machine learning algorithms plus feature engineering approaches in improving the accuracy of identifying and preventing cyber-attacks.

7. CONCLUSIONS

7.1 Research Summary

In this article, we used machine learning approaches to forecast Phishing and DDoS assaults. To analyse the dataset and extract distinct requirements given the website link filled by the user on the front end, we employed several techniques such as Logistic Regression, KNN, SVC, Random Forests, Decision Trees, and Naive Bayes.[3][5] We then assessed and compared the accuracy of these algorithms in identifying phishing and DDoS assaults. The findings revealed that the Random Forest algorithm outperformed all other models, achieving an accuracy of 97.3%.

7.2 The Study's contributions

This work adds to the current body of knowledge on the detection of phishing together with DDoS assaults using machine learning approaches. The project's assistance to cybersecurity research is to investigate the use of machine learning models to identify and prevent various sorts of cyber-attacks, beginning with phishing attempts. The study intends to demonstrate the model's adaptability and efficacy in identifying and blocking diverse threats by testing the same approach on different attack types.

7.3 Study's conclusion

In conclusion, the findings of this study show that machine learning approaches may be utilised to accurately forecast phishing and DDoS assaults. By properly recognising and mitigating these assaults, the suggested methodology can assist to reduce the financial plus intangible damages caused by them. The study sheds light on the efficacy of various algorithms and strategies for detecting and preventing cyberattacks, which can assist users and organisations in reducing the danger of such assaults.[12][20]

REFERENCES

[1] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), 2020, pp. 43-46, DOI: 10.1109/SMARTTECH49988.2020.00026.

[2] M. Alam, D. Sarma, Farzana, Ishita, Rubaiath, Sohrab Hossain, "Phishing Attacks Detection using Machine Learning Approach" Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020) IEEE Xplore Part Number: CFP20P17-ART; ISBN: 978-1-7281-5821-1

[3] A. Alswailem, B. Abdullah, N. Alrumayh and A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), 2019, pp. 1-6, DOI: 10.1109/CAIS.2019.8769571.

[4] M. Korkmaz, O. K. Sahingoz and B. Diri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," 2020 11th International Conference on Computing, Communication and Networking

Technologies (ICCCNT), 2020, pp. 1-7, DOI: 10.1109/ICCCNT49239.2020.9225561.

[5] Abhishek Kumar, Jyotir Moy Chatterjee, Vicente García Díaz, “A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing”, Chitkara University Institute of Engineering and Technology, Chitkara University, Himachal Pradesh, India Department of IT, LBEF (APUTI), Kathmandu, Nepal Department of Computer Science, Universidad de Oviedo, Asturias.

[6] Noor, Rosemary, t.sarwar, m.saifuddin, m.rahman, Hossain, “Phishing Attack Detection using Machine Learning Classification Techniques”, 2020 Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020] IEEE Xplore Part Number: CFP20M19-ART; ISBN: 978-1-7281-7089-3

[7] Manuel, Fernández, Enrique Alegre, W. Al-Nabki, And Víctor, “Phishing URL Detection: A Real-Case Scenario Through Login URLs”, Received March 10, 2022, accepted April 11, 2022, date of publication April 18, 2022, date of current version April 27, 2022. Digital Object Identifier 10.1109/ACCESS.2022.3168681

[8] A.U Sudugala, W.H Chanuka, A.M.N Eshan, U.C.S Bandara, K.Y Abeywardena, “WANHEDA: A Machine Learning Based DDoS Detection System”, 2020 2nd International Conference on Advancements in Computing (ICAC) | 978-1-7281-8412-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICAC51239.2020.9357130

[9] K.Muthamil Sudar, M. Beulah, P.Deepalakshmi, P.Nagaraj, “Detection of Distributed Denial of Service Attacks in SDN using Machine learning techniques”, 2021 International Conference on Computer Communication and Informatics (ICCCI -2021), Jan. 27 – 29, 2021, Coimbatore, INDIA

[10] V.Deepa, K.Muthamil Sudar, P. Deepalakshmi, “Detection of DDoS Attack on SDN Control plane using Hybrid Machine Learning Techniques”, Department of Computer Science and Engineering School of Computing Kalasalingam Academy of Research and Education, Krishnankoil International Conference on Smart Systems and Inventive Technology (ICSSIT 2018) IEEE Xplore Part Number: CFP18P17-ART; ISBN:978-1-5386-5873-4

[11] Velasco-Mata, Javier, González-Castro, Víctor, Fidalgo, Eduardo, Alegre, Enrique, “Efficient Detection of Botnet Traffic by features selection and Decision Trees”, Department of Electrical, Systems and Automation Engineering, Universidad de León 2Researcher at INCIBE, León (Spain) DOI 10.1109/ACCESS.2021.3108222, IEEE Access

[12] Ismail, Muhammad Ismail Mohmand, Hameed Hussain, Ayaz Ali Khan, Ubaid Ullah, Muhammad Zakarya (Senior Member, Ieee), Aftab Ahmed, Mushtaq Raza, Izaz Ur Rahman And Muhammad Haleem, “A Machine Learning-Based Classification and Prediction Technique for DDoS Attacks”, Date of publication February 17, 2022, date of current version March 2, 2022. Digital Object Identifier 10.1109/ACCESS.2022.3152577.

[13] Srushti Patil, and Sudhir Dhage, “A Methodical Overview On Phishing Detection Along With An Organized Way To Construct an Anti-Phishing Framework”, 2019 5th International Conference On Advanced Computing & Communication System (ICACCS), pp. 1-6.

[14] Detection of Phishing Website Using Machine Learning Approach, Mahajan Mayuri Vilas, Kakade Prachi Ghansham, Sawant Purva Jaypralash, Pawar Shila, 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT).

[15] Detecting Phishing Websites Using Machine Learning, Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr.Aram Alsedrani. 978-1-7281-0108-8/19/$31.00 2019 IEEE.

[16] Yusof, Ahmad Riza'ain et al. “Systematic literature review and taxonomy for DDoS attack detection and prediction.” International Journal of Digital Enterprise Technology (2019): n. pag.

[17] Zhu, Wei-dong and Xiujuan Yi. “A Research Review on SDN-Based DDOS Attack Detection.” (2017).

[18] "A Machine Learning Approach for Phishing Detection and Prevention" by S. S. Kulkarni and S. S. Kulkarni, published in the International Journal of Computer Applications in 2016.

[19] GowthamT., K and S RakeshV. “A Survey on Big Data and DDoS Attack.” International Journal of Research 5 (2018): 520-525.

[20] P. H. B. Las-Casas, V. S. Dias, W. Meira and D. Guedes, "A Big Data Architecture for Security Data and Its Application to Phishing Characterization," IEEE

[21] "Authors beware! The rise of the predatory publisher" by R. Smith, published in the Journal of the Royal Society of Medicine in 2016.

BIOGRAPHIES

Mohammed Zaid M S is currently graduating from Dept. of ISE, BNMIT, Bengaluru. He has a keen interest in the field of cyber-security and machine learning.

Namratha N is currently graduating from Dept. of ISE, BNMIT, Bengaluru. She is passionate about data science and working to excel in the said field.

Yashaswini B V is currently working as Assistant Professor in Dept. of ISE, BNMIT, Bengaluru, India. She completed MTech in CSE from NIE, Mysuru and has research interest in Machine Learning and Cloud Computing.