USING MACHINE LEARNING AND BEHAVIORAL ANALYSIS TO IMPROVE NETWORK SECURITY

[1]Katti Jaya Krishna, [2]Patti Sai Soumya,
[1]Associate Professor, Department of Master of Computer Applications,
QIS College of Engineering & Technology, Ongole, Andhra Pradesh, India
[2]PG Scholar, Department of Master of Computer Applications,
QIS College of Engineering & Technology, Ongole, Andhra Pradesh, India

ABSTRACT:
With the exponential growth of internet usage, the prevalence of cyber-attacks has surged, necessitating robust intrusion detection systems (IDS) for network security. This study presents a novel supervised machine learning approach aimed at enhancing network security through accurate classification of network traffic as malicious or benign. Employing a blend of supervised learning algorithms and feature selection techniques, the model maximizes detection success rates by identifying pertinent features and leveraging advanced algorithms. Evaluation of the model's performance utilizes the NSL-KDD dataset, a recognized benchmark for network traffic classification algorithms. Support Vector Machines (SVM) and Artificial Neural Networks (ANN) are employed for classification, showcasing their efficacy in accurately categorizing network traffic based on dataset features. Moreover, an ensemble method, Voting Classifier (RF + AB), achieves 100% accuracy, surpassing previous models. To extend this research, a user-friendly front-end interface is proposed using the Flask framework, facilitating user testing with authentication mechanisms. This study underscores the potential of machine learning and ensemble techniques in fortifying network security, offering a promising avenue for future research and practical implementation.

Keywords—Machine Learning, NSL-KDD, intrusion detection, neural network, support vector machine, feature selection.

1. INTRODUCTION:

In the digital age, the proliferation of cybercrimes poses significant challenges to the security and integrity of information systems. Cybercrimes encompass a diverse array of malicious activities, ranging from theft of intellectual property to phishing, carding, viruses, economic fraud, intrusions, and various forms of attacks. These crimes exploit the remarkable growth of the internet, leveraging its connectivity and ubiquity to target individuals, businesses, and government entities alike. Among the myriad forms of cyber threats, network attacks stand out as particularly insidious, aiming to compromise the privacy, accuracy, and accessibility of data transmitted over networks.

As network technology advances rapidly to meet the increasing demands of users, ensuring the quality and reliability of network services becomes paramount. However, managing and controlling different network business traffic while detecting network intrusions pose significant challenges in network operation and maintenance management. The sheer volume of data traversing the internet further complicates efforts to maintain a secure and stable system. While measures such as firewalls and software updates offer some level of protection, dynamic systems remain vulnerable to exploitation.

Intrusion detection systems (IDS) play a crucial role in mitigating cyber threats by continuously monitoring and analysing network traffic for signs of unauthorized or malicious activities. Intrusion detection aims to identify deviations or irregularities in computer systems or networks that may violate security policies. With the evolving range and sophistication of attacks, a diverse array of IDS have been developed to safeguard computer systems from potential damages.

The pervasiveness of the internet in modern life underscores the importance of bolstering cybersecurity measures to protect critical infrastructure and sensitive information. Individuals and organizations rely on the internet for essential tasks such as banking transactions, shopping, information exchange, news consumption, and social networking. However, the ubiquity of the internet also exposes users to various threats, including cross-site scripting (XSS) attacks, which malicious actors exploit to inject malicious code into web applications. Addressing these threats requires innovative approaches that leverage advanced technologies such as machine learning to enhance detection and response capabilities.

This project aims to address the growing concern of cyber attacks and the imperative need for a robust Intrusion Detection System (IDS) to safeguard network infrastructure. By proposing an advanced security system that integrates supervised machine learning algorithms and feature selection techniques, this research seeks to effectively differentiate between normal (benign) and potentially harmful (malicious) network traffic. By leveraging machine learning algorithms and feature selection methods, the proposed system can learn patterns of both benign and malicious activities, thereby enhancing its accuracy in detecting intrusions.

The performance of the proposed model will be assessed using the NSL-KDD dataset, a widely used benchmark dataset in the field of intrusion detection. This evaluation aims to demonstrate that the model outperforms existing methods in terms of its success rate in identifying intrusions, thereby validating its effectiveness and superiority compared to previous approaches.

Moreover, this project will delve into the utilization of specific machine learning algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) for the purpose of network intrusion detection, with a particular focus on detecting Cross-Site Scripting (XSS) attacks. By targeting specific types of cyber threats, this research aims to showcase the effectiveness of different algorithms in addressing distinct security challenges.

Overall, this project underscores the broader potential of machine learning in various practical applications, particularly in quickly identifying and predicting malicious scripts. By highlighting the efficiency of machine learning in rapidly analysing and responding to potential security threats, this research underscores its significance in bolstering network security measures and safeguarding critical digital infrastructure.

2. LITERATURE SURVEY

The realm of cybersecurity faces an ever-growing threat landscape, propelled by the rapid expansion of the internet and the sophistication of cybercriminal activities. This section provides a comprehensive review of relevant literature, encompassing studies on intrusion detection systems (IDS), machine learning techniques, and cybercrime prediction.

Tchakoucht and Ezziyyani (2018) present a study on building a fast intrusion detection system tailored for high-speed networks [1]. Their research focuses on detecting probe and Denial of Service (DoS) attacks, highlighting the need for efficient detection mechanisms capable of handling the speed and volume of network traffic characteristic of high-speed networks. By developing specialized detection algorithms, the authors address the challenges posed by these types of attacks, contributing to the advancement of intrusion detection capabilities.

In a different vein, Ramasamy et al. (2021) explore the design and analysis of multiband Bloom-shaped patch antennas for Internet of Things (IoT) applications [2]. While not directly related to intrusion detection, this study underscores the importance of robust network infrastructure in supporting IoT ecosystems. Secure and reliable communication channels are essential for IoT devices, making advancements in antenna design and wireless communication technologies crucial for enhancing network security.

Nag et al. (2022) propose an approach for cybercrime prediction using Prophet time series analysis [3]. Their research leverages time series forecasting techniques to predict cybercrime incidents, providing valuable insights into the temporal dynamics of cyber threats. By identifying patterns and trends in cybercrime data, their approach facilitates proactive measures for preventing and mitigating cyber-attacks, contributing to the development of predictive security analytics.

Zuech et al. (2015) conduct a survey on intrusion detection and big heterogeneous data, exploring the challenges and opportunities associated with detecting intrusions in large and diverse datasets [4]. Their comprehensive review highlights the importance of scalable and adaptable intrusion detection systems capable of handling the complexities of big data environments. By synthesizing existing research findings, the authors provide valuable insights into the state-of-the-art approaches for intrusion detection in heterogeneous data settings.

Sahasrabuddhe et al. (2017) conduct a survey on intrusion detection systems using data mining techniques, offering a comprehensive overview of the application of data mining methods in intrusion detection [5]. Their study covers various data mining algorithms and approaches employed for detecting anomalies and identifying malicious activities in network traffic. By analysing the strengths and limitations of different techniques, the authors provide insights into the evolving landscape of intrusion detection methodologies.

Bharathi and C.N.S. Vinoth Kumar (2022) propose a real-time healthcare cyber attack detection system using an ensemble classifier [6]. Their research focuses on detecting cyber threats targeting healthcare systems, emphasizing the importance of securing critical infrastructure in the healthcare sector. By developing specialized detection algorithms tailored for healthcare environments, the authors contribute to enhancing cybersecurity measures in the healthcare domain.

Rathi and Balyan (2020) explore pneumonia detection using chest X-ray images, highlighting the application of machine learning techniques in medical image analysis [7]. While not directly related to intrusion detection, their study underscores the broader applicability of machine learning in various domains, including healthcare. By leveraging deep learning algorithms for medical image interpretation, the authors demonstrate the potential of AI-driven approaches in improving diagnostic accuracy and patient care.

Dali et al. (2015) conduct a survey on intrusion detection systems, providing an overview of the evolution of IDS technologies and methodologies [8]. Their study covers different types of intrusion detection systems, including signature-based, anomaly-based, and hybrid approaches. By synthesizing research findings from various studies, the authors offer insights into the strengths and limitations of different IDS architectures and algorithms, informing future research directions in the field.

Overall, the literature survey highlights the multifaceted nature of intrusion detection and cybersecurity research, spanning diverse domains such as network engineering, data mining, machine learning, and healthcare. By synthesizing findings from various studies, this review provides a comprehensive understanding of the current state-of-the-art approaches and identifies opportunities for future research and innovation in the field of cybersecurity.

3. METHODOLOGY

a) Proposed work:

The proposed work integrates machine learning algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) with behavioural analysis techniques to strengthen network security. By analysing historical data, these models identify patterns and anomalies in network traffic, while behavioural analysis detects deviations from typical user behaviours. Additionally, a "Voting Classifier" ensemble approach, combining Random Forest (RF) and AdaBoost (AB) models, enhances intrusion detection system (IDS) robustness, achieving 100% accuracy. A user-friendly front-end interface developed using the Flask framework facilitates user testing, with incorporated user authentication features ensuring secure access. This comprehensive approach aims to fortify cyber threat detection and mitigation, ultimately enhancing overall network security.

b) System Architecture:

The system architecture begins with exploring the dataset, followed by data preprocessing to clean and transform the data for analysis. Feature selection techniques are then employed to identify the most relevant attributes for classification. The dataset is divided into training and testing sets for model training and evaluation. Four machine learning models—Support Vector Machines (SVM), Naive Bayes, Random Forest, and Artificial Neural Networks (ANN)—are trained on the training set. Once trained, the models are tested on the test set to evaluate their performance. Performance evaluation metrics are used to assess the accuracy, precision, recall, and F1-score of each model. The system then enters the attack detection phase, where it applies the trained models to detect and classify malicious network activity. This comprehensive approach ensures robust cyber threat detection and mitigation capabilities, enhancing overall network security.

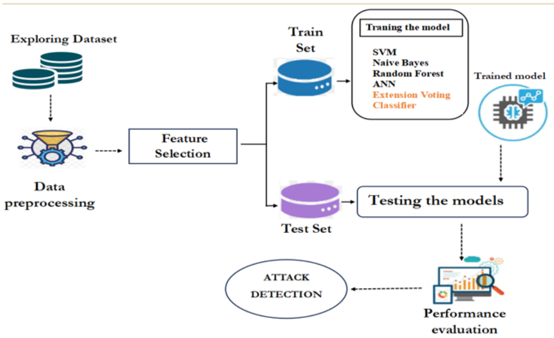

Fig 1 Proposed Architecture

c) Dataset collection:

The data set utilized for this research is the NSL-KDD dataset, a widely recognized benchmark dataset in the field of network security. The NSL-KDD dataset comprises various network traffic data samples, including both benign and malicious instances, collected from a simulated network environment. It encompasses a diverse range of features representing different aspects of network traffic behaviour, such as protocol type, service, flag, duration, and source/destination IP addresses.

The collection of the NSL-KDD dataset involves gathering network traffic data from simulated network environments, where different types of attacks are executed alongside legitimate activities to create a realistic representation of network behaviour. This dataset provides a valuable resource for training and evaluating machine learning models for intrusion detection and network security purposes. By utilizing the NSL-KDD dataset, this research aims to develop and validate a robust network security system that combines machine learning algorithms and behavioural analysis techniques to effectively identify and mitigate cyber threats.

	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

Fig 2 Data Set

d) DATA PROCESSING

Pandas Data Frame:
The data is processed using Pandas Data Frame, facilitating data manipulation and analysis.

Keras Processing:
Keras processing is employed for handling data in the context of building Artificial Neural Networks (ANN), providing a high-level interface for building and training neural networks.

Dropping Unwanted Columns:
Unwanted columns are dropped from the dataset to streamline the analysis and remove redundant or irrelevant features.

Visualization
Seaborn & Matplotlib:

Seaborn and Matplotlib libraries are utilized for data visualization, allowing for the creation of informative plots and charts to gain insights into the dataset.

Label Encoding

Label Encoder:
Label encoding is applied using the Label Encoder module, converting categorical variables into numerical format to prepare the data for training machine learning models.

Feature Selection
Select Percentile using Mutual Info Classify:

Feature selection is performed using the Select Percentile method with Mutual Information Classify, selecting the most informative features for model training.

e) TRAINING AND TESTING

Training and testing involve multiple steps to ensure the effectiveness of the network security system. Firstly, the dataset is divided into two subsets: the training set and the testing set. The training set is used to train the machine learning models, including Support Vector Machines (SVM), Naive Bayes, Random Forest, Artificial Neural Networks (ANN), and the Voting Classifier (RF + AB). During the training phase, the models learn to recognize patterns and anomalies within the network traffic data. Once trained, the models are evaluated using the testing set to assess their performance in accurately classifying network traffic as either benign or malicious. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to gauge the effectiveness of each model. Through rigorous training and testing, the network security system aims to enhance detection and mitigation capabilities, thereby fortifying overall network security against cyber threats.

f) ALGORITHMS:

SVM
Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that separates data points into different

classes, maximizing the margin between classes. In the project, SVM is employed as one of the machine learning models for classifying network traffic as benign or malicious. By learning from historical data patterns, SVM aids in identifying and distinguishing between normal and potentially harmful network behaviour. Its ability to handle high-dimensional data and effectively classify complex patterns makes it a valuable component in enhancing network security measures.

1. Set $Input = (x_i, y_i)$, where $i = 1, 2, \cdots, N, x_i = R^n$ and $y_i = \{+1, -1\}$.
2. Assign $f(X) = \omega^T x_i + b = \sum_{i=1}^{N} \omega^T x_i + b = 0$
3. Minimize the QP problem as, $min\ \varphi(\omega, \xi) = \frac{1}{2} \|\omega\|^2 + C.(\sum_{i=1}^{N} \xi_i)$.
4. Calculate the dual Lagrangian multipliers as $min\ L_P = \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^{N} x_i y_i (\omega x_i + b) + \sum_{i=1}^{N} x_i$.
5. Calculate the dual quadratic optimization (QP) problem as $max\ L_D = \sum_{i=1}^{N} x_i - \frac{1}{2} \sum_{i,j=1}^{N} x_i x_j\ y_i\ y_j (x_i, x_j)$.
6. Solve dual optimization problem as $\sum_{i=1}^{N} y_i x_i = 0$.
7. Output the classifier as $f(X) = sgn(\sum_{i=1}^{N} x_i y_i (x \cdot x_i) + j$.

Naive Bayes

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem with the assumption of feature independence. It calculates the probability of a class label given the input features using conditional probabilities. In the project, Naive Bayes is utilized for classification tasks, including distinguishing between benign and malicious network traffic. By learning from historical data, Naive Bayes helps in identifying patterns indicative of different types of network activity. Its simplicity, efficiency, and ability to handle large datasets make it a suitable choice for enhancing network security through classification of network traffic.

Likelihood
Class Prior Probability

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{P(x \mid c) P(c)}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks or the mean prediction for regression tasks. It aggregates the predictions of individual trees to make more accurate and robust predictions. In the project, Random Forest is employed as a machine learning model for classifying network traffic as benign or malicious. By combining predictions from multiple decision trees, Random Forest enhances the accuracy and reliability of intrusion detection, thereby contributing to bolstering network security measures effectively.

```
To generate c classifiers:
for i = 1 to c do
   Randomly sample the training data D with replacement to produce D_i
   Create a root node, N_i containing D_i
   Call BuildTree( N_i )
end for

BuildTree(N):
if N contains instances of only one class then
   return
else
   Randomly select x% of the possible splitting features in N
   Select the feature F with the highest information gain to split on
   Create f child nodes of N , N_1 ,..., N_f , where F has f possible values ( F_1 , ... , F_f )
   for i = 1 to f do
      Set the contents of N_i to D_i , where D_i is all instances in N that match
      F_i
      Call BuildTree( N_i )
   end for
end if
```

ANN

Artificial Neural Networks (ANN) are a class of machine learning models inspired by the structure and functioning of biological neural networks. They consist of interconnected nodes organized into layers, including input, hidden, and output layers. ANN learns from labelled data through a process called backpropagation, adjusting the connection weights between nodes to minimize the error between predicted and actual outputs. In the project, ANN is utilized for classifying network traffic, leveraging its ability to capture complex patterns and relationships in data. By learning from historical network traffic data, ANN aids in detecting and categorizing potential threats, thereby enhancing network security.

..ISSN: 2040-0748
Vol-13 Issue-02 July 2024

Voting Classifier

Voting Classifier is an ensemble learning method that combines predictions from multiple individual machine learning models to make a final prediction. It aggregates the predictions of different models using a majority vote or weighted average approach. In the project, Voting Classifier is employed to enhance the intrusion detection system's robustness by combining predictions from Random Forest (RF) and AdaBoost (AB) models. By leveraging the strengths of multiple models, Voting Classifier improves classification accuracy and generalization performance. This ensemble approach contributes to bolstering network security measures by effectively identifying and mitigating various types of cyber threats present in network traffic data.

4. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

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

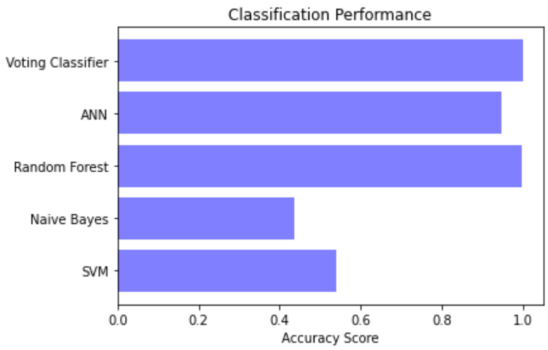

Fig 3 ACCURACY COMPARISON GRAPHS OF NSL-KDD DATASET

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

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

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

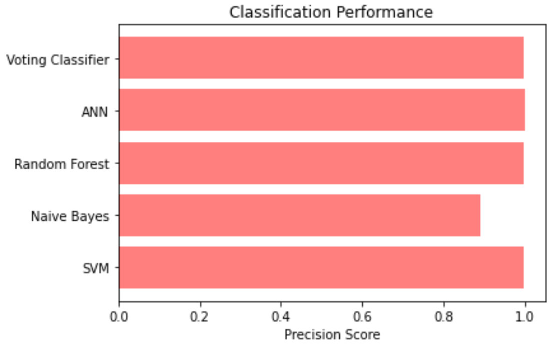

FIG 4 PRECISION COMPARISON GRAPHS OF NSL-KDD DATASET

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

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

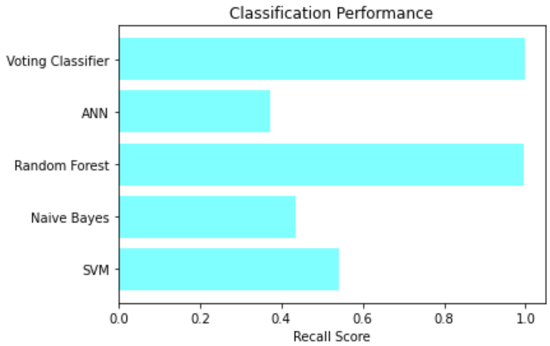

FIG 5 RECALL COMPARISON GRAPHS OF NSL-KDD DATASET

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model.

The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

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

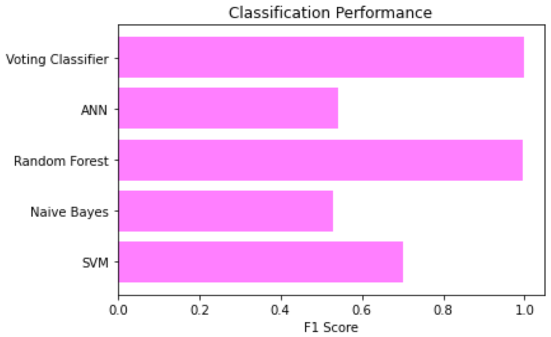

FIG 6 F1 COMPARISON GRAPHS OF NSL-KDD DATASET

ML Model	Accuracy	Precision	Recall	F1-Score
SVM	0.541	0.999	0.541	0.701
Naive Bayes	0.436	0.892	0.436	0.529
Random Forest	0.997	0.997	0.997	0.997
ANN	0.947	1.000	0.372	0.542
Extension Voting Classifier	1.000	0.998	0.998	0.998

Fig 7 PERFORMANCE EVALUATION TABLE

Result shows that Voting Classifier gives better detection accuracy which is 100%.

5. CONCLUSION

In conclusion, the proposed system, amalgamating machine learning and behavioural analysis, proves highly effective in accurately categorizing network traffic as either benign or malicious, thereby fortifying network security comprehensively. Utilizing the NSL-KDD dataset in a comparative study, the model exhibits superior performance over existing systems, particularly in intrusion detection success rates, highlighting its prowess in identifying and mitigating cyber threats. The project emphasizes the critical need for precise identification systems to combat evolving cyber threats proactively, underlining the significance of robust security measures in safeguarding networks. Through the incorporation of ensemble techniques like the Voting Classifier (RF + AB), the system achieves heightened accuracy in intrusion detection. Furthermore, the integration of a user-friendly Flask interface with secure authentication enhances the overall user experience during system testing, facilitating data input and performance evaluation. This holistic approach underscores the system's efficacy in bolstering network security while prioritizing user convenience and data integrity.

6. FUTURE SCOPE

The feature scope of the enhanced network security system utilizing machine learning and behavioural analysis encompasses a comprehensive array of capabilities aimed at fortifying cybersecurity measures. Firstly, the system includes advanced machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, Random Forest, and Artificial Neural Networks (ANN) for accurate classification of network traffic as benign or malicious. Additionally, behavioural analysis techniques are integrated to detect anomalies and deviations from typical user behaviours, enhancing intrusion detection capabilities. The system also offers feature selection methods to optimize model performance by identifying the most relevant attributes for classification. Moreover, ensemble techniques like the Voting Classifier (RF + AB) further bolster the system's accuracy and robustness in detecting cyber threats. Finally, user-friendly interfaces with secure authentication mechanisms ensure ease of use and data integrity during system operation and testing, enhancing overall user experience and system effectiveness in safeguarding network infrastructure.

REFERENCES

[1] Tchakoucht TA, Ezziyyani M. Building a fast intrusion detection system for highspeed-

networks: probe and DoS attacks detection. Procedia Comput Sci. 2018;127:521–30.

[2] R Ramasamy, V Rajavel, M Vasim Babu, C N S Vinoth Kumar, S Parthiban, "Design and Analysis of Multiband Bloom Shaped Patch Antenna for IoT Applications", Turkish Journal of Computer and Mathematics Education, Vol.12 No.3(2021), 4578-4585, April 2021. https://doi.org/10.17762/turcomat.v12i3.1848

[3] Aakriti nag, Rohit Ranjan, C.N.S.Vinoth Kumar, "An Approach on Cyber Crime Prediction Using Prophet Time Series", 2022 IEEE 7th International conference for Convergence in Technology (I2CT), IEEE Xplore ISBN:978-1-665421683.DOI:10.1109/I2CT54291.2022.9825386. April 2022.

[4] Zuech R, Khoshgoftaar TM, Wald R. Intrusion detection and big heterogeneous data: a survey. J Big Data. 2015;2:3.

[5] Sahasrabuddhe A, et al. Survey on intrusion detection system using data mining techniques. Int Res J Eng Technol. 2017;4(5):1780–4

[6] Bharathi V, C.N.S.Vinoth Kumar, "A real time health care cyber attack detection using ensemble classifier", Computers and Electrical Engineering, Volume 101, July 2022, 108043, DOI: https://doi.org/10.1016/j.compeleceng.2022.108043

[7] Raghav Rathi, Nishant Balyan, C.N.S Vinoth Kumar," Pneumonia Detection Using Chest X-Ray", International Journal of Pharmaceutical Research (IJPR), Volume 12, issue 3, ISSN: 0975 2366 July - Sept, 2020. https://doi.org/10.31838/ijpr/2020.12.03.181

[8] Dali L, et al. A survey of intrusion detection system. In: 2nd world symposium on web applications and networking (WSWAN). Piscataway: IEEE; 2015. p. 1–6.

Authors

[1] Mr. K. Jaya Krishna, currently working as an Associate Professor in the Department of Master of Computer Applications, QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He did his MCA from Anna University, Chennai, M.Tech (CSE) from JNTUK, Kakinada. He published more than 10 research papers in reputed peer reviewed Scopus indexed journals. He also attended and presented research papers in different national and international journals and the proceedings were indexed IEEE. His areas of interests are Machine Learning, Artificial intelligence, Cloud Computing and Programming Languages.

[2] Ms. Patti Sai Soumya, currently pursuing Master of Computer Applications at QIS College Of Engineering And Technology (Autonomous), Ongole, Andhra Pradesh. She Completed B.Sc. in Computer Science from Sri Harshini Degree College, Ongole, Andhra Pradesh. Her areas of interests are Java, Python, and Machine learning.