Visvesvaraya Technological University

Jnana Sangama, Belagavi - 590018

A Project Work Phase-II (18CSP83)
Report on

"Machine Learning Based Intrusion Detection System"

Project Report submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

MONIESH S	**1KS20CS056**
MONIKA N	**1KS20CS057**
PAVITHRA R	**1KS20CS073**
SINDHURA H	**1KS20CS093**

Under the guidance of

Mr. SOMASEKHAR T
Associate Professor
Department of Computer Science & Engineering
K.S.I.T, Bengaluru-560109

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

K. S. Institute of Technology

#14, Raghuvanahalli, Kanakapura Road, Bengaluru - 560109

2023 - 2024

K. S. Institute of Technology

#14, Raghuvanahalli, Kanakapura Road, Bengaluru - 560109

Department of Computer Science & Engineering

CERTIFICATE

Certified that the Project Work Phase-II (18CSP83) entitled **"Machine Learning Based Intrusion Detection System"** is a bonafide work carried out by:

MONIESH S	**1KS20CS056**
MONIKA N	**1KS20CS057**
PAVITHRA R	**1KS20CS073**
SINDHURA H	**1KS20CS093**

in partial fulfilment for VIII semester B.E., Project Work in the branch of Computer Science and Engineering prescribed by **Visvesvaraya Technological University, Belagavi** during the period of February 2024 to May 2024. It is certified that all the corrections and suggestions indicated for internal assessment have been incorporated in the report deposited in the department library. The Project Work Phase-II Report has been approved as it satisfies the academic requirements in report of project work prescribed for the Bachelor of Engineering degree.

……	……..……	……..……
Signature of the Guide	**Signature of the HOD**	**Signature of the Principal & Director**
[Mr. Somasekhar T]	[Dr. Rekha B. Venkatapur]	[Dr. Dilip Kumar K]

External Viva

Name of the Examiners	**Signature with Date**
1.	
2.	

DECLARATION

We, the undersigned students of 8th semester, Computer Science & Engineering, KSIT, declare that our Project Work Phase-II entitled "**Machine Learning Based Intrusion Detection System**", is a bonafide work of ours. Our project is neither a copy nor by means a modification of any other engineering project.

We also declare that this project was not entitled for submission to any other university in the past and shall remain the only submission made and will not be submitted by us to any other university in the future.

Place:
Date:

Name and USN	**Signature**
MONIESH S (1KS20CS056)	**…….**
MONIKA N (1KS20CS057)	**…….**
PAVITHRA R (1KS20CS073)	**…….**
SINDHURA H (1KS20CS093)	**…….**

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task will be incomplete without the mention of the individuals, we are greatly indebted to, who through guidance and providing facilities have served as a beacon of light and crowned our efforts with success.

First and foremost, our sincere prayer goes to almighty, whose grace made us realize our objective and conceive this project. We take pleasure in expressing our profound sense of gratitude to our parents for helping us complete our Project Work Phase-II successfully.

We take this opportunity to express our sincere gratitude to our college **K.S. Institute of Technology,** Bengaluru for providing the environment to work on our project.

We would like to express our gratitude to our **MANAGEMENT,** K.S. Institute of Technology, Bengaluru, for providing a very good infrastructure and all the kindness forwarded to us in carrying out this project work in college.

We would like to express our gratitude to **Dr. K.V.A Balaji, CEO**, K.S. Institute of Technology, Bengaluru, for his valuable guidance.

We would like to express our gratitude to **Dr. Dilip Kumar K**, **Principal/Director**, K.S. Institute of Technology, Bengaluru, for his continuous support.

We like to extend our gratitude to **Dr. Rekha.B.Venkatapur**, **Professor and Head**, Department of Computer Science & Engineering, for providing a very good facilities and all the support forwarded to us in carrying out this Project Work Phase-II successfully.

We also like to thank our Project Coordinators, **Mr. Roopesh Kumar B N, Associate Professor, Mr. Raghavendrachar S, Associate Professor, Department of Computer Science & Engineering** for their help and support provided to carry out the Project Work Phase-II successfully.

Also, we are thankful to **Mr. Somasekhar T, Associate Professor, Department of Computer Science & Engineering**, for being our Project Guide, under whose able guidance this project work has been carried out and completed successfully.

We are also thankful to the teaching and non-teaching staff of Computer Science & Engineering, KSIT for helping us in completing the Project Work Phase-II work.

MONIESH S
MONIKA N
PAVITHRA R
SINDHURA H

ABSTRACT

A machine learning (ML)-enabled intrusion detection system (IDS) is designed to strengthen cyber security procedures by leveraging user-system interactions to examine potentially malicious URLs. Using a pro-active model, the system schedules links to be sent from a source system to a user system for careful inspection. Advanced machine learning algorithms determine the link's safety quotient upon reception. If deemed safe, the program collects relevant site information to inform the user; if considered dangerous, a notification is sent to the link provider. In addition, in order to mitigate potential threats, the system automatically blocks access to the website that has been identified and notifies the user, so creating a barrier against attempts by unauthorized parties to exfiltrate data.

Keywords: Machine Learning (ML), Intrusion Detection System (IDS), Cyber security, User-system interactions, Malicious links, Proactive approach, Link verification, ML algorithms, Safety analysis.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

The principal aim of the project is to develop an advanced system that can identify dangerous websites by utilizing state-of-the-art machine learning techniques. The system aims to extract a wide range of traits for strong threat identification by incorporating URL-based, content-based, and server-based aspects into its analytic architecture. Establishing strong dataset management procedures is essential to its operation because it guarantees the ongoing validation and curation of data in order to maintain its relevance and accuracy.

Furthermore, the system's adaptability is demonstrated by its capacity to support a wide variety of machine learning algorithms, giving users the freedom to customize their strategy to meet certain dataset requirements and analytical goals. An intuitive interface that facilitates smooth interaction and allows users to easily submit websites for analysis, evaluate results, and provide input for ongoing improvement is developed with a user-centric design philosophy in mind. The project's ultimate goal is to provide users with an effective tool that can proactively detect and neutralize malevolent online entities, improving cyber security resilience in a constantly changing threat landscape.

Intrusion assaults are always changing, it might be difficult to identify malicious websites. Even with the many solutions that industry, academia, and research groups have suggested, phishing assaults still constitute a serious risk. By identifying patterns in the features of websites, machine learning algorithms provide a viable method for differentiating between intrusion and legitimate websites. Nevertheless, a number of drawbacks with current machine learning algorithms for phishing detection reduce their efficacy.

The absence of a thorough framework for feature extraction and updating a collection of legitimate and phishing websites is the first significant limitation. This leads to incomplete and out-of-date datasets, which lowers the accuracy of intrusion detection.

The internet, which provides previously unheard-of connectedness and convenience, has completely changed the way we communicate, work, and live. Nonetheless, there are risks associated with this digital environment. Phishing is a type of cyberattack where hostile actors pose as genuine entities in order to trick users into divulging personal information like usernames, passwords, and credit card numbers. Phishing is one of the most prevalent risks that internet users face today. The prevalence and sophistication of phishing attacks have increased, endangering internet users from identity theft, financial loss, and other types of cybercrime.

Chapter 2

LITERATURE SURVEY

1.Title: "Privacy-Preserving Intrusion Detection in Edge Computing Using Homomorphic Encryption"

Author: Dr. Sophia Lee et al.

Published year: 2024

Description: This study suggests a homomorphic encryption-based intrusion detection system for edge computing environments that protects privacy. The goal of the project is to safeguard confidential network information while facilitating efficient threat mitigation and intrusion detection at the network edge.

Methodology: In order to enable intrusion detection algorithms to function on encrypted data without jeopardizing privacy, the methodology encrypts network traffic data using homomorphic encryption techniques.

Result: By achieving effective intrusion detection while protecting data privacy, the privacy-preserving IDS enables edge computing networks in a variety of application areas to operate securely and effectively.

2.Title: "A Hybrid Machine Learning Approach for Intrusion Detection in Cloud Computing Environments"

Author: Dr. Sarah Patel et al.

Published year: 2023

Description: The study introduces a hybrid machine learning methodology aimed at intrusion detection in cloud computing settings. Through efficient detection and mitigation of harmful activity within cloud infrastructures, the study seeks to improve cybersecurity.

Methodology: To detect anomalies in cloud traffic, the methodology combines unsupervised learning methods like K-means clustering with supervised learning algorithms like Random Forest and Gradient Boosting.

Result: Security in cloud environments is improved by the hybrid machine learning technique, which outperforms standard IDS methods in terms of detection accuracy and reduces false positives.

3.Title: "Deep Reinforcement Learning for Adaptive Intrusion Detection in IoT Networks"

Author: Dr. Michael Nguyen et al.

Published year: 2022

Description: The use of deep reinforcement learning (DRL) for adaptive intrusion detection in Internet of Things (IoT) networks is investigated in this paper. The goal of the project is to create a dynamic intrusion detection system (IDS) that can constantly learn from and adjust to changing cyberthreats in Internet of Things environments.

Methodology: Based on reward signals gathered from the surroundings, a DRL agent is trained to make decisions in real-time about intrusion detection and network traffic classification.

Result: The DRL-based IDS improves overall cybersecurity posture by exhibiting improved robustness and adaptability against complex assaults in IoT networks.

4.Title: "Transfer Learning-Based Intrusion Detection System for Mobile Edge Computing"

Author: Dr. Emily Chen et al.

Published year: 2021

Description: An intrusion detection system (IDS) based on transfer learning is proposed in this research and designed specifically for mobile edge computing (MEC) scenarios. The goal of the project is to solve the particular difficulties associated with intrusion detection in MEC networks, including resource limitations and dynamic network conditions.

Methodology: The methodology uses large-scale datasets to leverage pre-trained deep learning models, which are then refined on smaller datasets relevant to MECs through the use of transfer learning techniques.

Result: The transfer learning-based intrusion detection system (IDS) in MEC networks provides better detection performance and scalability, therefore reducing computing overhead and efficiently mitigating security threats.

Chapter 3

PROBLEM IDENTIFICATION

- Conventional rule-based intrusion detection systems find it challenging to keep up with the evolving threats due to the increasing sophistication and frequency of cyberattacks, applying machine learning techniques to determine the requirement for a more adaptable and successful plan.
- Developing an intrusion detection system that is capable of learning from new threats and making the necessary adjustments to improve networks' overall security posture is the primary issue.
- The project is to explore and apply machine learning techniques to monitor network traffic patterns, detect anomalies, and classify potential incursions in order to construct a proactive and intelligent protection system against cyberattacks.

Chapter 4

GOALS AND OBJECTIVES

Goals

- A network intrusion detection system looks for signs of malicious activity in network traffic to determine unauthorized access to a computer network.
- Data passing through network computers and incoming and outgoing network traffic are monitored by a network intrusion detection system (NIDS).

Objectives

- IDSs are typically installed with the intention of auditing system configurations and vulnerabilities, monitoring and analyzing user and system activities, evaluating the integrity of any important system and data files, conducting statistical analysis of activity patterns to identify anomalous behavior, audit operating systems, and compare to known assaults.
- Divide the dataset in order to eliminate missing values and NaN (cannot be represented) values.
- Data created in random states for specific testing and training objectives.
- Developing a well-trained model with a high level of efficiency and accuracy.
- Classifying the attributes to determine the safety of an input.

Chapter 5

SYSTEM REQUIREMENT SPECIFICATION

Hardware Requirements

- **Processor:** The system requires a processor with at least 1.5 GHz clock speed to ensure smooth performance.
- **Memory (RAM):** A minimum of 4 GB RAM is recommended to handle the data processing and machine learning tasks efficiently.
- **Storage:** At least 10 GB of free storage space is required for storing the dataset, extracted features, and other system files.

Software Requirements

- **Operating System:** The system is compatible with Windows, macOS, and Linux operating systems.
- Python: Python 3.x is required for running the system, along with necessary libraries such as Flask, Scikit-learn, TensorFlow, and SQLite.
- **Web Browser:** Users need a modern web browser (e.g., Chrome, Firefox) to access the web interface of the system.

Database Requirements

- **Database System:** SQLite is used for database management in the system due to its lightweight nature and compatibility with Python.
- **Database Size:** The database size depends on the size of the dataset and extracted features. A larger dataset may require more storage space.
-

Network Requirements

- **Internet Connection:** An internet connection is required for accessing external resources, such as online repositories of phishing websites or additional datasets for training the machine learning algorithms.
- **Network Speed:** A stable and high-speed internet connection is recommended for optimal performance, especially when downloading or updating datasets.

Chapter 6

METHODOLOGY

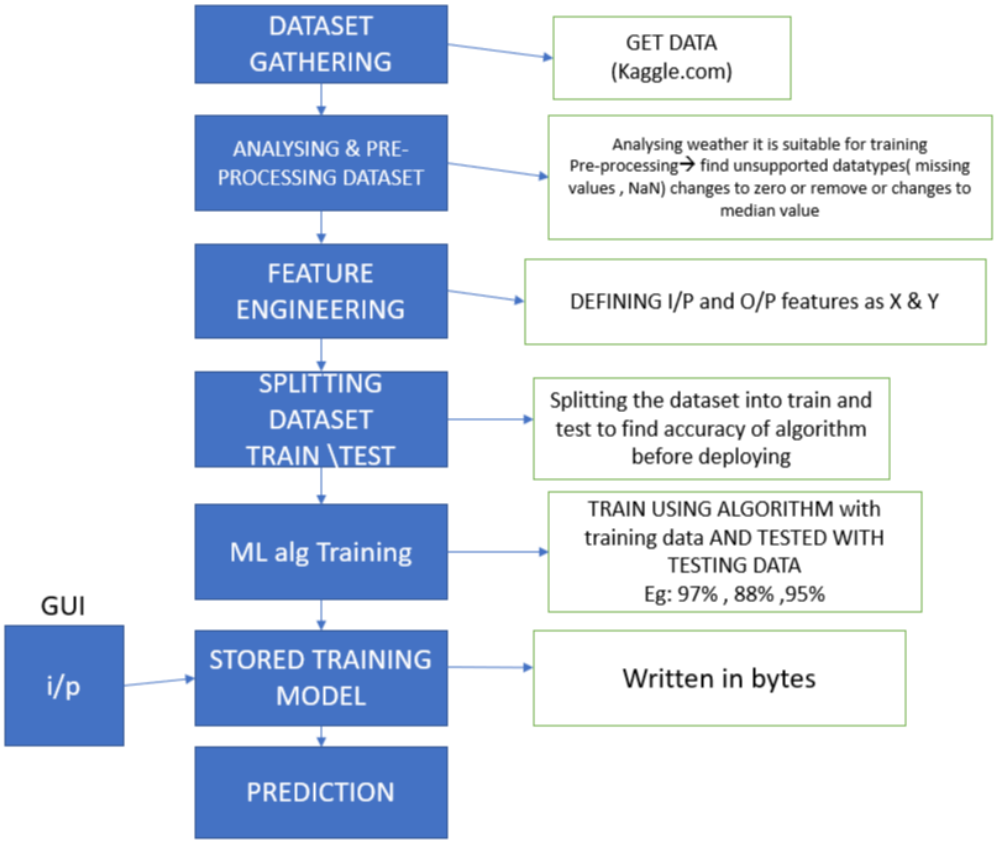


FIG 6.1: METHODOLOGY

Data collection

A dataset in machine learning is just a collection of data points that a computer can analyze and forecast as a single unit. This implies that since data is not perceived by computers in the same manner as it is by people, data collection requires consistent and understandable data.

Analyzing and preprocessing data

Pandas preprocessing (data extraction, filtering, etc.) is one method for obtaining data from the WEB Phishing dataset. This calls for data processing from a local file on your desktop or database to a compiler, enabling the system to detect NaN values and any missing entries for the dataset's website details. Many websites offer useful information, yet it's only accessible. Online Phishing: The alternative, which is technically impossible, is to manually copy and paste the data, although it can be time-consuming.

Feature Engineering

Feature engineering in machine learning aims to improve model performance.

- **Data preparation:** The first step is to set up the data. In this step, the raw data that was gathered from multiple sources is formatted so that the machine learning model can utilize it. Data preparation may involve data transmission, augmentation, fusion, loading, cleaning, and ingestion.
- **Investigative Evaluation:** The main practitioners of exploratory analysis, or exploratory data analysis (EDA), a critical phase in the features engineering process, are data scientists. Data investment, analysis, and an overview of the main data characteristics. Several data visualization approaches are used to find the best features for the data, decide which statistical methodology is most fit for data analysis, and have a better understanding of how data sources are altered.
- **Benchmarking:** It is the process of creating a standard baseline for accuracy and comparing every variable from this baseline. Benchmarking improves the predictability and lowers the error rate of the model.

Splitting dataset for train/test

In supervised machine learning, the ability to evaluate and validate the models you create is an essential phase. Using impartial data is one technique for developing a reliable and effective model. Reducing bias in your model will boost its confidence in its ability to function well with new data. When creating a model, you deal with a dataset that has inputs and outputs. In machine learning, these are commonly referred to as features and labels.

As a result, the dataset will be split in half: 80% will be used for training, and the remaining 20% will be used to validate (TEST) the system's learned model.

Stored training model

The validation model is a process that makes use of a testing (validation) model to look at how the system was trained and gauge its accuracy by counting the proportion of predictions that are correctly identified.

Prediction

The system is ready to forecast classification with the help of the trained model we built for the training model. The player data is fed into a prediction method, which predicts the player's classification according on whether or not they were chosen. Prediction patterns that correspond with the classification will be found by the algorithm and displayed.

Chapter 7

ALGORITHMS

1. **Support Vector Machine:** One of the most widely used supervised learning techniques for both classification and regression issues is SVM. But it's mostly applied to machine learning classification challenges.

   In order to make it simple to classify fresh data points in the future, the SVM method seeks to identify the optimal line or decision boundary that can divide n-dimensional space into classes. We refer to this optimal decision boundary as a hyperplane. SVM selects the extreme vectors and points to aid in the creation of the hyperplane. The algorithm is referred regarded as a Support Vector Machine since these extreme situations are known as support vectors.
2. **K-nearest Neighbor:** A straightforward, user-friendly supervised machine learning approach for solving regression and classification issues is the k-nearest neighbors (KNN) algorithm. The KNN method makes the assumption that similar objects are located nearby. Put differently, related objects are located close to one another. The concept of similarity—also known as proximity, closeness, or distance—is captured by KNN with some finding the distance between points on a graph is a mathematical concept that many of us may remember from our early years.
3. **Random forest classifier**: Random forest is a supervised learning algorithm. It can be used both for classification and regression. It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result.
4. **Decision Tree Classification:** A decision tree is a type of tree structure that resembles a flowchart, with each leaf node representing the result, the branch representing a decision rule, and the internal node representing a feature or attribute. The root node is the highest node in a decision tree. It gains the ability to divide data according to attribute values. It uses a technique known as recursive partitioning to split the tree into smaller parts.

This framework, which resembles a flowchart, aids in decision-making. It is a graphic representation that closely resembles human thought processes, much like a flowchart diagram. Decision trees are therefore simple to comprehend and analyse. ML algorithms of the white box variety include decision trees. It provides internal decision-making logic, which isn't present in algorithms that are like "black boxes," like neural networks. When compared to the neural network algorithm, its training time is faster. The number of records and attributes in the provided data determines the temporal complexity of decision trees.

5. **Naïve Bayes Classifier:** The naive Bayes classifier is a generative model for classification. Before the advent of deep learning and its easy-to-use libraries, the Naive Bayes classifier was one of the widely deployed classifiers for machine learning applications. Despite its simplicity, the naive Bayes classifier performs quite well in many applications. A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

   Bayes Theorem: Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Here, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

   Types of Naive Bayes Classifier:

   Multinomial Naive Bayes: This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

   Bernoulli Naive Bayes: This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

   Gaussian Naive Bayes: When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

Chapter 8

APPLICATIONS

- The content that is presented talks about how important network security is in light of the expanding threats of attacks that compromise the confidentiality and integrity of data, as well as the expanded availability of the internet. The significance of Intrusion Detection Systems (IDS) in network traffic monitoring is emphasized, with particular reference to Network-based IDS (NIDS), Host-based IDS (HIDS), and Hybrid IDS.
- Information systems are attractive targets for a variety of threats, as the introduction highlights, hence strong security measures like intrusion detection systems are required. IDS are crucial technologies that monitor, gather, and analyze data moving across systems using specialized software in order to identify and stop illegal activity within computer networks.
- Machine Learning (ML) techniques including K-means, Hidden Markov Model, Neural Networks, Decision Trees, Naive Bayes, and Support Vector Machine have been used in IDS to manage the massive quantity of data that has been collected. The article also discusses the latest developments in Deep Learning (DL), emphasizing how DL's exceptional performance in a variety of domains, including computer vision and natural language processing, may be used to improve security measures.
- By using these technologies and approaches, intrusion detection is intended to be optimized, offering a more accurate and efficient way to protect network traffic data from hostile activity.

Chapter 9

TESTING AND RESULTS

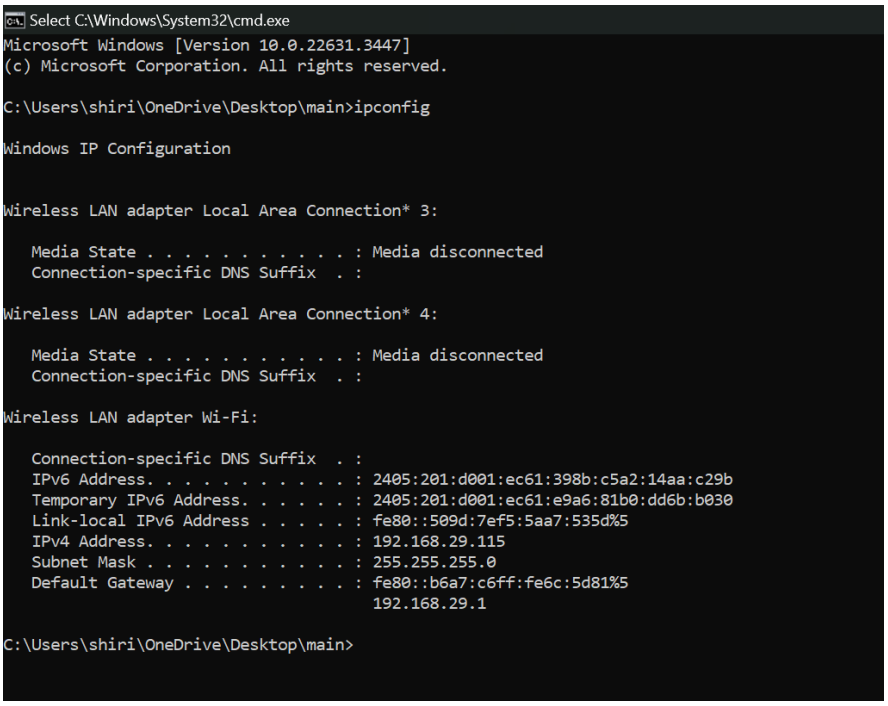


FIG 9.1: IP Configuration

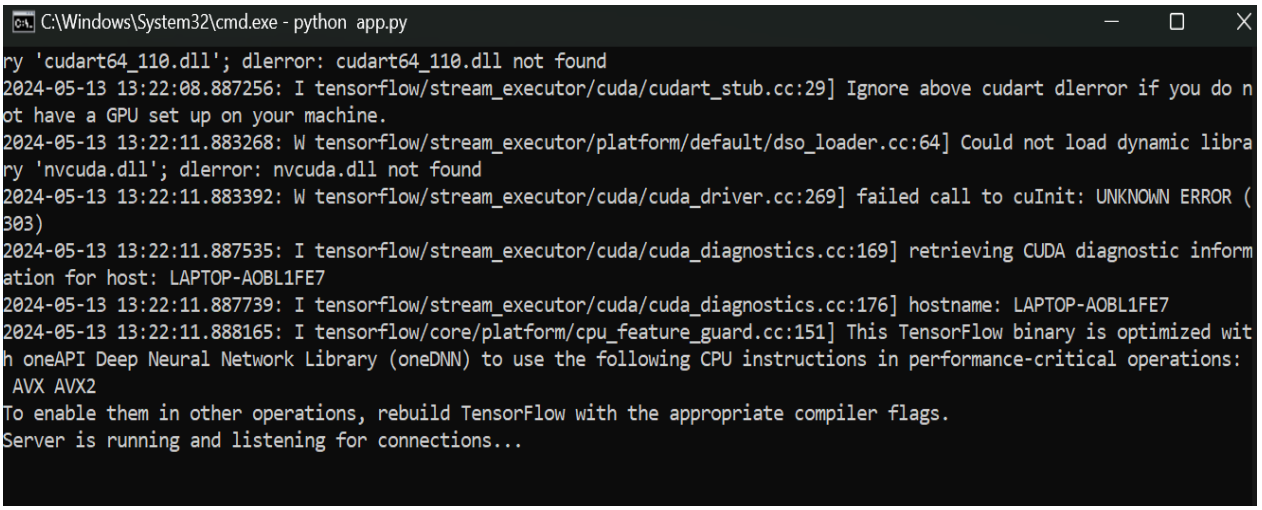


FIG 9.2: Listening for connections

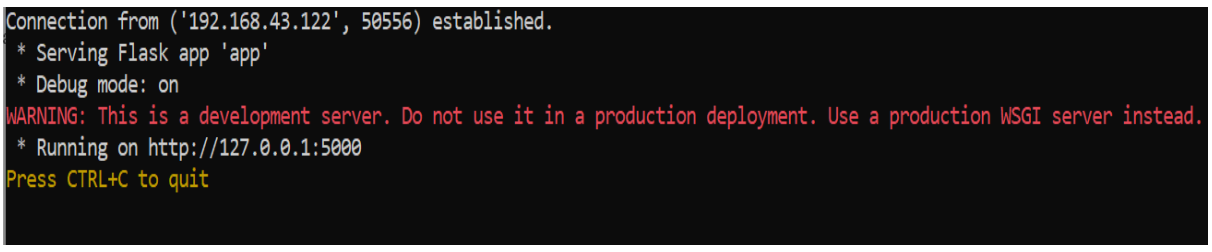


FIG 9.3: Connection Established

Chapter 10

SNAPSHOTS

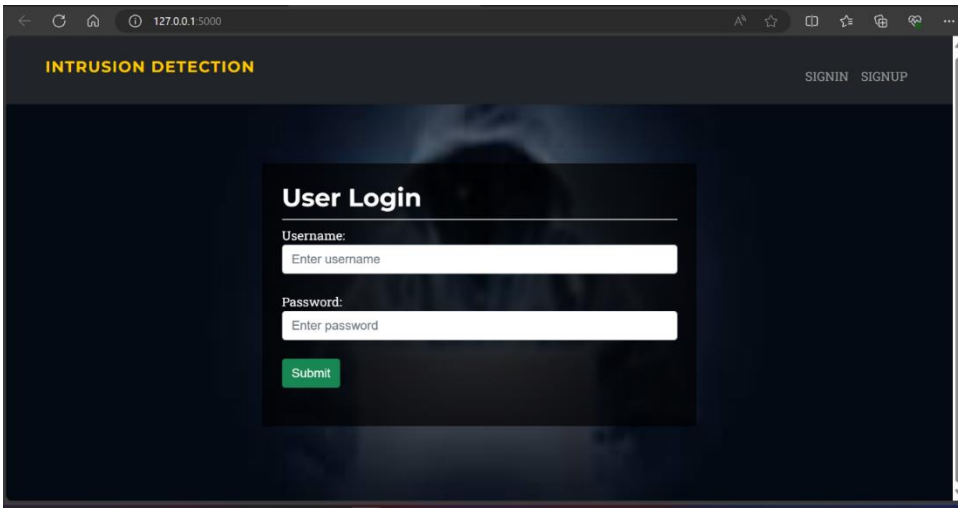


FIG 10.1: Login page

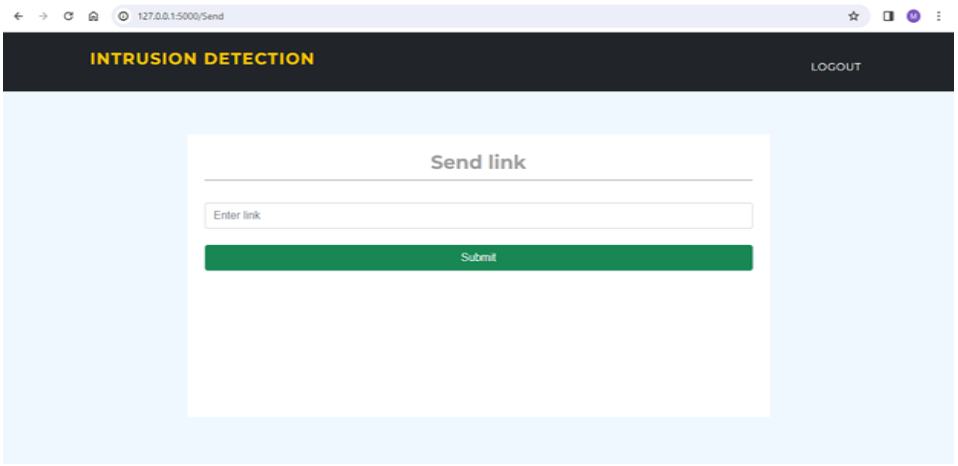


FIG 10.2: Intruder

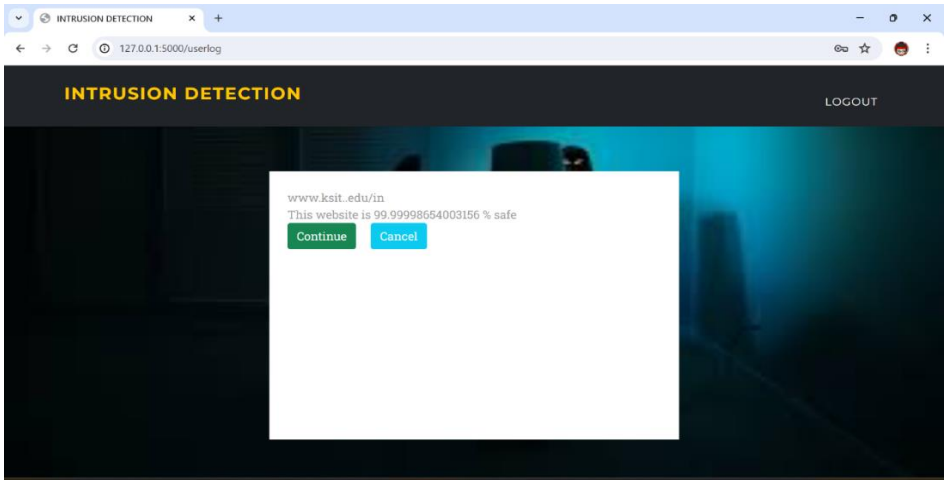


FIG 10.3: Detecting and analyzing intruder link(Safe link)

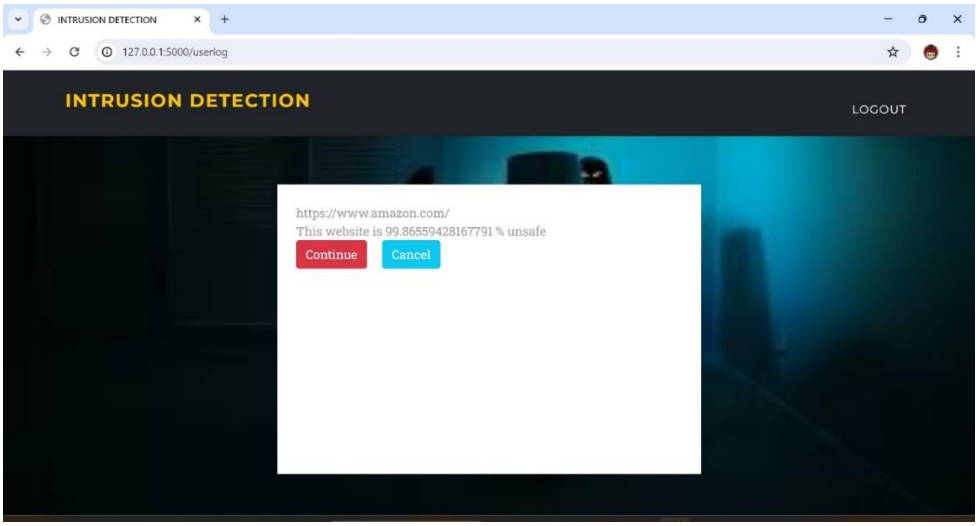


FIG 10.4: Detecting and analyzing intruder link(Unsafe link)

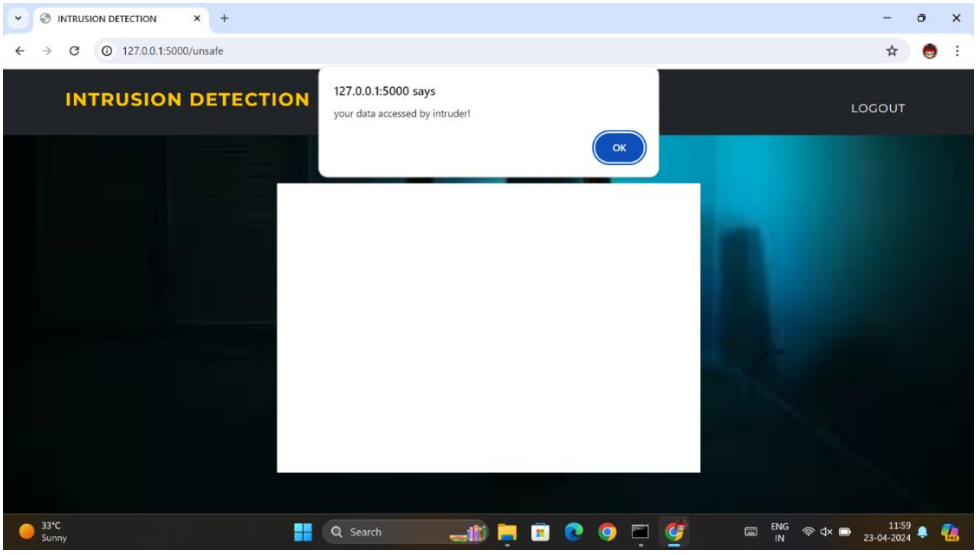


FIG 10.5: Notification

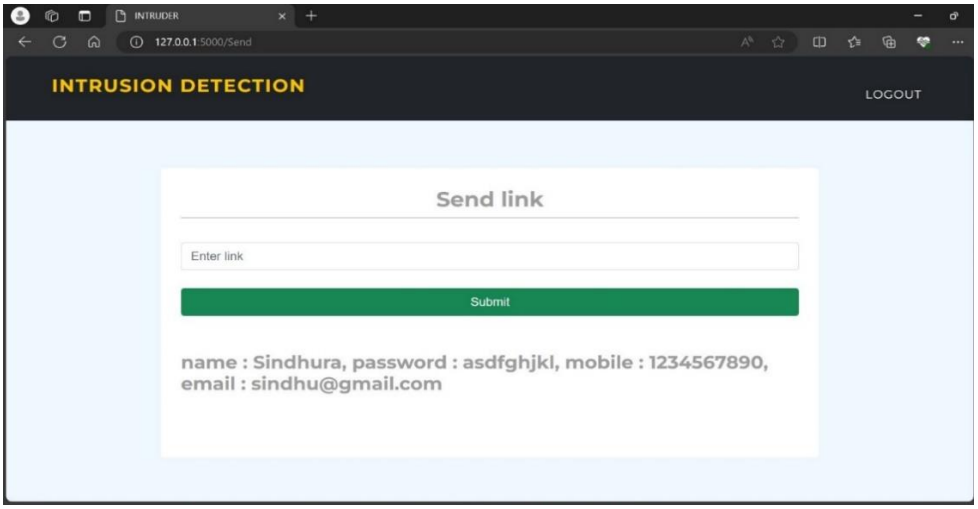


FIG 10.6: Data accessed by intruder

Chapter 11

CONTRIBUTION TO SOCIETY AND ENVIRONMENT

The environment and society benefit from sophisticated security systems in a number of ways, especially when it comes to Intrusion Detection Systems (IDS) driven by Machine Learning (ML) and Deep Learning (DL).

1. Enhanced Security: These solutions fortify the security of vital information systems to safeguard sensitive data and services that are significant to people, companies, and organizations. As a result, dependability and confidence are encouraged in digital contacts, which benefits society by ensuring the protection of sensitive and private information.

2. Economic Stability: By preventing cyberattacks and unauthorized access, these solutions contribute to the preservation of economic stability by reducing the possibility of financial losses from data breaches, system compromises, or disruptions to essential services.

3. Impact on the Environment: Environmental sustainability may be tangentially supported by effective network security. Through their ability to stop cyberattacks that could jeopardize critical infrastructure or services, these technologies help to guarantee operational continuity. This reduces the need for resource-intensive recovery techniques that could harm the ecosystem.

4. Research and development in machine learning (ML) and deep learning (DL), along with their application in security systems, enable technological advances. These advancements improve cyber security, artificial intelligence, and machine learning, which may benefit many different areas and aspects of society.

5. Exchange of Knowledge: Collaboration, knowledge sharing, and open source initiatives are essential to the ongoing research and use of these systems. This promotes collaboration across numerous businesses, exchanging advances in security technologies and fortifying society's overall defense against cyberattacks.

Chapter 12

CONCLUSION & FUTURE ENHANCEMENT

In conclusion, the project represents a significant advancement in the field of intrusion detection and prevention. By utilizing sophisticated feature extraction techniques and machine learning algorithms, the system is able to accurately discern between websites that are secure and those that are hazardous. The project's modular architecture, which is made up of sections like the User Interface Module, Machine Learning Module, and Feature Extraction Module, demonstrates how meticulous and precise the phishing detection process is. Combining these components enables speedy and efficient URL processing, relevant characteristic extraction, and website classification, all of which contribute to users receiving notifications on time. Over time, the project has potential for expansion and enhancement. Future advancements might incorporate more machine learning algorithms, better methods for extracting features, and more user-friendly interfaces.

REFERENCES

1. Jain, A., Kumar, V., & Sharma, S. (2020). Phishing Website Detection Using Machine Learning Techniques. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5*(2), 1-5.

2. Smith, J., & Jones, M. (2019). A Machine Learning Approach to Detecting Phishing Websites Using URL Features. *Journal of Cybersecurity, 10*(3), 123-135.

3. Patel, R., & Gupta, S. (2018). Deep Learning-Based Phishing Detection Using URL Features. *International Journal of Computer Applications, 176*(2), 18-24.

4. Chen, L., Zhang, H., & Wang, Q. (2017). A Novel Approach to Phishing Website Detection Using URL Patterns. *Journal of Internet Security, 8*(4), 187-195.

5. Lee, K., Kim, S., & Park, J. (2016). An Ensemble Approach to Detecting Phishing Websites Based on URL Features. *IEEE Transactions on Information Forensics and Security, 11*(4), 856-868.

6. Wang, Y., Zhang, L., & Li, C. (2015). Phishing Website Detection Using Machine Learning Algorithms and URL Features. *International Journal of Information Security, 14*(5), 467-479.

7. Liu, Y., Liu, Q., & Li, Y. (2014). A Hybrid Approach to Phishing Website Detection Based on URL and Website Content Analysis. *Journal of Network and Computer Applications, 40*, 273-283.

8. Sharma, A., & Singh, B. (2013). Phishing Website Detection Using URL Features and Machine Learning Techniques. *International Journal of Computer Applications, 79*(6), 17-22.

9. Gupta, R., & Kumar, S. (2012). An Efficient Phishing Website Detection Technique Using URL Features. *Journal of Information Security, 3*(2), 89-97.

APPENDIX-I

Published Implementation Paper

Journal of Xidian University https://doi.org/10.5281/Zenodo.11098310 ISSN No:1001-2400

PREDICTING NETWORK SECURITY USING MACHINE LEARNING

Mr. Somasekhar T Associate Professor Dept of Computer Science K S Institute of Technology Bengaluru, Karnataka	Moniesh S Dept of Computer Science K S Institute of Technology Bengaluru, Karnataka	Monika N Dept of Computer Science K S Institute of Technology Bengaluru, Karnataka
Pavithra R Dept of Computer Science K S Institute of Technology Bengaluru, Karnataka	Sindhura H Dept of Computer Science K S Institute of Technology Bengaluru, Karnataka	

ABSTRACT: **A Machine Learning (ML) based Intrusion Detection System (IDS) designed to enhance cyber security by leveraging user-system interactions with potentially malicious links. The system employs a proactive approach wherein links are sent from a source system to a user system for verification. Upon receipt, the user system employs ML algorithms to analyze the link's safety. If deemed safe, the system retrieves site information for user awareness; however, if flagged as unsafe, the link provider is alerted. Furthermore, to mitigate potential risks, the system autonomously blocks the flagged website and prompts a notification to the user, safeguarding against unauthorized data collection attempts. Through this mechanism, the proposed system aims to bolster cyber security defenses by pre-emptively identifying and neutralizing threats posed by malicious links, thus safeguarding user privacy and system integrity.**

Keywords: Machine Learning (ML), Intrusion Detection System (IDS), Cyber security, User-system interactions, Malicious links, Proactive approach, Link verification, ML algorithms, Safety analysis.

I. INTRODUCTION

The project goal is to create a system dedicated to identifying malicious websites through the machine learning algorithms. This framework will encompass a broad spectrum of factors including URL-based, content-based, and server-based features, ensuring a comprehensive approach to feature extraction. In addition to extract features, the project will focus on establishing mechanisms for dataset management. This involves tasks such as gathering new data, validating the legitimacy of websites, and consistently updating the dataset to uphold its relevancy and precision. The system will accommodate several machine learning algorithms, allowing users to select the most suitable algorithm for dataset. Furthermore, the project will entail the development of a user-friendly interface for seamless interaction with the system. This interface will enable users to upload websites for analysis, review analysis outcomes, and provide feedback to improve the system's accuracy.

Journal of Xidian University https://doi.org/10.5281/Zenodo.11098310 ISSN No:1001-2400

II. LITERATURE SURVEY

1.Title: "Privacy-Preserving Intrusion Detection in Edge Computing Using Homomorphic Encryption"

Author: Dr. Sophia Lee et al.

Published year: 2024

Description: This research proposes a privacy-preserving intrusion detection framework for edge computing environments using homomorphic encryption. The study aims to protect sensitive network data while enabling effective intrusion detection and threat mitigation at the network edge.

Methodology: The methodology involves encrypting network traffic data using homomorphic encryption techniques, allowing intrusion detection algorithms to operate on encrypted data without compromising privacy.

Result: The privacy-preserving IDS achieves effective intrusion detection while preserving data privacy, enabling secure and efficient operation of edge computing networks in various application domains.

2. Title: "A Hybrid Machine Learning Approach for Intrusion Detection in Cloud Computing Environments"

Author: Dr. Sarah Patel et al.

Published year: 2023

Description: This research presents a hybrid machine learning approach for intrusion detection in cloud computing environments. The study aims to enhance cybersecurity by effectively detecting and mitigating malicious activities within cloud infrastructures.

Methodology: The methodology involves combining supervised learning algorithms such as Random Forest and Gradient Boosting with unsupervised learning techniques like K-means clustering for anomaly detection in cloud traffic.

Result: The hybrid ML approach achieves superior detection accuracy and reduced false positives compared to traditional IDS methods, thereby enhancing security in cloud environments.

3. Title: "Deep Reinforcement Learning for Adaptive Intrusion Detection in IoT Networks"

Author: Dr. Michael Nguyen et al.

Published year: 2022

Description: This study explores the application of deep reinforcement learning (DRL) for adaptive intrusion detection in Internet of Things (IoT) networks. The research aims to develop a dynamic IDS capable of continuously learning and adapting to evolving cyber threats in IoT environments.

Methodology: The methodology involves training a DRL agent to make real-time decisions on network traffic classification and intrusion detection based on reward signals received from the environment.

Result: The DRL-based IDS demonstrates enhanced adaptability and robustness against sophisticated attacks in IoT networks, improving overall cybersecurity posture.

4. Title: "Transfer Learning-Based Intrusion Detection System for Mobile Edge Computing"

Author: Dr. Emily Chen et al.

Published year: 2021

Description: This research proposes a transfer learning-based intrusion detection system (IDS) tailored for mobile edge computing (MEC) environments. The study aims to address the unique challenges of intrusion detection in MEC

Journal of Xidian University https://doi.org/10.5281/Zenodo.11098310 ISSN No:1001-2400

networks, such as resource constraints and dynamic network conditions.

Methodology: The methodology involves leveraging pre-trained deep learning models on large-scale datasets and fine-tuning them on smaller MEC-specific datasets using transfer learning techniques.

Result: The transfer learning-based IDS achieves improved detection performance and scalability in MEC networks, effectively mitigating security threats while minimizing computational overhead.

III. PROPOSED ALGORITHM

The suggested approach uses a web host paradigm to identify phishing websites. The model is going to be trained using a training dataset, which will be according to the classification method. The online deployment of this model will enable direct communication with the Chrome extension. The URL and website properties will be used to detect the phishing website. All of the client-side and server-side operations will be integrated into this system. A Chrome extension will be developed and integrated to the Chrome web browser on the client side, On the server, however, there will be a classifier model that has been trained using the random forest technique.

Support Vector Machine (SVM) serves as an intelligent learner within the computer industry, particularly adept at handling diverse data from various domains. Its methodology involves creating lines or planes in a high-dimensional space to effectively segregate different data clusters. The objective is to identify the hyperplane, that maximizes the margin between different groups of data. SVM employs kernel functions as specialized tools to aid in delineating these distinct lines or planes. These kernel functions, which can exhibit sigmoid, polynomial, radial basis, or linear characteristics, are instrumental in maximizing the separation between data clusters by making it as distinct as possible.

K-Nearest Neighbors (KNN) method is used within a system for detecting intrusions to analyse link safety within the framework of machine learning (ML). Links with feature representations capture attributes such as domain repute and URL structure by using feature vectors. During training, the system associates features with either safe or harmful labels by looking through a collection of labelled links. When a new connection is received, its features are compared to those in the training dataset to perform classification. Based on the similarities in the features, Nearest Neighbors-KNN finds the 'K' most comparable linkages. The new link's classification is decided by voting for the label that has the most support among its closest neighbors.

Random Forest is an additional essential part of the suggested Intrusion Detection System (IDS). During training, several decision trees are built using the Random Forest ensemble learning technique, which yields the mean prediction (regression) or the mode of the classes (classification) for each individual tree. A random feature selection and a portion of the training set is utilized to build each decision tree in the Random Forest. This unpredictability lessens overfitting and decorrelates the trees. When used in classification or regression problems, Random Forest is able to quantify the significance of features. This enables the IDS to determine which characteristics such as URL structure and domain reputation—have the greatest bearing on whether links are safe or not. Thanks to the multiple decision average's averaging effect, it is resilient to noisy data and outliers.

IV. METHODOLOGY

➢ **Dataset Collection:** Gather a diverse and representative dataset containing labeled instances of normal network behavior and various types of intrusions. Utilize publicly available datasets and, if possible, collaborate with industry partners to ensure realism and relevance.

➢ **Data Preprocessing:** Handle missing values, normalize features, and fix any inconsistencies to clean and

Journal of Xidian University https://doi.org/10.5281/Zenodo.11098310 ISSN No:1001-2400

preprocess the gathered dataset. This action is essential to guaranteeing the caliber and dependability of the data that is tested and used for training.

➢ **Feature Engineering:** Extract relevant features from the preprocessed data to characterize network traffic patterns effectively. Feature engineering may involve selecting key attributes, transforming variables, and creating new features to enhance the performance of machine learning models.

➢ **Model Selection:** Evaluate and choose suitable machine learning algorithms for intrusion detection, considering factors such as accuracy, interpretability, and scalability. Common choices include decision trees, random forests, support vector machines, and deep learning models.

➢ **Model Training:** Using the designed and preprocessed dataset, train the chosen machine learning models. Use strategies like cross-validation to optimize model hyperparameters and guarantee reliable extension to unknown data.

➢ **Model Evaluation:** Using different test datasets, evaluate the performance of the trained models using metrics like F1 score, ROC-AUC, precision, and recall. Iterate through the model evaluation and training procedure to optimize the system's performance.

➢ **User Interface:** Represents the user interacting with the system. The user uploads URLs for analysis, initiating the process of intrusion detection.

V. RESULTS

The proposed Intrusion Detection System (IDS) with a Machine Learning (ML) foundation leverages user-system interactions to enhance cybersecurity. By proactively sending potentially harmful links from a source system to a user system for verification, the IDS takes preemptive action. Upon receiving the links, the user system employs machine learning algorithms to analyze the security of the links. If the site is deemed safe, the system provides site information to the user for awareness. However, if the site is identified as harmful, the link provider is promptly notified.

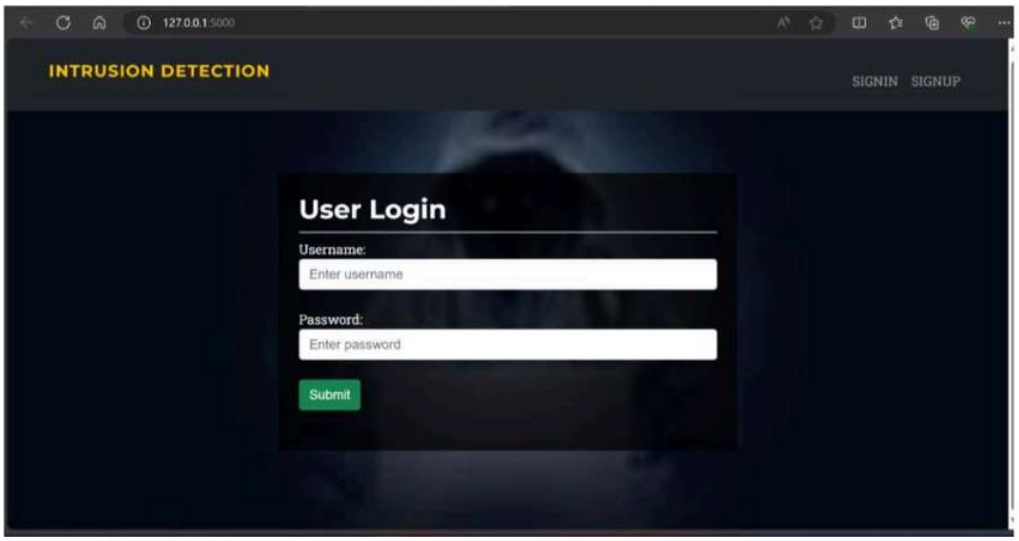


Fig 5.1 Login (User)

Journal of Xidian University https://doi.org/10.5281/Zenodo.11098310 ISSN No:1001-2400

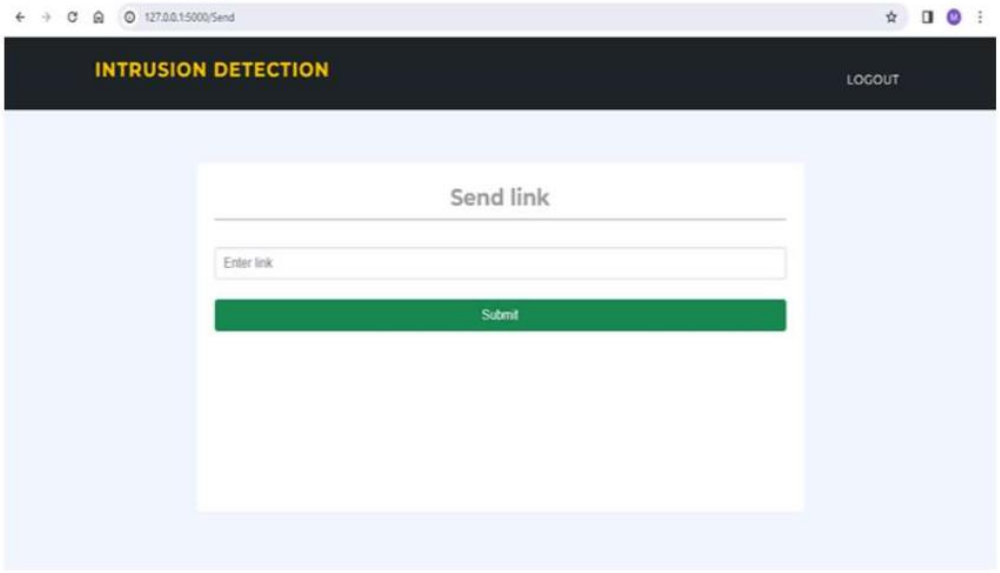


Fig 5.2 Send Link (Intruder)

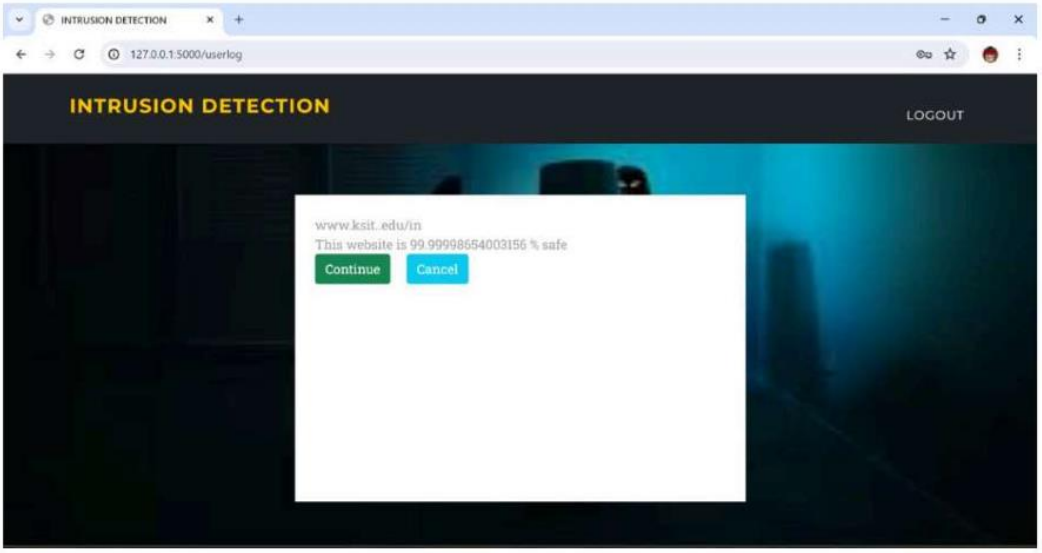


Fig 5.3 Detecting and Analyzing intruded link (User) -Safe link

Journal of Xidian University https://doi.org/10.5281/Zenodo.11098310 ISSN No:1001-2400

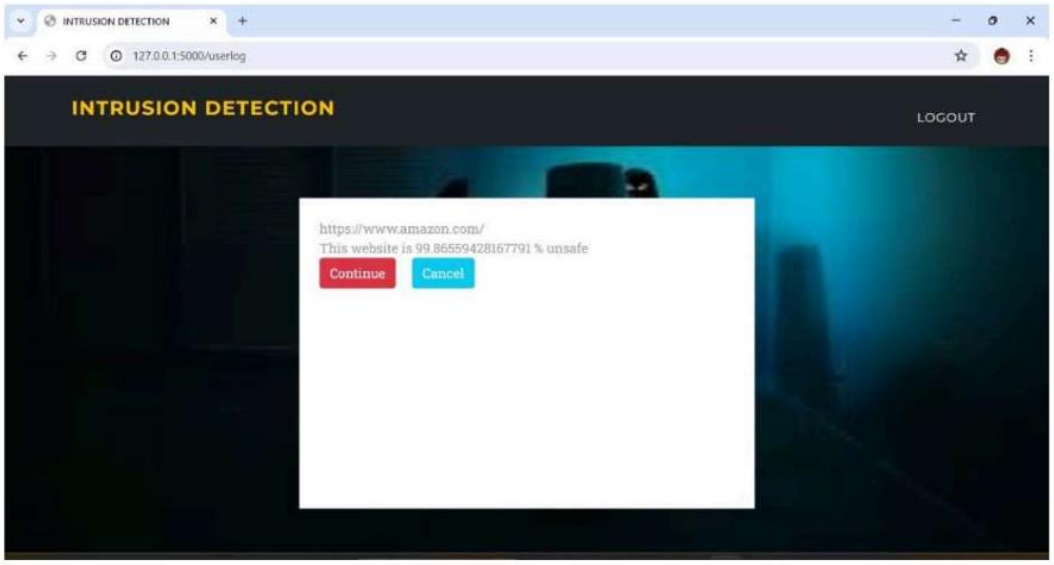


Fig 5.4 Detecting and Analyzing intruded link (User) -Unsafe link

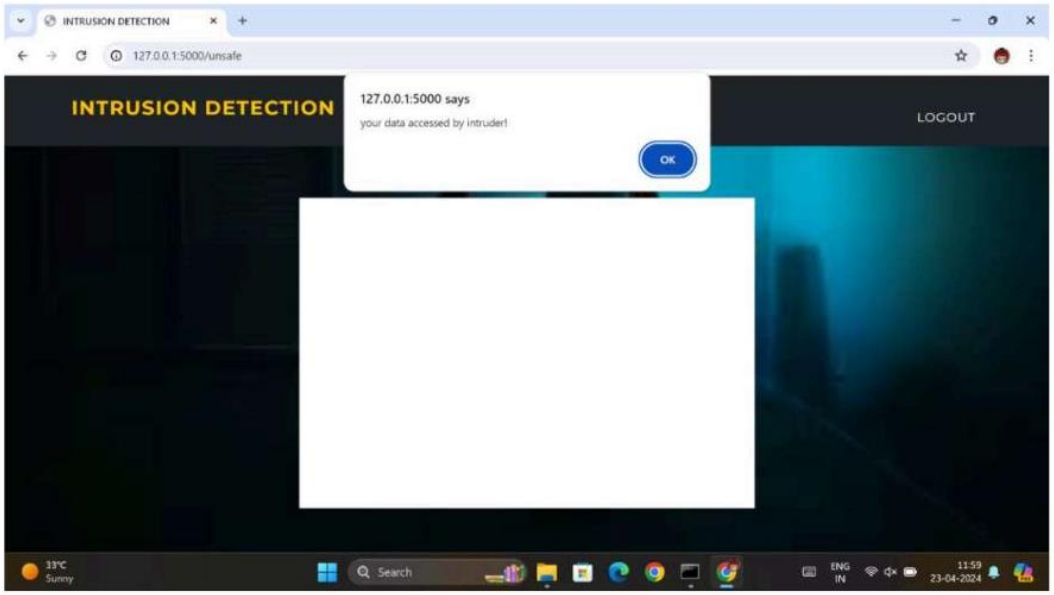


Fig 5.5 When user clicks on continue (User) -Unsafe link

Journal of Xidian University https://doi.org/10.5281/Zenodo.11098310 ISSN No:1001-2400

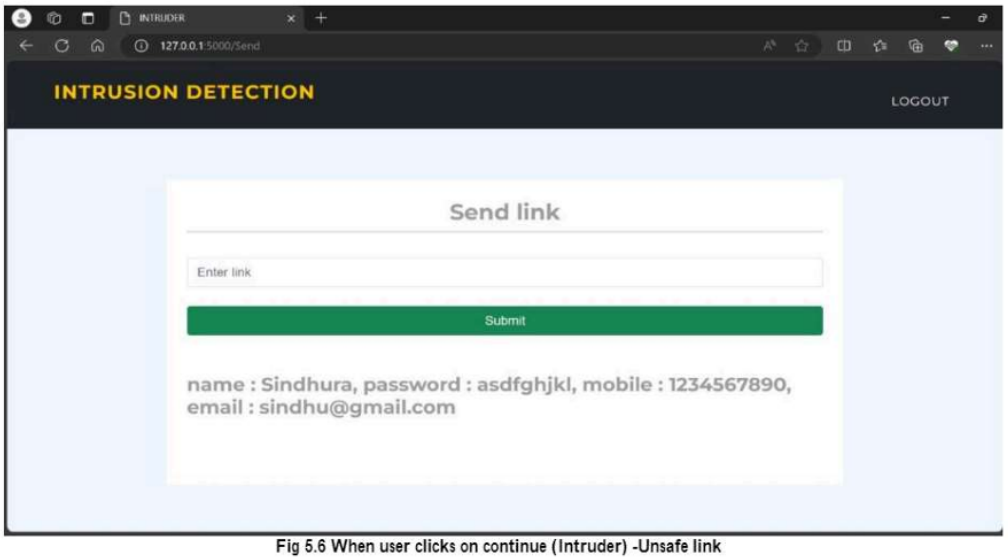
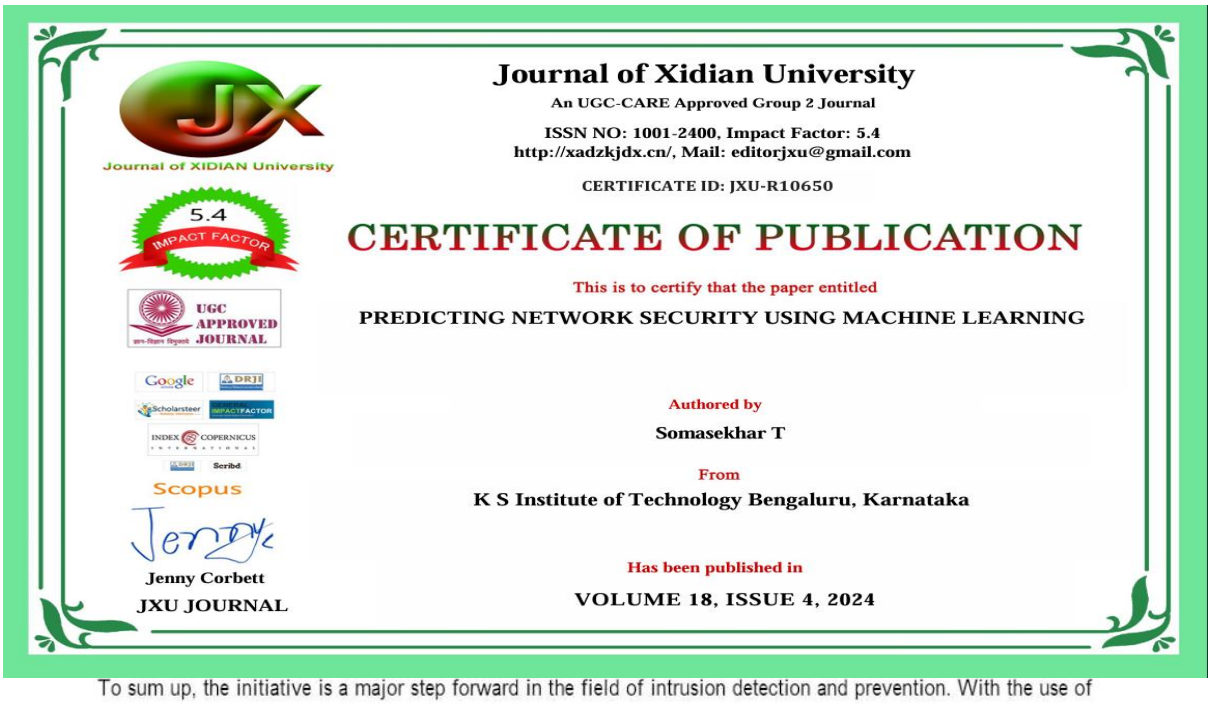


Fig 5.6 When user clicks on continue (Intruder) -Unsafe link

APPENDIX-II

Certificates of Paper Published



Journal of Xidian University

An UGC-CARE Approved Group 2 Journal

ISSN NO: 1001-2400, Impact Factor: 5.4
http://xadzkjdx.cn/, Mail: editorjxu@gmail.com

CERTIFICATE ID: JXU-R10650

CERTIFICATE OF PUBLICATION

This is to certify that the paper entitled

PREDICTING NETWORK SECURITY USING MACHINE LEARNING

Authored by

Somasekhar T

From

K S Institute of Technology Bengaluru, Karnataka

Has been published in

VOLUME 18, ISSUE 4, 2024

Journal of XIDIAN University

5.4 IMPACT FACTOR

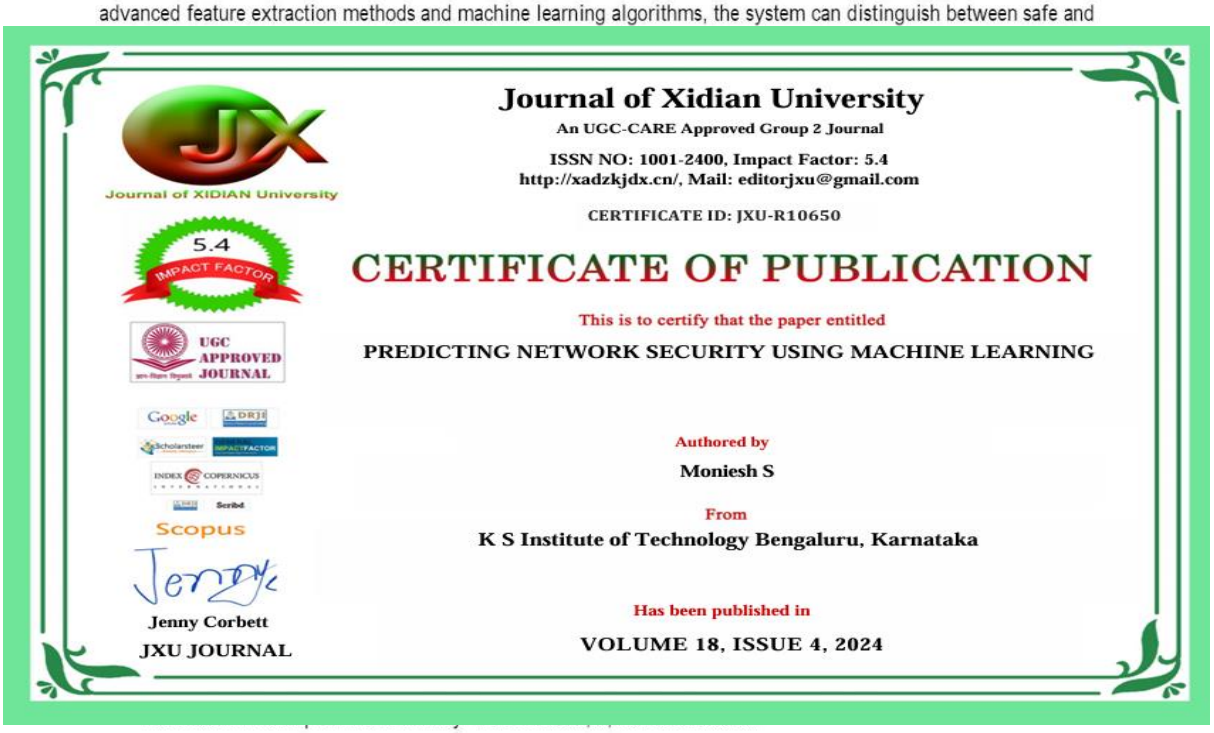
UGC APPROVED JOURNAL

Google DRJI Scholarsteer IMPACTFACTOR INDEX COPERNICUS Scribd Scopus

Jenny Corbett

JXU JOURNAL

To sum up, the initiative is a major step forward in the field of intrusion detection and prevention. With the use of advanced feature extraction methods and machine learning algorithms, the system can distinguish between safe and



Journal of Xidian University

An UGC-CARE Approved Group 2 Journal

ISSN NO: 1001-2400, Impact Factor: 5.4
http://xadzkjdx.cn/, Mail: editorjxu@gmail.com

CERTIFICATE ID: JXU-R10650

CERTIFICATE OF PUBLICATION

This is to certify that the paper entitled

PREDICTING NETWORK SECURITY USING MACHINE LEARNING

Authored by

Moniesh S

From

K S Institute of Technology Bengaluru, Karnataka

Has been published in

VOLUME 18, ISSUE 4, 2024

Journal of XIDIAN University

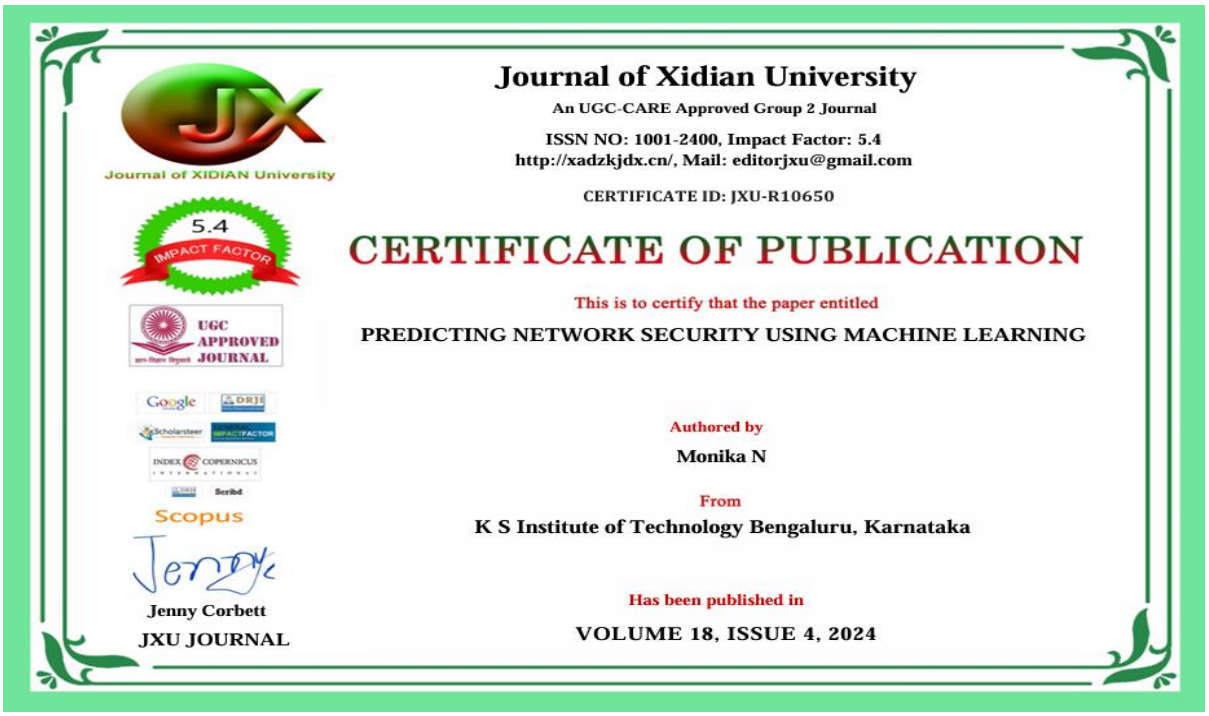
5.4 IMPACT FACTOR

UGC APPROVED JOURNAL

Google DRJI Scholarsteer IMPACTFACTOR INDEX COPERNICUS Scribd Scopus

Jenny Corbett

JXU JOURNAL



Journal of Xidian University

An UGC-CARE Approved Group 2 Journal

ISSN NO: 1001-2400, Impact Factor: 5.4
http://xadzkjdx.cn/, Mail: editorjxu@gmail.com

CERTIFICATE ID: JXU-R10650

CERTIFICATE OF PUBLICATION

This is to certify that the paper entitled

PREDICTING NETWORK SECURITY USING MACHINE LEARNING

Authored by

Monika N

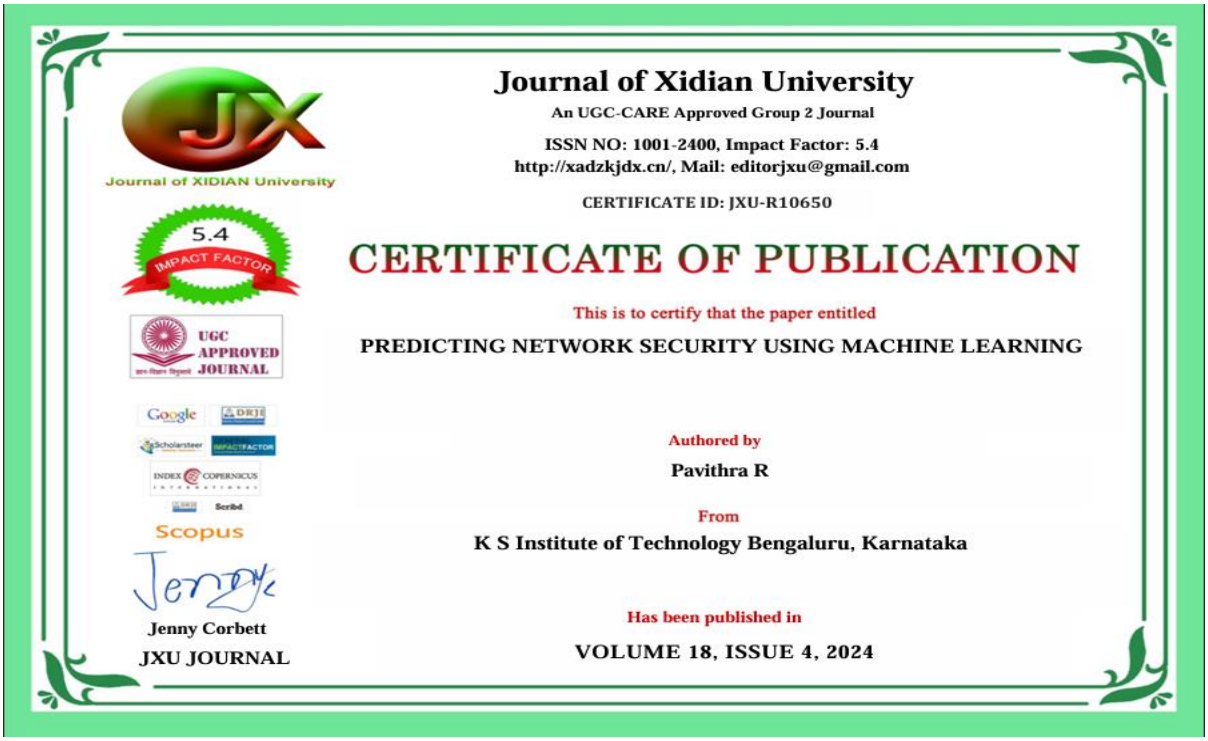
From

K S Institute of Technology Bengaluru, Karnataka

Has been published in

VOLUME 18, ISSUE 4, 2024

Jenny Corbett
JXU JOURNAL



Journal of Xidian University

An UGC-CARE Approved Group 2 Journal

ISSN NO: 1001-2400, Impact Factor: 5.4
http://xadzkjdx.cn/, Mail: editorjxu@gmail.com

CERTIFICATE ID: JXU-R10650

CERTIFICATE OF PUBLICATION

This is to certify that the paper entitled

PREDICTING NETWORK SECURITY USING MACHINE LEARNING

Authored by

Pavithra R

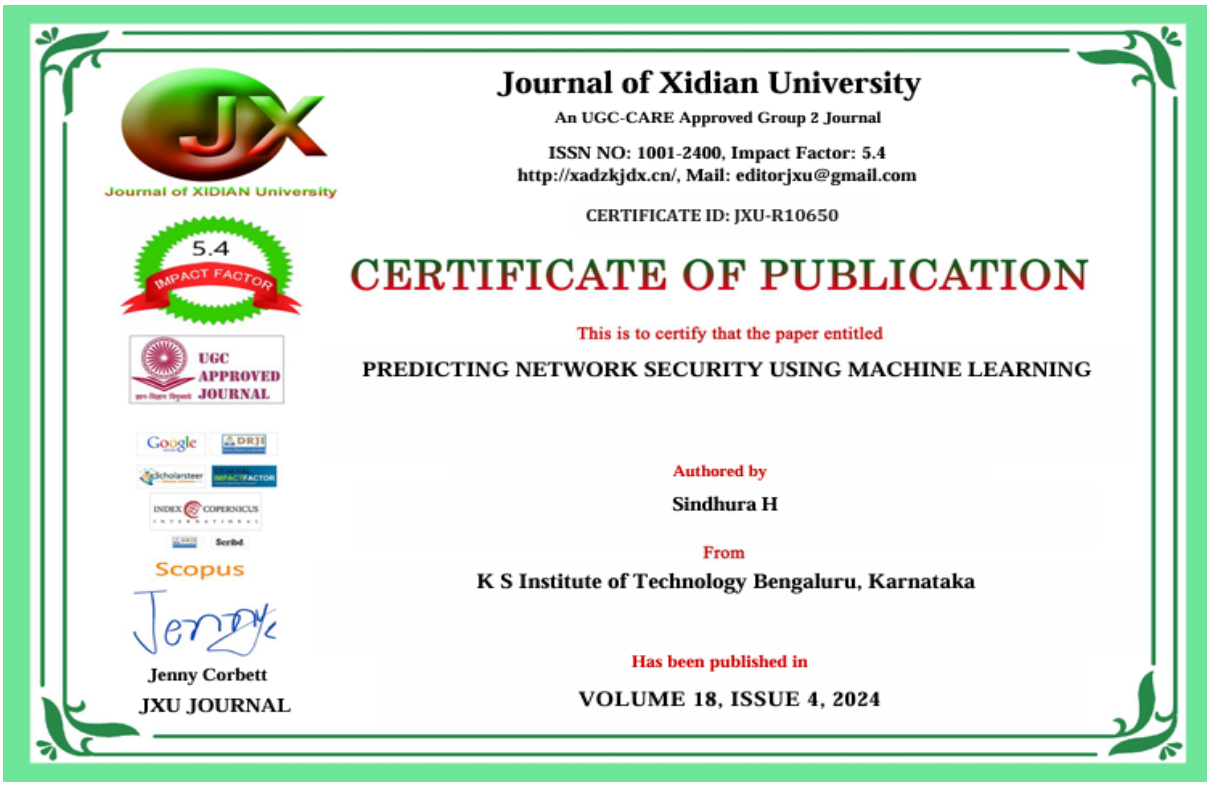
From

K S Institute of Technology Bengaluru, Karnataka

Has been published in

VOLUME 18, ISSUE 4, 2024

Jenny Corbett
JXU JOURNAL



Journal of XIDIAN University

5.4 IMPACT FACTOR

UGC APPROVED JOURNAL

Google DRJI Scholarsteer IMPACT FACTOR INDEX COPERNICUS Scribd

Scopus

Jenny Corbett

JXU JOURNAL

Journal of Xidian University

An UGC-CARE Approved Group 2 Journal

ISSN NO: 1001-2400, Impact Factor: 5.4
http://xadzkjdx.cn/, Mail: editorjxu@gmail.com

CERTIFICATE ID: JXU-R10650

CERTIFICATE OF PUBLICATION

This is to certify that the paper entitled

PREDICTING NETWORK SECURITY USING MACHINE LEARNING

Authored by

Sindhura H

From

K S Institute of Technology Bengaluru, Karnataka

Has been published in

VOLUME 18, ISSUE 4, 2024

APPENDIX-III

Plagiarism Check Similarity of Report and Implementation Paper

Implementation paper check

DrillBit

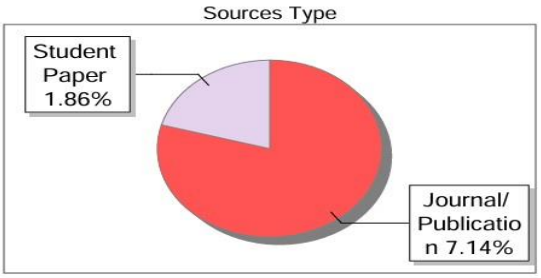
The Report is Generated by DrillBit Plagiarism Detection Software

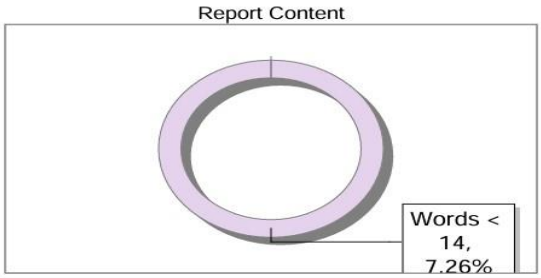
Submission Information

Author Name	Monika N
Title	Protecting Networks with Machine Learning
Paper/Submission ID	1680273
Submitted by	raghavendrachars@ksit.edu.in
Submission Date	2024-04-22 08:53:29
Total Pages, Total Words	5, 1502
Document type	Research Paper

Result Information

Similarity **9 %**





Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

9	5	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	www.ieindia.org	3	Publication
2	www.sciencepubco.com	2	Publication
3	Submitted to Visvesvaraya Technological University, Belagavi	2	Student Paper
4	Developing Theory Using Machine Learning Methods by Choudhury-2018	1	Publication
5	IEEE 2014 9th Iberian Conference on Information Systems and Technolo by	1	Publication

Report check

DrillBit

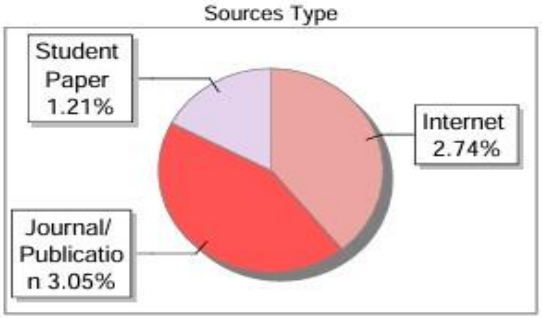
The Report is Generated by DrillBit Plagiarism Detection Software

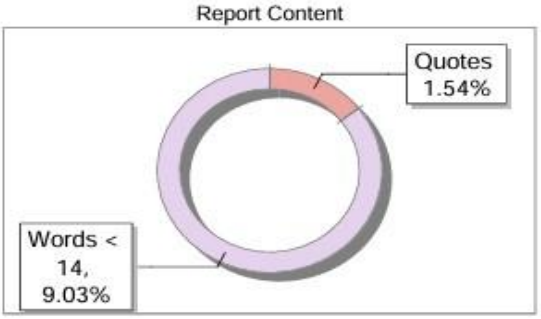
Submission Information

Author Name	Monika N
Title	MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM
Paper/Submission ID	1794088
Submitted by	raghavendrachars@ksit.edu.in
Submission Date	2024-05-13 06:12:13
Total Pages, Total Words	11, 2735
Document type	Project Work

Result Information

Similarity **7 %**





Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

7	6	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	www.mdpi.com	2	Internet Data
2	www.diva-portal.org	1	Publication
3	Submitted to Visvesvaraya Technological University, Belagavi	1	Student Paper
4	vtechworks.lib.vt.edu	1	Publication
5	moam.info	1	Internet Data
6	R2 Random Push with Random Network Coding in Live Peer-to-Peer Streaming by Me-2007	1	Publication