*Section A -Research paper*

# PHISHING LINK DETECTION USING MACHINE LEARNING

**[1] Divyansh Sharma, [2]Vishu Sharma, [3]Yashashvi Dixit, [4]Sonal Pahwa, [5]Amit Kumar Saini**

[1]divyansh.sharma.cs.2019@miet.ac.in, [2]vishu.sharma.cs.2019@miet.ac.in
[3]yashashvi.dixit.cs.2019@miet.ac.in, sonal.pahwa.cs.2019@miet.ac.in, amit.cs@miet.ac.in

Meerut Institute of Engineering and Technology, Meerut, India
**DOI:10.48047/ecb/2023.12.si4.739**

**Abstract – In this paper safety and security for the user is provided by preventing them from accessing any harmful or suspicious link that may possibly be a phishing link while browsing the internet. As most of the financial and work-related activities have been moved to the internet, that makes us more exposed to cybercrime. Phishing attacks are one of the common threats to any internet user out there. In this the attacker exploits human vulnerability by tricking a person into revealing sensitive information through a URL or link that looks secured. A phisher can target both individuals and organizations, the main aim of phishers is to acquire a user's sensitive and critical information such as banking details, username and password, etc. To overcome this problem, we are using a Random Forest algorithm for detecting any phishing URL link based on the features of the URL, the extension will report and alert the user about the URL being a phishing link and not to be processed further. After testing various algorithms Random Forest Algorithm is applied to our dataset and create a user-friendly chrome based plugin also called as chrome extension.**

**Keywords—Chrome Extension, Phishing Detection, Random Forest Algorithm, Web security.**

## 1. Introduction

In today's scenario, when everyone is working from home and have moved most of their work online it's become necessary to protect the users credentials and sensitive information from getting exploited by the hackers. Phishing is a common and prevalent cyber attack that a phisher or hacker does by stealing sensitive information through various manipulation techniques, it has interfered with people's lives and had an impact on a number of organizations. Phishing can be done in a number of ways, including via email, clickbait advertisements, whaling, spear-phishing, and more. An official website is copied by attackers into a website that appears authentic to the user but is actually run by the attacker. The users eventually give their login

information and other sensitive information to the cloned website, which is subsequently exploited by the attacker [1][2].

Phishing is not just restricted to emails and pop-ups on websites. Links in tweets, Facebook posts, status updates, and advertisements online might take you to malicious websites built to steal your financial data. Several strategies have been used to prevent phishing, however they haven't really proven that successful. Campaigns on cybersecurity precautions to be taken against phishing have been organized by a wide variety of organizations and businesses. But this conventional method hasn't been very effective, which has prompted the development of numerous phishing detection software and plugins using a variety of technologies year after year [3][4].

Example of a typical attack:

1. An attacker impersonating a member of the IT department sends a bulk email to employees.

2. The email or a notification that serves as a reminder for an employee to complete an important task or course, but the course is under the control of an attacker.

3. The affected user is instructed to input their employee credentials during the course completion, which are then sent directly to the attacker [5][6].

## 2. Related Work

The various methods and technologies used to create systems and software to prevent or detect have been discussed in numerous publications we have read so far [7].

Network Phishing Detection database mining is used in [2] Zou Futai, Gang Yuxiang, Pei Bei, Pan Li, and Li Linsen's "Web Phishing Detection Based on Graph Mining," published in IEEE in 2016. This will identify any potential phishing attempts that the analysis of the URLs cannot. utilizes a user-website browsing setup. to obtain the information gathered from actual traffic of a Big ISP. The client consumer receives a unique AD, but the ISP chooses a generic address from its own Address database. With the AD, we consequently establish a visit relationship graph.

They proposed a framework that examines the idea of ACT-R cognitive-behavioral architecture in [3] Nick Williams and Shujun Li's article "Simulating Human detection of phishing websites: An investigation into the applicability of ACT-R cognitive behavior architecture model," published in IEEE in 2017. Simulate the brain processes that are used to determine the legitimacy of a certain website using features mostly found in HTTPS padlock secure predictors. Further research to more accurately reflect the range of human security awareness and activities in an ACT-R system may lead to deeper insights into how to integrate technology and human protection to reduce the likelihood of phishing attacks for users. ACT-R has good abilities that model the phishing use case well.

## 3. Methodology Used

The dataset used has been recommended from the UCI machine learning repository for our suggested system. The phishing website attribute collection consists of around 11,000 URLs (examples), which include 6,000 phishing cases and 5,000 real incidents. Each of these instances has about twenty-five aspects, each of which is linked to a set of choices and proceeds in accordance with a set of rules [8][9].

Figure 3.1 below illustrates how the qualities and features are divided into various groups:

I Address-bar-based features

ii) Anomaly-based characteristics

iii) JavaScript and HTML-based features

iv) Domain system-based features

Fig.3.1 - Attributes in Dataset

## 4. IMPLEMENTATION DETAILS

The dataset was arranged in a 7:3 ratio for planning and verification [10][11].

In the field of tests, the results of our research will be presented.

### A. Classifier based on Random Forest (Algorithm)

Based on the values of the random variable collected independently, random forests are optimization methods that include multiple tree predictors [12][13].

The distribution of all forest trees actually follows the same pattern. For the majority of the data, the random forest algorithm can correctly identify a class. However, there are a few errors that

occasionally even trees make. As a result, we decide to conduct a vote for each observation in order to observe the class on the poll result and help the model forecast outcomes more accurately [14][15].

This algorithm is used in our suggested method to determine whether or not web URLs are phishing attempts. In order to predict the class from the data, we utilise this approach to classify the characteristics on the data. There are numerous situations, each with a range of possible consequences. On this foundation, many decision trees are developed based on various features and outcomes as well as the same features with various outcomes. The class prediction is based on the final outcome, which is based on the majority pole of all the multiple decision trees. Additionally, while building the forest, an internal, impartial calculation of the generalization error is generated. Actually, incomplete data can be calculated effectively. Below is the figure of our system's flow diagram [16][17].

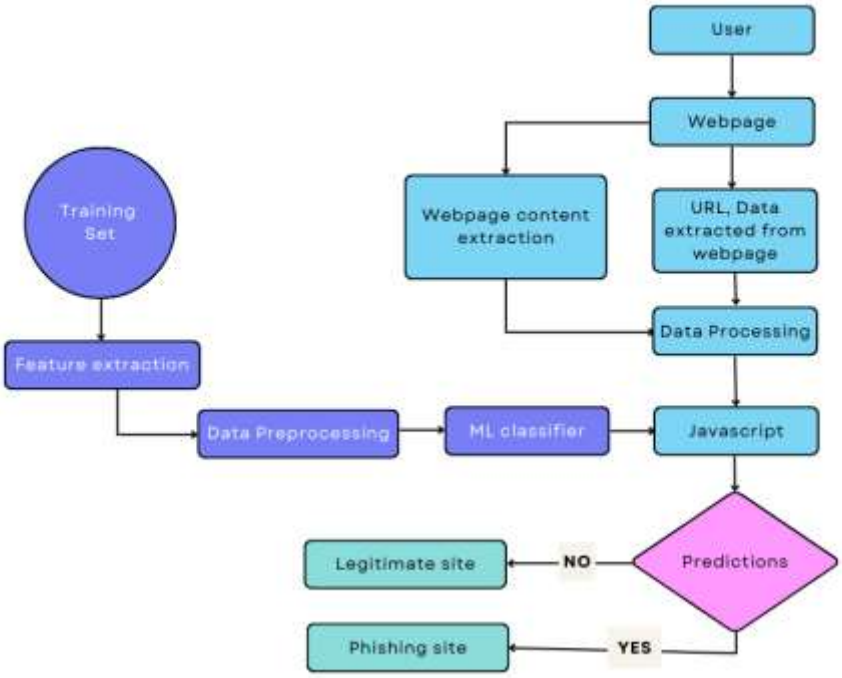


**Fig.2 System Architecture**

**B. Results :**

**Pre–Processing of Dataset (Non-functional Result):**

Sci-kit Learn library loads the phishing dataset that was referred to from the repository into the array. About 25 occurrences in the phishing dataset need to be extracted using the classifier. Because each characteristic has potential outcomes, this feature extraction is crucial in identifying the class. More features prepared for classification aid in more precise class prediction. The mentioned phishing dataset is divided into training and testing sets in a 7:3 ratio using the k-fold validation technique [18][19].

The project consists of a Random Forest algorithm using the Confusion Matrix. We can see the accuracy and runtime from the result. After examining Random Forest, Support Vector Machine, and Neural Network, Random Forest was chossen since it produced accurate results and required less processing time [20][21].

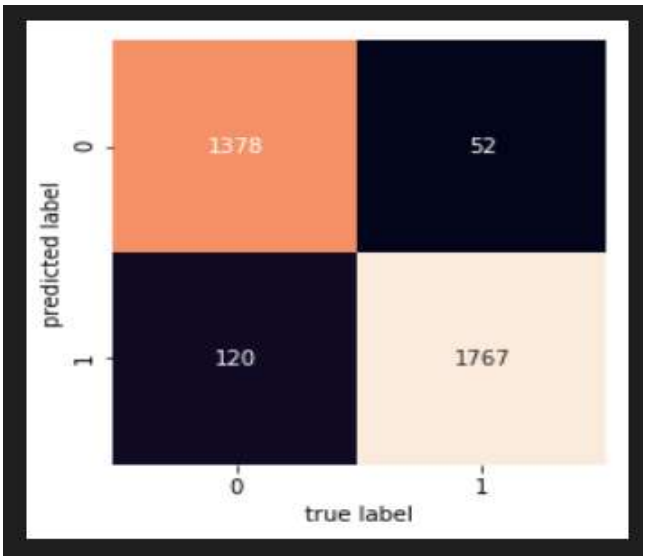


**Fig. Confusion Matrix**

| | ANN | k-NN | SVM | C4.5 | Random Forest | Rotation Forest |
|---|---|---|---|---|---|---|
| **ROC Area** | 0.995 | 0.989 | 0.97 | 0.984 | 0.996 | 0.994 |
| **F – measure** | 0.969 | 0.972 | 0.972 | 0.959 | 0.974 | 0.968 |
| **Accuracy** | 96.91% | 97.18% | 97.17% | 95.88% | **97.36%** | 96.79% |

**Fig.Performance Matrix**

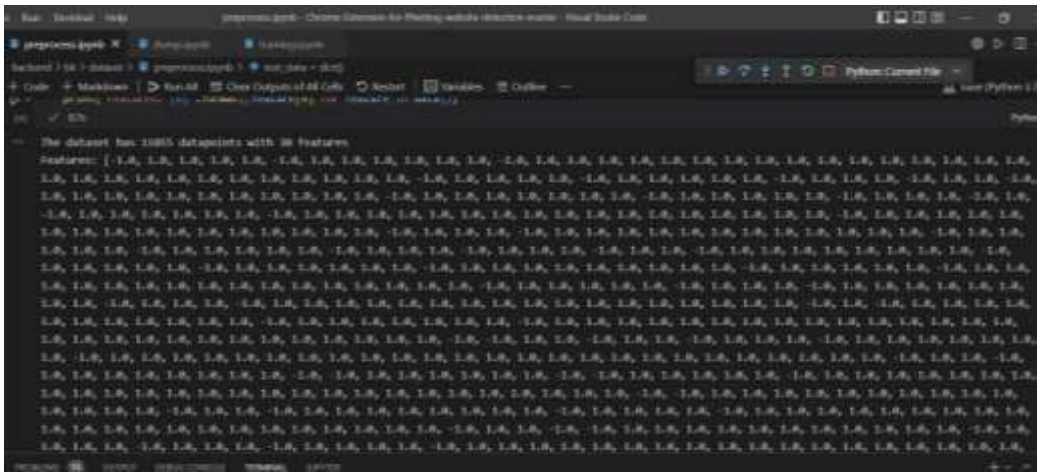

**Fig.Pre-processed data**

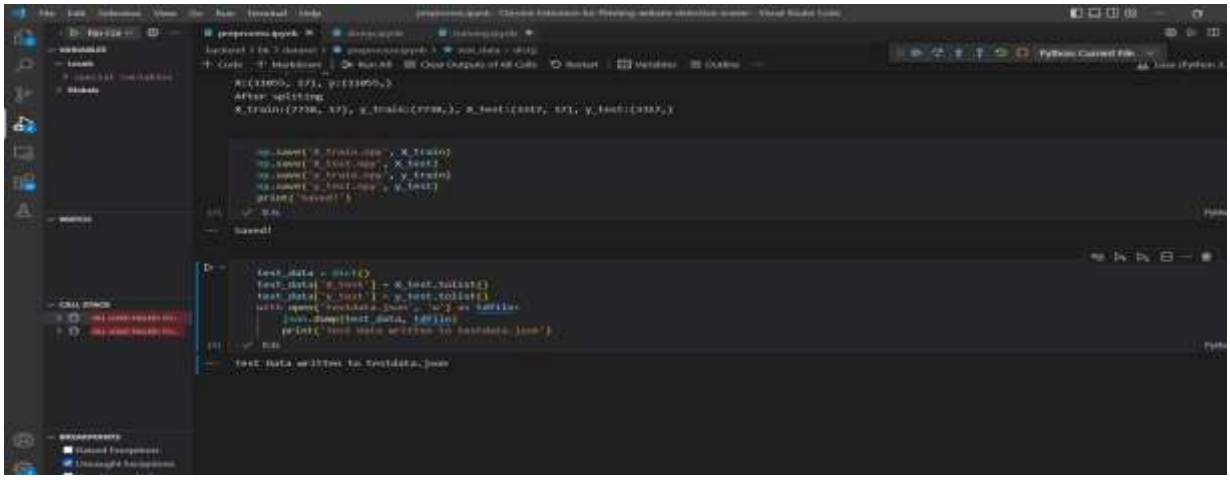

**Fig. Preprocessing output**

**Precision** is described as the percentage of relevant objects among those that were retrieved. In this instance, the percentage of URLs that are correctly identified as phished are the ones that are actually phished [22][23].

$$\text{Precision} = TP/TP+FP$$

Recall can be defined as the ratio of the total number of correctly classified positive examples divided to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN) [24].

$$\text{Recall} = TP/TP+FN$$

F Measure: The harmonic mean of recall and precision both followed by the formula below:

F-measure = 2*Recall*Precision / Recall + Precision

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1.0 | 0.92 | 0.96 | 0.94 | 1430 |
| 1.0 | 0.97 | 0.94 | 0.95 | 1887 |
| micro avg | 0.95 | 0.95 | 0.95 | 3317 |
| macro avg | 0.95 | 0.95 | 0.95 | 3317 |
| weighted avg | 0.95 | 0.95 | 0.95 | 3317 |

Accuracy Score: 95.0 %

**Fig: Evaluation of the model.**

**Training:**

The phishing dataset will now be used for model training after being preprocessed. The classifier is chosen by analysis and based on the confusion matrix, accuracy, turnaround time, and efficiency. The phishing dataset will be trained using the appropriate split ratio using the k-fold method in our case. This prevents the trained model from having problems with under-fitting and over-fitting. The file will be saved separately for the training process after the dataset has been split along the X and Y axes.

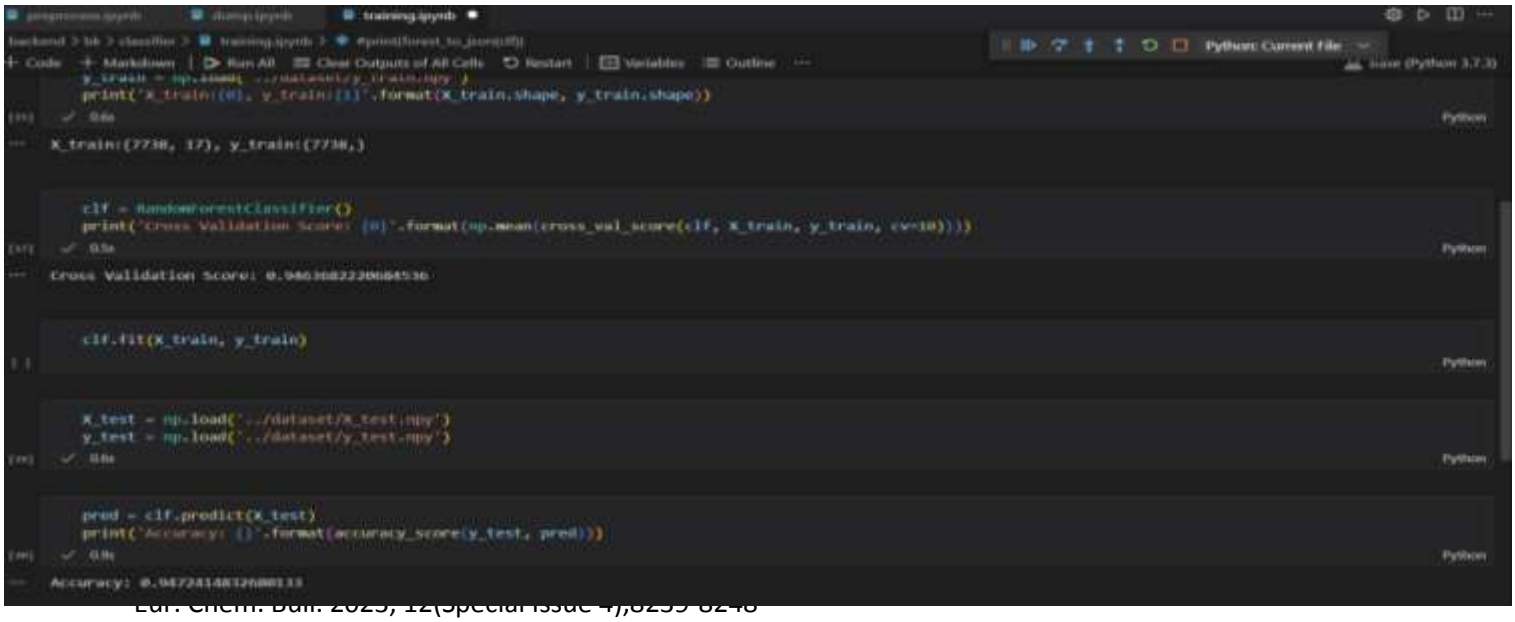

**Fig: Training output**

**Plugin Feature Extraction and Classification:**

It makes use of a content script to gain access to the webpage's DOM. Each page automatically contains the content script as it loads. The features must be gathered by the content script and sent to the plugin. The major goal of this work is to avoid utilizing any external online services, and the features must be independent of network latency and quick in terms of extraction. All of these are taken into consideration while creating strategies for feature extraction. A feature is retrieved and then encoded into the values "-1, 0, 1" (Legitimate, Suspicious, Phishing).

After the feature extraction process the data is sent to the model which is using a random forest algorithm for classification of the link.

**C. Chrome Plugin API:**

The strategy focuses on building the model using the already-collected data and Random Forest. discriminative classifier, extracting key features from the phishing dataset, and setting up the platform for Chrome. JavaScript was used during the plugin's development. Python and other libraries added support for the integration of JavaScript and JSON objects and vectors for the efficient processing of the whole functionality of distinguishing between legal and phished URLs. The user or any other person from any organization can utilize it as an automatic extension. We investigated various approaches throughout the project's first phase, and after weighing the benefits, drawbacks, and resource availability, we came to the conclusion that the above-mentioned technique would be most effective.

**5. CONCLUSION:**

An enormous global impact of phishing, a social engineering attack, is the destruction of the financial and economic value of enterprises, government sectors, and individuals by using various phishing techniques to abuse the data. Due to thorough searches and the upkeep and upgrading of a blacklist database with new links, which is not an appropriate and accurate method of identifying phishing attempts, we have overcome various traditional ways to phishing detection. So, using a machine learning algorithm, we have suggested a Chrome plugin that automates the process of identifying and defeating conventional approaches. After reviewing a number of other algorithms, we chose random forest as our primary algorithm based on its execution rate and confusion matrix using the k-fold validation. The predicted phishing link or website will be reported to the website blacklist database of google.

## 6. References

[1]https://www.transunion.com/blog/identityprotection/7-facts-about-cyber-security-and-phishing

[2]Zou Futai, Gang Yuxiang, Pei Bei, Pan Li, Li Linsen "Web Phishing Detection Based on Graph Mining", 2016.

[3]Nick Williams, Shujun Li "Simulating Human detection of phishing websites: An investigation into the applicability of ACT-R cognitive behavior architecture model", 2017.

[4] Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., ... & Ragan, E. (2021, May). Machine learning explanations to prevent overtrust in fake news detection. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 421-431).

[5] Narayan, Vipul, et al. "Enhance-Net: An Approach to Boost the Performance of Deep Learning Model Based on Real-Time Medical Images." Journal of Sensors 2023 (2023).

[6] Babu, S. Z., et al. "Abridgement of Business Data Drilling with the Natural Selection and Recasting Breakthrough: Drill Data With GA." Authors Profile Tarun Danti Dey is doing Bachelor in LAW from Chittagong Independent University, Bangladesh. Her research discipline is business intelligence, LAW, and Computational thinking. She has done 3 (2020).

[7] NARAYAN, VIPUL, A. K. Daniel, and Pooja Chaturvedi. "FGWOA: An Efficient Heuristic for Cluster Head Selection in WSN using Fuzzy based Grey Wolf Optimization Algorithm." (2022).

[8] Faiz, Mohammad, et al. "IMPROVED HOMOMORPHIC ENCRYPTION FOR SECURITY IN CLOUD USING PARTICLE SWARM OPTIMIZATION." Journal of Pharmaceutical Negative Results (2022): 4761-4771.

[9] Narayan, Vipul, A. K. Daniel, and Pooja Chaturvedi. "E-FEERP: Enhanced Fuzzy based Energy Efficient Routing Protocol for Wireless Sensor Network." Wireless Personal Communications (2023): 1-28.

[10] Tyagi, Lalit Kumar, et al. "Energy Efficient Routing Protocol Using Next Cluster Head Selection Process In Two-Level Hierarchy For Wireless Sensor Network." Journal of Pharmaceutical Negative Results (2023): 665-676.

[11] Paricherla, Mutyalaiah, et al. "Towards Development of Machine Learning Framework for Enhancing Security in Internet of Things." Security and Communication Networks 2022 (2022).

[12] Sawhney, Rahul, et al. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." Decision Analytics Journal 6 (2023): 100169.

[13] Srivastava, Swapnita, et al. "An Ensemble Learning Approach For Chronic Kidney Disease Classification." Journal of Pharmaceutical Negative Results (2022): 2401-2409.

[14] Mall, Pawan Kumar, et al. "FuzzyNet-Based Modelling Smart Traffic System in Smart Cities Using Deep Learning Models." Handbook of Research on Data-Driven Mathematical Modeling in Smart Cities. IGI Global, 2023. 76-95.

[15] Mall, Pawan Kumar, et al. "Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique." Journal of Pharmaceutical Negative Results (2022): 4784-4792.

[16] Pramanik, Sabyasachi, et al. "A novel approach using steganography and cryptography in business intelligence." Integration Challenges for Analytics, Business Intelligence, and Data Mining. IGI Global, 2021. 192-217.

[17] Narayan, Vipul, et al. "Deep Learning Approaches for Human Gait Recognition: A Review." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.

[18] Narayan, Vipul, et al. "FuzzyNet: Medical Image Classification based on GLCM Texture Feature." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023

[19] Mahadani, Asim Kumar, et al. "Indel-K2P: a modified Kimura 2 Parameters (K2P) model to incorporate insertion and deletion (Indel) information in phylogenetic analysis." Cyber-Physical Systems 8.1 (2022): 32-44.

[20] Singh, Mahesh Kumar, et al. "Classification and Comparison of Web Recommendation Systems used in Online Business." 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM). IEEE, 2020.

[21] Awasthi, Shashank, Naresh Kumar, and Pramod Kumar Srivastava. "A study of epidemic approach for worm propagation in wireless sensor network." Intelligent Computing in Engineering: Select Proceedings of RICE 2019. Springer Singapore, 2020.

[22] Srivastava, Arun Pratap, et al. "Stability analysis of SIDR model for worm propagation in wireless sensor network." Indian J. Sci. Technol 9.31 (2016): 1-5.

[23] Ojha, Rudra Pratap, et al. "Global stability of dynamic model for worm propagation in wireless sensor network." Proceeding of International Conference on Intelligent Communication, Control and Devices: ICICCD 2016. Springer Singapore, 2017.

[24] Shashank, Awasthi, et al. "Stability analysis of SITR model and non linear dynamics in wireless sensor network." Indian Journal of Science and Technology 9.28 (2016)