ICEST-2K24 [International Conference on Engineering, Science and Technology]
In Association with
International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN - 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Malware Detection for Web URL's and PE Files Using Machine Learning

Prof. Jalindar N. Ekatpure[1], Miss. Archana S Bichkule[2], Miss. Shital M kamble[2], Miss. Rutuja U Lawand[4]
*[1]Assistant Professor, S. B. Patil College of Engineering, Indapur, Maharashtra, India
[2]D S. B. Patil College of Engineering, Indapur, Maharashtra, India

ABSTRACT
Malware such as Viruses, Worms, Trojans, Backdoors are some of the threats to computer system and internet in recent years malware count is increased in millions. In the past few years millions of malwares were found in portable executable files which are downloaded from the internet. As the solution to this, it is highly desirable for users to detect such malware files, so that users can secure the devices as well as highly confidential data. Malware Detection System is an application which will detect the malwares from the portable executable files. The proposed system uses KNN algorithm to predict the malware files and legitimate files. so users can easily differentiate between them and secure their systems. The database will be generated by extracting maximum features of Portable Executable files which improves the accuracy of the model. The system implements pure machine learning algorithms to identify every malware file.
Keywords: Machine Learning, Malware, Portable Executable Files.

I. INTRODUCTION

The main domain for the paper is Machine Learning. Machine Learning is the sub- category of an Artificial Intelligence. AI gives ability to work computer as a human, in this a computer is able to do task which are usually done by human. ML specifically takes data as an input and finds pattern amongst them by using various algorithm. It may include supervised, unsupervised, reinforcement, semi-supervised learning models. The students can make malpractice for choosing answers of questions from assignments or from the online search engines like Google, Chrome, Windows Explore etc. Malware is a recent problem, which affects the data, devices, etc. Prevention of malware attack is important to save highly confidential files and the devices. In this section, let's have a quick look of the existing Malware detection methodologies and related works.

II. LITERATURE SURVEY

For this refer [1] The Online Proctoring Exam System includes as per the term "malicious malware", or malware, is used to refer to computer software that is developed for illegal purposes such as stealing data, corrupting data, damaging computers and computer systems of certain individuals and organizations. As per paper [2] both static malware examination and element malware investigation as indicated by the component of the biologic resistant framework that can shield us from disease by creatures. In this model, the static marks and dynamic marks of malware are separated, and in view of the genuine esteemed vector encoding, the

antigens are produced.[3] Malware is a program or file which harms the computer, network or server intentionally. “Malware Intrusion Detection for System Security” proposed by Mrs. Ashwini Katkar, Ms. Sakshi Shukla and Mr. Danish Shaikh in year 2021. [4]“A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection” proposed by Luca Demetrio, Scott E. Coull, Battista Biggio, Giovanni Lagorio, Alessandro Armando, Fabio Roli in year 2020. [5] This paper provides the functionality of preserving manipulations to the Windows Portable Executable (PE) file format. [6] This paper has the limitations that they didn’t uses Random Forest and Decision tree so the accuracy might vary and also this is powerful in case of Denial-of-Service attacks (DOS) only. [7]“Malware Detection using Honeypot and Malware Prevention” proposed by Dhruvi Vadaviya, Mahesh Panchal, Dr. Abdul Jhummarwala and Dr. M. B. Potdar in year 2019. [8]The main intension of this paper is to elaborate the seriousness of Malware problem and paper the importance of online malware analysis. This paper explains only about the protection regarding the network attacks.[9] Pharmaceutical innovation faces challenges. Research merges quantum computing and machine learning to revolutionize drug discovery, simulation, and safety assessment for expedited progress.[15] In this paper authors has used Honeypot system to trace the details about the hacker or the unauthorized user who is accessing the details. Proposed paper only explains about the network safety includes recording and analysis of network activities and captures, and capture, to uncover evidence of the origin of device security attacks. As per paper [10] authors had explained about identification of various harmful URLs through use of Machine Learning techniques. Author presented an algorithm for detecting and preventing Node isolation attack where attacker become the sole MPR of victim and isolated the victim from the rest of the network.[11]

III.PROPOSED SYSTEM

A. Problem Statement

Problem Statement: The problem statement of malware detection using machine learning is to develop effective and efficient techniques that can accurately identify and classify malicious software based on patterns and characteristics in the data. It's essential to keep in mind that malware detection is an ongoing arms race, with attackers constantly evolving their techniques to evade detection. Therefore, a robust and adaptive machine learning approach is crucial for effective malware detection. Additionally, privacy and ethical considerations should be taken into account when collecting and handling malware samples for training purposes.

B. Block Diagram

The block diagram is a visual representation of a system, emphasizing overall structure and functions. It features three main components Data collection Visualization, Preprocessing, Dataset Clean, Learning.

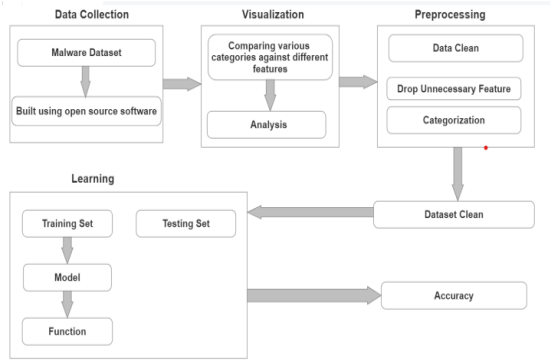

Figure 1: Block Diagram

C. Software Requirement

- Operating System - Windows 11
- Application Server – Django or Flask
- Programming Language - Python
- Front End - HTML, CSS, JavaScript
- IDE - Visual Studio Code
- Database – MySQL

D. Hardware Requirement

- Processor - Intel i5/i7
- Speed - 3.1 GHz
- RAM - 4 GB(min)
- Hard Disk - 20 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor – SVGA

E. Sequence Diagram

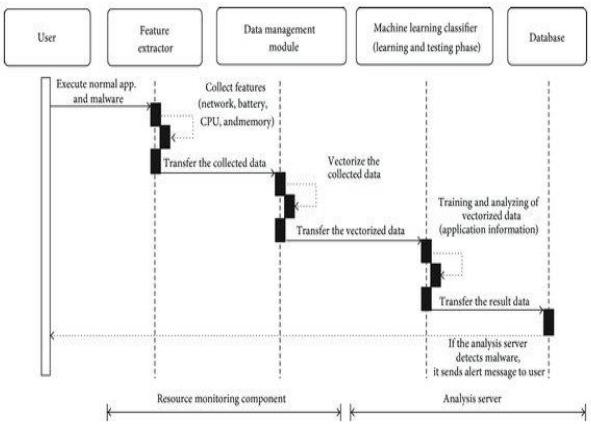

Figure 2: Use Case Diagram

IV. ALGORITHM

KNN Algorithm K-Nearest Neighbor is the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN working is based on the below algorithm: K-NN working can be explained on the basis of the below algorithm:

- Step-1: firstly, Select the value of K that denotes the number of neighbors
- Step-2: Then, Calculate the Euclidean distance of K no. of neighbors
- Step-3: Take the K nearest neighbors based on the calculated Euclidean distance.
- Step-4: Then, Count the number of the points in each category, among these K neighbors.
- Step-5: Assign the new data points to that category to which the number of the neighbor is maximum.
- Step-6: Our model is ready

A. Random Forest Algorithm

Random Forests algorithm learns from a weak model (like DT) to create a. more robust one and to avoid over-adjustment with a minimum cost. The for est is built using bootstrap techniques that is well known. The main idea behind the bootstrapping is to combine learning models by increasing the overall result of classification. To achieve RF, the following steps are performed after a dataset XN of size N is splitted using the bootstrap technique:

- Draw n size random bootstrap sample, where randomly choose n samples from the training set with replacement
- Construct a decision tree from the random bootstrap sample.

At each node:

- Randomly select d features without replacement
- Use the feature to split the node, that provides the best split 3) Repeat the steps 1 to k times. 4) To assign the class label by a majority vote, aggregate the prediction by each tree:
- Select D(xN) = xb along with N samples without replacement.
- Create a bootstrap dataset B in 1, ..., N With the previously assumptions is computed Equation: P rob(K) = N!(K!(N K))(1N)K N1N; 0 k N (13) where P rob(K) is a estimation of probability, N states the number of samples and K is an iteration.

B. Decision Tree

Decision Tree is a supervised classification algorithm that begins by growing with a single-leaf tree i.e. root and through all classes of the training samples assigned to its own sheet with class by a majority voting technique. DT algorithm starts with a one node, which is evaluated by computing possible outcomes. Each outcome inserts the additional nodes, which branch off into other possibilities. The main idea of DT algorithm is to divide the dataset into smaller sets depends on the most descriptive features until is reached the smallest set containing data points that fall under one label. At the end the iterations stops until the zero impurity, this is when entropy (the degree uncertainty) is minimized. The results are demonstrated by identifying which features maximizes the gain of information as depicted in Equation

IG(Dp, xi) = I(Dp) NleftNp I(Dleft) NrightNp I(Dright) where IG(Dp, xi) states the Gini impurity which is a measure to calculate probabilistically different outputs, Dp is the dataset of the parent and child nodes, Np is the total of samples at the parent node, I is the impurity measure, D left and Dright are two child nodes.

V. RESULT DISCUSSION

Our experiments revealed that the:

- The Module is very helpful and productive for learning the detection of malware from the specified features.
- Module has flexible GUI which is understandable to any user.
- The module is divided into four sub modules as upload dataset, show dataset, clean dataset and prediction for test dataset. The visualization of each and every sub modules is clear and easy to understand for any user.
- Each sub module performs its specified functions as upload dataset upload the dataset from system by opening window, show dataset shows the uploaded dataset, clean dataset shows the cleaned dataset and after all the functioning it predicts the result according to the features as it is malware or not.

VI. RESULT

Here in this section we will discuss about the result of our proposed system.
Let us know about the input and output to the proposed system. In this proposed system Malware are present or not in the test file or URL's when we can upload test file or URL's then shows the test file are legitimate or not and URL's are safe or not for the system.

A. Input

Create an account and register with our email id and enter you're password after that open the option test file and upload the file with extension .exe and .dll etc.

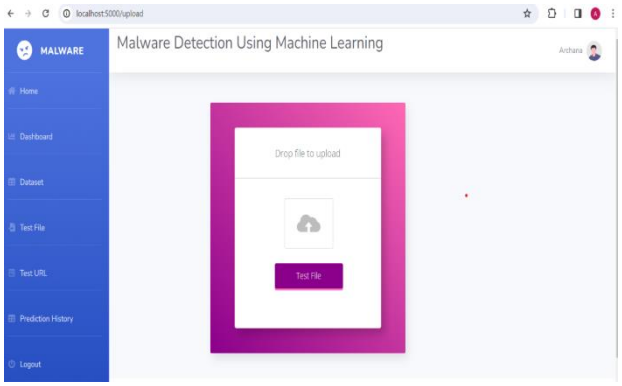

Figure 3: Test file

B. Output

Shows the output malware are present or not after upload the test file.

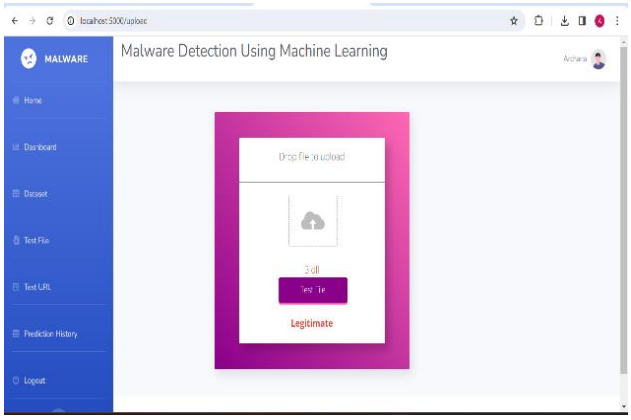

Figure 4: output

C. Input

Upload any type of URL like www.youtube.com,www.indiapost.gov.inetc.

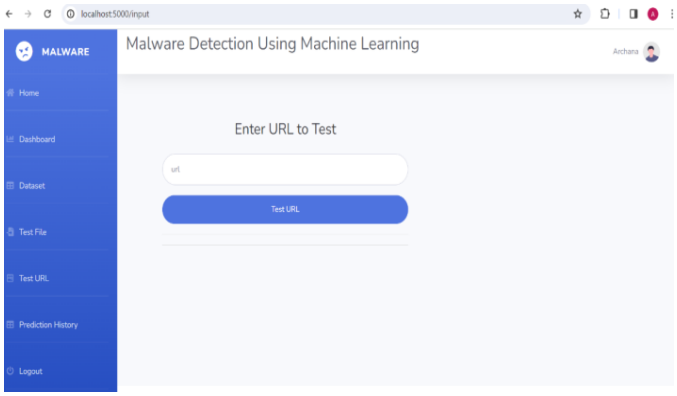

Figure 5: Input

D. Output

Show the website or URL are safe or not for the system.

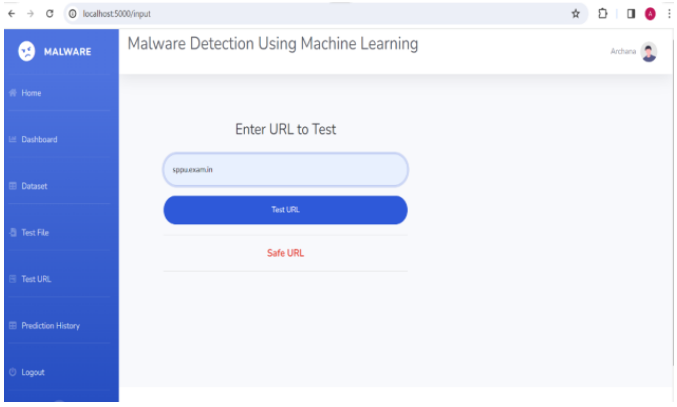

Figure 6: URL's output

VII. CONCLUSION

Image to image translation is a powerful deep learning technique that has numerous applications in computer vision. In this paper, we aim to implement image to image translation for face image synthesis from sketches.

The paper methodology involves various phases, including research and planning, data collection and preparation, model selection and training, evaluation and optimization, software development, and deployment and maintenance. The expected outcomes of the paper are a software product that can generate realistic face images from sketches, improved performance of the image-to-image translation model, and features like face morphing and face copy-paste. Overall, this paper will provide a valuable contribution to the field of text to image translation and have numerous practical application.

VIII. REFERENCES

[1]. G.D. Penna, L.D. Vita and M.T. Grifa, "MTA-KDD'19: A Dataset for Malware Traffic Detection" in ITASEC - 2020.

[2]. M. Gao, Li Ma, H. Liu, Z. Zhang, Z. Ning and J. Xu, "Malicious Network Traffic Detection Based on Deep Neural Networks and Association Analysis"in Sensors (Basel) - 6 March 2020.

[3]. Sudarshan N P.Dass, "Malicious Traffic Detection System using Publicly Available Blacklist's" in IEEE Conference of International Journal of Engineering and Advanced Technology (IJEAT) - August 2019.

[4]. Paul Prasse, Lukas Machlica, Tomas Pevny, Jiri Havelka and Tobias Scheffer,"Malware Detection by Analyzing Network Traffic with Neural Networks" inIEEE Conference of Symposium on Security and Privacy Workshops – May2017.

[5]. Nancy Agarwal and Syed Zeeshan Hussain, "A Closer Look at Intrusion Detection System for Web Applications" in IEEE Conference of Security and Communication Networks Volume - 2018.

[6]. Gonzalo Marin, Pedro Casas, German Capdehourat, " DeepMal - Deep Learning Models for Malware Traffic Detection and Classification" on 10 March2020.

[7]. Felipe N. Ducau, Ethan M. Rudd, Tad M. Heppner, Alex Long, Konstantin Berlin" Automatic Malware Description via Attribute Tagging and Similarity Embedding " on 15 May 2019.

[8]. Luca Demetrio, Scott E. Coull, B. Biggio, G. Lagorio, A. Armando, FabioRoli "A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection" on 17 Aug 2020.

[9]. T. M. Mohammed, L. Nataraj, S. Chikkagoudar, S. Chandrasekaran, B. S. Manjunath "Malware Detection Using Frequency Domain-Based Image Visualization and Deep Learning" on 26 Jan 2021.

[10]. Karve, S. M, Kakad , S., SwapnajaAmol, Gavali, A. B. ., Gavali , S. B. ., &Shirkande, S. T. . (2024). An Identification and Analysis of Harmful URLs through the Application of Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 12(17s), 456–468. https://www.ijisae.org/index.php/IJISAE/article/view/4905

[11]. K. S. Gaikwad and S. B. Waykar, "Detection and Removal Of Node Isolation Attack In OLSR Protocol Using Imaginary Nodes with Neighbour Response in MANET," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2017, pp. 1-5, doi: 10.1109/ICCUBEA.2017.8463762.

[12]. Sairise, Raju M., Limkar, Suresh, Deokate, Sarika T., Shirkande, Shrinivas T., Mahajan, RupaliAtul& Kumar, Anil(2023) Secure group key agreement protocol with elliptic curve secret sharing for authentication in distributed environments, Journal of Discrete Mathematical Sciences and Cryptography, 26:5, 1569–1583.

[13]. Upadhye, P. A., Ghule, G., Tatiya, M., Shirkande, S. T., Kashid, S. M., &Bhamare, D. Optimizing Communication Systems with Applied Nonlinear Analysis Techniques, Communications on Applied Nonlinear Analysis, https://doi.org/10.52783/cana.v30.277

[14]. Nalawade, V. S., Ashok, G. K., Hanumant, B. A., &Reshma, G. (2021). ENCRYPTION THEN COMPRESSION BASED SYSTEM USING GRAYSCALE BASED IMAGE ENCRYPTION FOR JPEG IMAGES.

[15]. Ekatpure, J. N., Jadhav, P., Gavali, R., Kale, P., & Padasalkar, S. (2023). Pharmaceutical Data Optimisation Using Quantum Machine Learning.