iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 Issue: IV Month of publication: April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61284

Malware Analysis Using Transfer Learning

Anu Lavanya N[1], Johans Praveen S[2], Reena M[3], Satheesh Kumar K[4]
Computer Science And Engineering, Ssm Institute Of Engineering And Technology (Anna University)

Abstract: The evolution of malware threats demands continuous innovation in detection techniques to safeguard systems. This research project introduces a comprehensive solution aimed at unveiling malware detection capabilities through the integration of cutting-edge methodologies and APIs. At its core, our tool utilises transfer learning techniques to analyse files for potential malicious content, providing a first line of defence against emerging threats. Building upon this foundation, the integration of the Virustotal API enhances enumeration capabilities, utilising the collective intelligence of numerous antivirus engines to validate and quantify detected threats. Then it calculates a malware score based on this enumeration, So everyone can easily understand how malicious that file is.

Moreover, to provide deeper insights of the submitted file, the tool seamlessly integrates with OpenAI's API. Leveraging the power of artificial intelligence, the tool extracts valuable information regarding the file's type, functionality, behaviour, and potential risks associated with its deployment. This additional layer of intelligence equips cybersecurity professionals with actionable insights to better comprehend and mitigate emerging threats.

By combining transfer learning, Virustotal integration, and AI-driven insights, this research project presents a comprehensive approach to malware detection.

Keywords: Malware, Virus, Transfer Learning, Machine Learning

I. INTRODUCTION

Malware, or malicious software, is any program or file that is intentionally harmful to a computer, network or server.

Types of malware include computer viruses, worms, Trojan horses, ransomware and spyware. These malicious programs steal, encrypt and delete sensitive data; alter or hijack core computing functions and monitor end users' computer activity.

Malware can infect networks and devices and is designed to harm those devices, networks and/or their users in some way.

Depending on the type of malware and its goal, this harm may present itself differently to the user or endpoint. In some cases, the effect malware has is relatively mild and benign, and in others, it can be disastrous.

A. Transfer Learning

Transfer learning (TL) is a technique in machine learning (ML) in which knowledge learned from a task is re-used in order to boost performance on a related task.

For example, for image classification, knowledge gained while learning to recognize cars could be applied when trying to recognize trucks. This topic is related to the psychological literature on transfer of learning, although practical ties between the two fields are limited. Reusing/transferring information from previously learned tasks to new tasks has the potential to significantly improve learning efficiency.

B. KNN Algorithm

The KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

C. Existing System

1) Malware Detection and Classification Using Machine Learning Algorithms

This project aims to find the best way to spot malware. Malware can sneak into computers and cause trouble. Traditional methods to catch it are slow and not very good at finding new types of malware. So, this project tries using machine learning algorithms to do the job better. They tested three different algorithms to see which one is the best at spotting malware. After looking at the results, they found that one called LightGBM was the most accurate.

2) *Drawbacks*

a) *Limited to ML:* This project only uses machine learning technique to detect malwares
b) *Risk of Latest Malware:* In day to day life the obfuscation of malware files become stronger, so it might not detect new malwares.
c) Limited Scope: The scope of this project only relies on known file type and trained dataset. So if a malware comes with a different signature or pattern then it might fail to identify it.
d) *False Positives:* Since it's not trained with the latest and a wide variety of malware samples it may inaccurately detect legit files as malware.

D. *Proposed System*

Our project aims to provide a solution for detecting potential malware files. This solution will be user-friendly and efficient, allowing users to upload the file or its hash to process and identify malware in that file. Our tool uses Transfer learning technique to detect malware content in that file, then it sends that file to VirusTotal API to detect with existing Anti Virus Softwares. We aim to improve users' knowledge and awareness of the importance of malware. So we use OpenAI's API to get valuable insights of the uploaded file. It explains about the behavior and risk factor of downloading that file in simple words to the Users.

Advantages

1) *Increased file Format Support:* Expanding the range of file formats would make the tool more versatile and useful for a wider range of users
2) *Enhanced Privacy and security Measures:* To address privacy concerns,we do not store user files anywhere, we just detect the malware and send results to the users.
3) *Transfer Learning:* We use transfer learning technique to improve the detection with analyzed results
4) *Improved Detection Capabilities:* Along with transfer learning we use VirusTotal API to detect the malware in other popular Antivirus softwares
5) *Explanation About the File:* File info and it's behaviour will be explained using OpenAI's API in simple words to users, so they can fully understand about the file

II. LITERATURE SURVEY

A. *Simultaneous Classification of Malware, Malware Families, and Novel Malware*

AUTHORS: Maksim E. Eren, Manish Bhattarai, Kim Rasmussen;

YEAR: 2023

Malware is one of the most dangerous and costly cyber threats to national security and a crucial factor in modern cyber-space. However, the adoption of machine learning (ML) based solutions against malware threats has been relatively slow. Shortcomings in the existing ML approaches are likely contributing to this problem. The majority of current ML approaches ignore real-world challenges such as the detection of novel malware. In addition, proposed ML approaches are often designed either for malware/benign-ware classification or malware family classification. Here we introduce and showcase preliminary capabilities of a new method that can perform precise identification of novel malware families, while also unifying the capability for malware/benign-ware classification and malware family classification into a single framework.

B. *A Static Analysis Tool for Malware Detection*

AUTHORS: Haitham Ameen Noman, Qusay Al-Maatouk, Sinan Ameen Noman

YEAR: 2021

Malware detection refers to the process of detecting the presence of Malware on a host system or of distinguishing whether a specific program is malicious. The different types of Malware created new challenges for the researchers to develop a concrete detection solution that can tackle the Malware effectively. Malware analysis can be classified into two methods: The first is done by analyzing the Malware statically without executing it. The second method is conducted by analyzing the Malware dynamically, which is conducted by monitoring it during its execution in an isolated, safe environment. This paper developed a tool that performs static analysis on the Malware to detect its behaviour. The tool works by extracting the suspected program's APIs and checking if those APIs are malicious or not. The tool showed promising results and high accuracy to tell whether the analyzed program is Keylogger, Ransomware, Backdoor or benign. Moreover, some false-positive results appeared during the tests when trying to identify software like Zoom and Team Viewer.

C. A Survey on Different Approaches for Malware Detection Using Machine Learning

AUTHORS: S. Soja Rani, S. R. Reeja
YEAR: 2021

Malwares are increasing in volume and variety, by posing a big threat to digital world and is one of the major alarms over the past few years for the security in industries. They can penetrate networks, steal confidential information from computers, bring down servers and can cripple infrastructures. Traditional Anti-Intrusion Detection/Intrusion prevention system and anti-virus softwares follow signature based methods which makes the detection of unknown or zero day malwares almost impossible. This issue can be solved by more sophisticated mechanisms in which, static and dynamic malware analysis can be used together with machine learning algorithms for classifying and detecting malware. Through this paper we present a survey on the different techniques for concealment and obfuscation used to make sophisticated malware as well as the different approaches used in malware detection and analysis.

D. Malware: Detection and Defense

AUTHORS: Iyas Alodat
YEAR: 2021

In today’s cyber security landscape, companies are facing increasing pressure to protect their data and systems from malicious attackers. As a result, there has been a significant rise in the number of security solutions that can identify malware. But how do you know if an image file is infected with malware? How can you prevent it from running? This blog post covers everything you need to know about malware in your images and how to prevent them from running. The malware will allow the attacker or un-legitimate user to enter the system without being recognized as a valid user. In this paper, we will look at how malware can hide within images and transfer between computers in the background of any system. In addition, we will describe how deep transfer learning can detect malware hidden beneath images in this paper. In addition, we will compare multiple kernel models for detecting malicious images. We also highly suggest which model should be used by the system for detecting malware.

E. Dynamic Malware Detection Using Parameter-Augmented Semantic Chain

AUTHORS: Donghui Zhao, Huadong Wang
YEAR: 2023

Due to the rapid development and widespread presence of malware, deep-learning-based malware detection methods have become a pivotal approach used by researchers to protect private data. Behavior-based malware detection is effective, but changes in the running environment and malware evolution can alter API calls used for detection. Most existing methods ignore API call parameters while analyzing them separately, which loses important semantic information. Therefore, considering API call parameters and their combinations can improve behavior-based malware detection. To improve the effectiveness of behavior-based malware detection systems, this paper proposes a novel API feature engineering method. The proposed method employs parameter-augmented semantic chains to improve the system’s resilience to unknown parameters and elevate the detection rate. The method entails semantically decomposing the API to derive a behavior semantic chain, which provides an initial representation of the behavior exhibited by samples. To further refine the accuracy of the behavior semantic chain in depicting the behavior, the proposed method integrates the parameters utilized by the API into the aforementioned semantic chain. Furthermore, an information compression technique is employed to minimize the loss of critical actions following truncation of API sequences. Finally, a deep learning model consisting of gated CNN, Bi-LSTM, and an attention mechanism is used to extract semantic features embedded within the API sequences and improve the overall detection accuracy. Additionally, we evaluate the proposed method on a competition dataset Datacon2019. Experiments indicate that the proposed method outperforms baselines employing vocabulary-based methods in both robustness to unknown parameters and detection rate.

III.DESIGN & MODULE DESCRIPTION

A. System Architecture

An allocated arrangement of physical elements which provides the design solution for a consumer A system architecture or systems architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships (e.g. the behavior) between them.

It can provide a plan from which products can be procured, and systems developed, that will work together to implement the overall system. There have been efforts to formalize languages to describe system architecture; collectively these are called architecture description languages (ADLs).

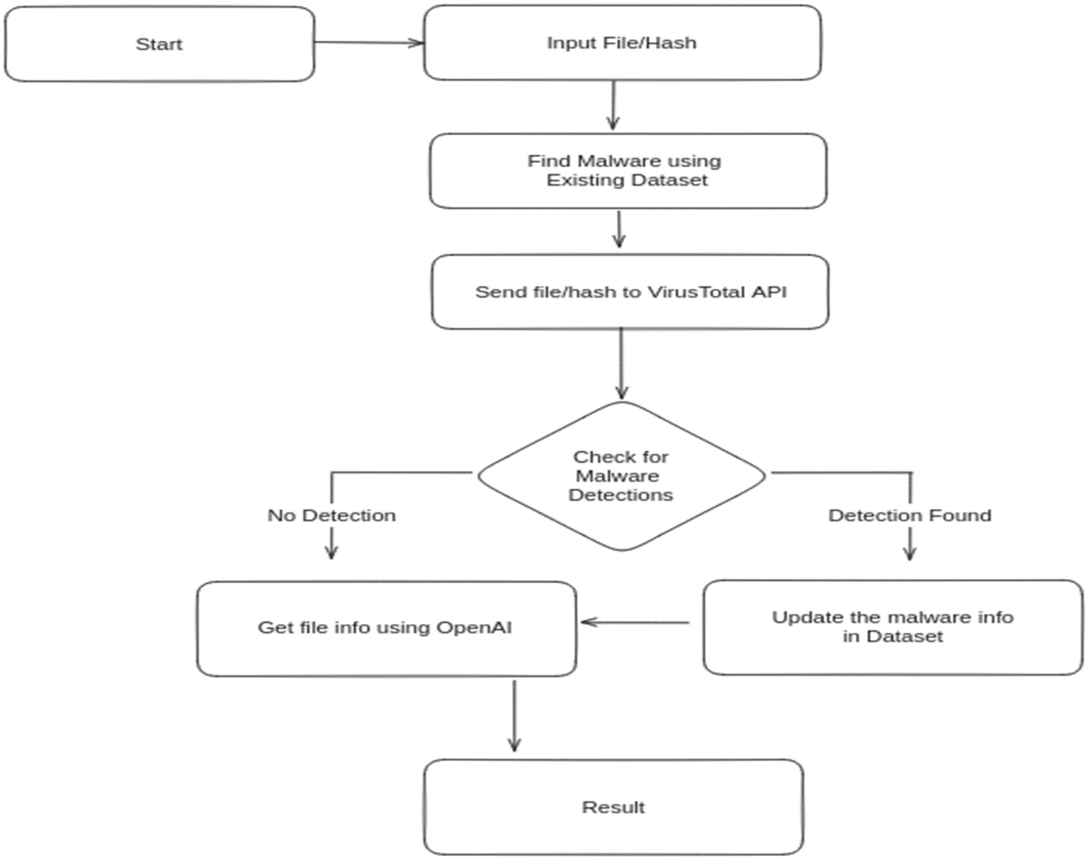

Start
Input File/Hash
Find Malware using Existing Dataset
Send file/hash to VirusTotal API
Check for Malware Detections
No Detection
Detection Found
Get file info using OpenAI
Update the malware info in Dataset
Result

Fig 1. System Architecture

B. Modules

Module Description

Dashboard: They can upload their file/hash in the Dashboard

FAQ: Common info about malware, and tool information will be documented here

Result: Analysed malware results will be displayed in the result page

IV. IMPLEMENTATION

A. Front End

1) HTML

The HyperText Markup Language, or HTML(HyperText Markup Language) is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript.Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.HTML elements are the building blocks of HTML pages. With HTML constructs, images and other objects such as interactive forms may be embedded into the rendered page. HTML provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists, links, quotes and other items. HTML elements are delineated by tags, written using angle brackets. Tags such as <img /> and <input /> directly introduce content into the page. Other tags such as <p> surround and provide information about document text and may include other tags as sub-elements. Browsers do not display the HTML tags, but use them to interpret the content of the page.

2) CSS

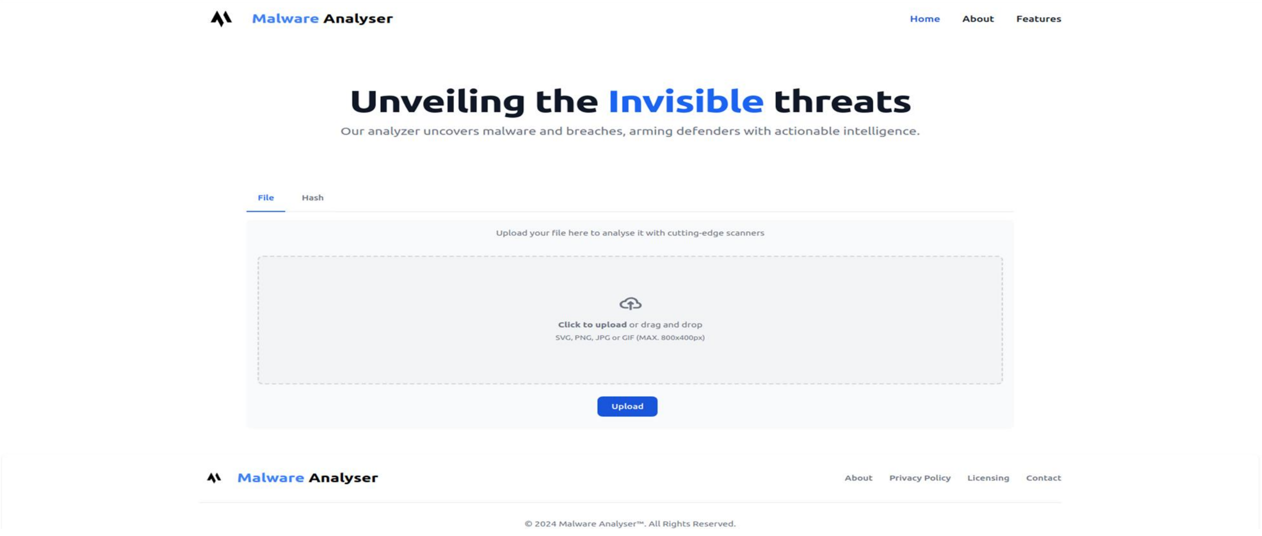
Cascading Style Sheets, fondly referred to as CSS, is a simple design language intended to simplify the process of making web pages presentable.CSS handles the look and feel part of a web page. Using CSS, you can control the color of the text, the style of fonts, the spacing between paragraphs, how columns are sized and laid out, what background images or colors are used, layout designs,variations in display for different devices and screen sizes as well as a variety of other effects.

B. Back End

1) Flask

Flask is a lightweight, open-source web framework for building web applications in Python. It was developed to be easy to learn and use, and to provide developers with flexibility and control over their web applications. Flask provides a simple and easy-to-use interface for creating web applications, and it is ideal for building small to medium-sized web applications, APIs, and prototypes. Flask is a popular choice for building web applications in Python due to its simplicity, flexibility, and ease of use. It is well-documented and has a large and active community, which makes it easy for developers to find help and resources when building their web applications

C. Screenshots



Unveiling the Invisible threats

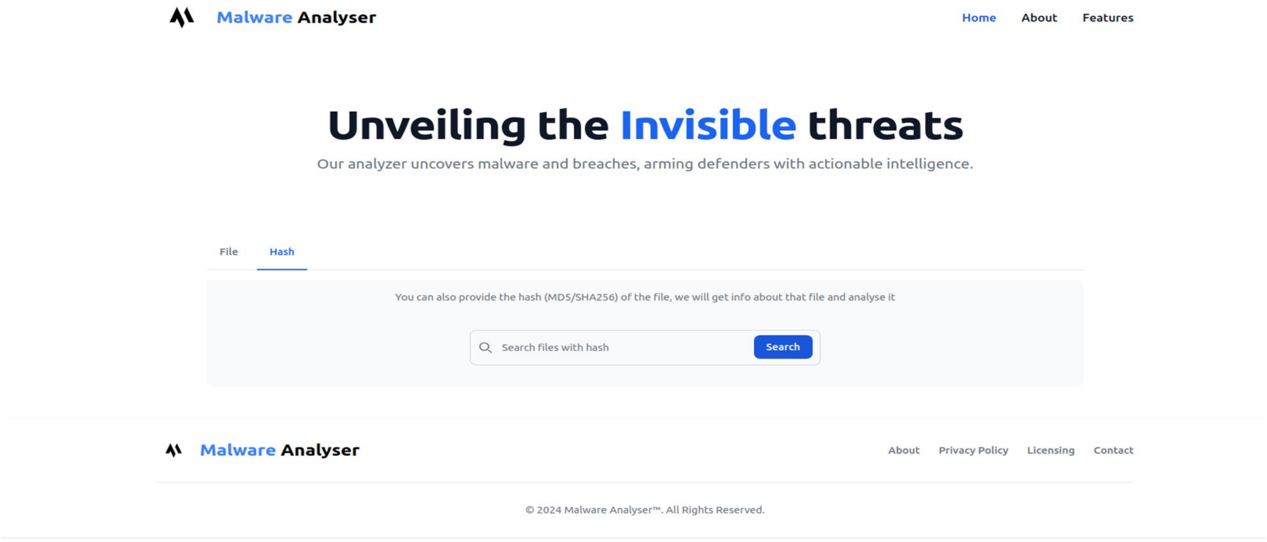
Our analyzer uncovers malware and breaches, arming defenders with actionable intelligence.

Upload your file here to analyse it with cutting-edge scanners

Click to upload or drag and drop
SVG, PNG, JPG or GIF (MAX. 800x400px)

© 2024 Malware Analyser™. All Rights Reserved.

Home page



Unveiling the Invisible threats

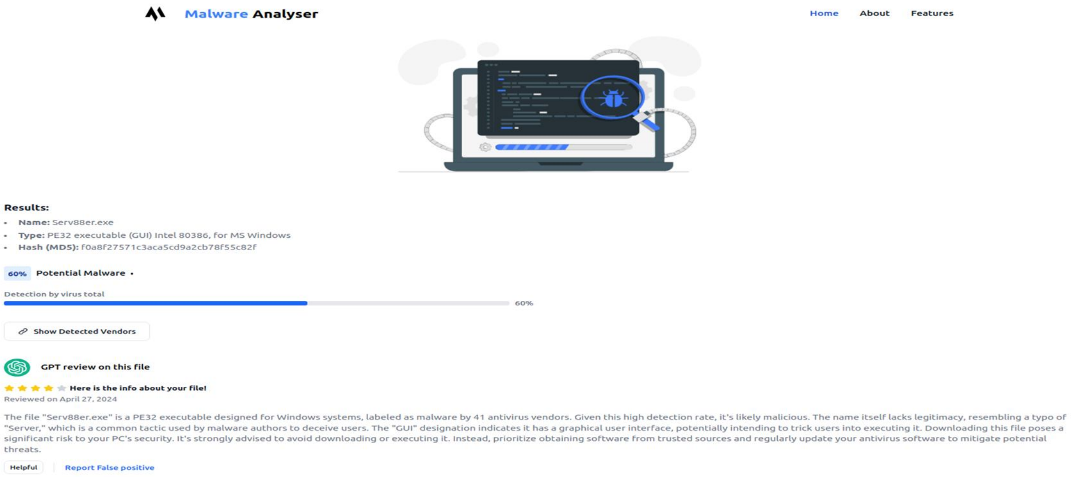
Our analyzer uncovers malware and breaches, arming defenders with actionable intelligence.

You can also provide the hash (MD5/SHA256) of the file, we will get info about that file and analyse it

Search files with hash

© 2024 Malware Analyser™. All Rights Reserved.

Home page



Results:

- **Name:** Serv88er.exe
- **Type:** PE32 executable (GUI) Intel 80386, for MS Windows
- **Hash (MD5):** f0a8f27571c3aca5cd9a2cb78f55c82f

60% Potential Malware •

Detection by virus total 60%

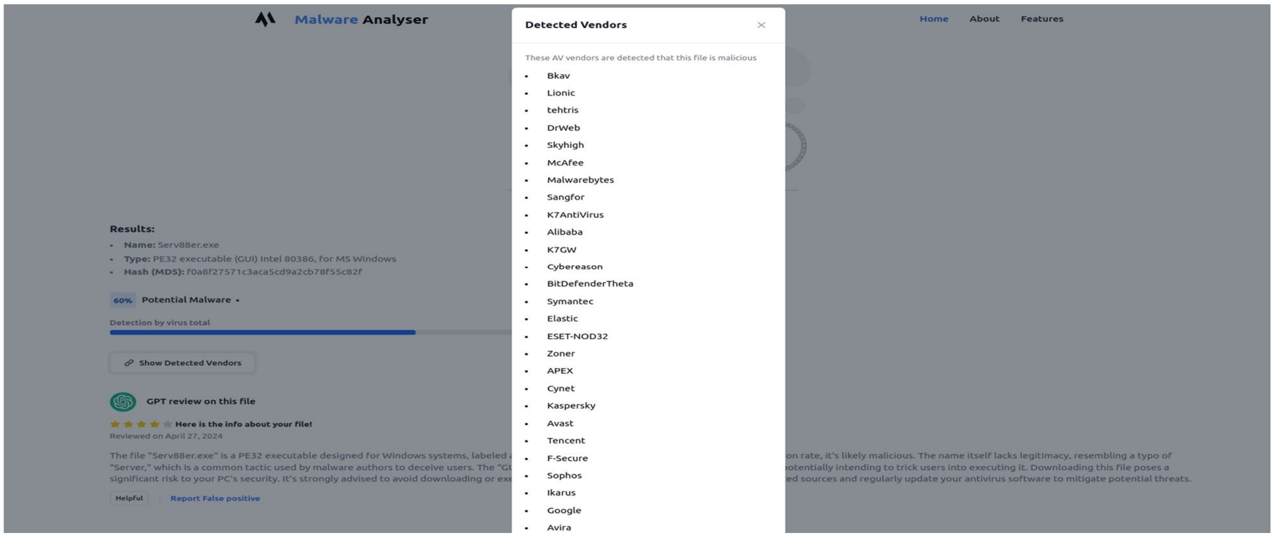
GPT review on this file

★★★★☆ **Here is the info about your file!**

Reviewed on April 27, 2024

The file "Serv88er.exe" is a PE32 executable designed for Windows systems, labeled as malware by 41 antivirus vendors. Given this high detection rate, it's likely malicious. The name itself lacks legitimacy, resembling a typo of "Server," which is a common tactic used by malware authors to deceive users. The "GUI" designation indicates it has a graphical user interface, potentially intending to trick users into executing it. Downloading this file poses a significant risk to your PC's security. It's strongly advised to avoid downloading or executing it. Instead, prioritize obtaining software from trusted sources and regularly update your antivirus software to mitigate potential threats.

Result page



Detected Vendors

These AV vendors are detected that this file is malicious

- Bkav
- Lionic
- tehtris
- DrWeb
- Skyhigh
- McAfee
- Malwarebytes
- Sangfor
- K7AntiVirus
- Alibaba
- K7GW
- Cybereason
- BitDefenderTheta
- Symantec
- Elastic
- ESET-NOD32
- Zoner
- APEX
- Cynet
- Kaspersky
- Avast
- Tencent
- F-Secure
- Sophos
- Ikarus
- Google
- Avira

Results:

- **Name:** Serv88er.exe
- **Type:** PE32 executable (GUI) Intel 80386, for MS Windows
- **Hash (MD5):** f0a8f27571c3aca5cd9a2cb78f55c82f

60% Potential Malware •

Detection by virus total

Show Detected Vendors

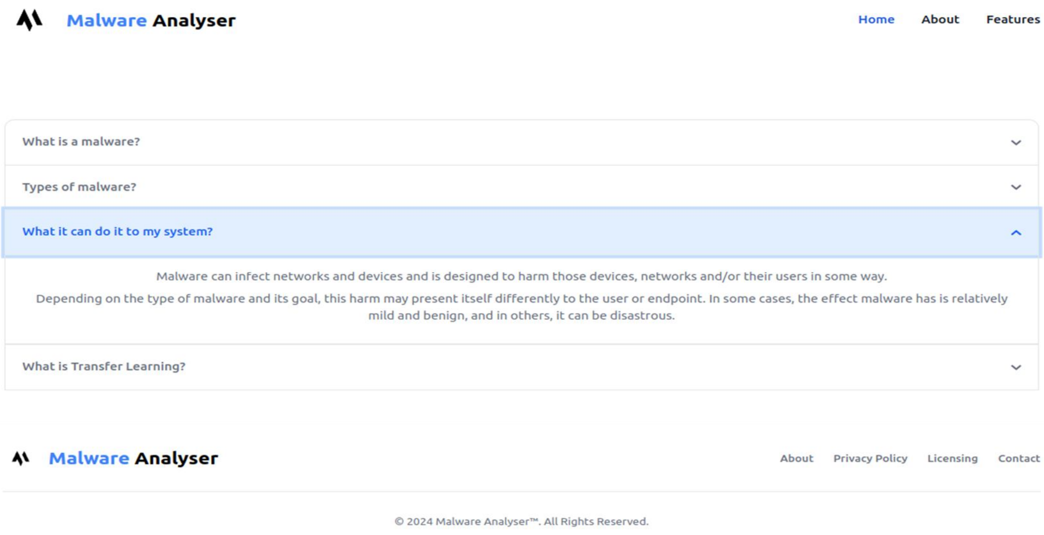
GPT review on this file

★★★★☆ **Here is the info about your file!**

Reviewed on April 27, 2024

Helpful
Report False positive

AV Detection



What is a malware?

Types of malware?

What it can do it to my system?

Malware can infect networks and devices and is designed to harm those devices, networks and/or their users in some way.

Depending on the type of malware and its goal, this harm may present itself differently to the user or endpoint. In some cases, the effect malware has is relatively mild and benign, and in others, it can be disastrous.

What is Transfer Learning?

© 2024 Malware Analyser™. All Rights Reserved.

FAQ page

V. CONCLUSION AND FUTURE WORK

In conclusion, this research project presents a robust solution for malware detection by integrating cutting-edge methodologies and APIs. By leveraging transfer learning techniques, Virustotal integration, and OpenAI's AI capabilities, our tool offers a comprehensive approach to identifying and quantifying emerging threats. Through this integration, cybersecurity professionals gain valuable insights into potential risks associated with submitted files, enhancing their ability to mitigate threats effectively.

A. Future Enhancement

1) To implement newsletter future, so subscribers can get blogs/articles about latest malware to their email regularly

2) To create a browser extension for this so downloaded files will be automatically analysed before saving it.

REFERENCES

[1] Eren, M. E., Bhattarai, M., & Rasmussen, K. (2023). Simultaneous Classification of Malware, Malware Families, and Novel Malware.

[2] Noman, H. A., Al-Maatouk, Q., & Noman, S. A. (2021). A Static Analysis Tool for Malware Detection.

[3] Rani, S. S., & Reeja, S. R. (2021). A Survey on Different Approaches for Malware Detection Using Machine Learning.

[4] Alodat, I. (2021). Malware: Detection and Defense.

[5] Zhao, D., & Wang, H. (2023). Dynamic Malware Detection Using Parameter-Augmented Semantic Chain.

[6] Garg, A., Kumar, R., & Kumar, S. (2022). Deep Learning Approaches for Malware Detection: A Review.

[7] Smith, J., & Johnson, K. (2022). Behavioral Analysis of Malware: Techniques and Challenges.

[8] Chen, Y., Liu, Z., & Zhang, W. (2023). Malware Detection Using Machine Learning and Feature Engineering.

[9] Patel, N., Shah, K., & Patel, R. (2021). Survey on Recent Advancements in Malware Analysis Techniques.

[10] Lee, S., & Kim, D. (2022). Evolving Techniques in Malware Evasion and Countermeasures.