Implementing Steganalysis using Machine Learning

Prasanta Kumar Sahoo
Prof. Department of CSE
Sreenidhi Institute of Science & Technology

Sai Manigopal Reddy
Student, Dept. of CSE
Sreenidhi Institute of Science & Technology

Kalyan Sai Chinthala
Student, Dept. of CSE
Sreenidhi Institute of Science & Technology

Badiga Srinivas
Student, Dept. of CSE
Sreenidhi Institute of Science &Technology

Karthik Lingala
Student Department of CSE
Sreenidhi Institute of Science and Technology

ABSTRACT
Nowadays hackers are using many novel techniques to hide malicious files inside an image and later sent these images to target users. These malicious image files allow the hackers to get access to the system and control the victim system from their remote location. In order to prevent this, a Django interface is proposed in this research work to verify whether the file is malicious or not? The interface allows the user to choose an image after downloading from an email or any other sources. The image is then sent to the back end server to verify if there is any malicious file hidden inside the image or not? If there is any hidden file then it is extracted from the image using steganalysis, which is a technique used to extract a hidden file. This classification is done by using machine learning technique where the machine is trained using the dataset containing features of malicious and non-malicious files. Random Forest algorithm is used for classification purpose to classify the file is malicious or not. So when an executable file is hidden inside an image, the file is extracted and the features of that executable file are then sent to the already trained machine learning model. If it is malicious then a message will appear displaying that the file is malicious and if not, a message will appear, displaying that the file is safe to use. This research work is implemented and is then deployed in IIS web server in windows server 2019, so that any user can access the website and can upload single image or multiple images. The website can be accessed through mobile phones or laptops.

Keywords
Malicious image, stenography, Machine learning

1. INTRODUCTION
There has been an escalated number of security attacks in the past decade. With the increasing number of vulnerabilities each passing day new type of attacks are being created toexploit them .There has been many security advancements to secure the data from attackers but some of the attacks can still sneak under those security systems without a trace. So it is becoming utmost importance to stay up to date with the new attacks. It is a very serious security problem that the modern world is facing today in cyber space which leads to financial losses for individuals and the society at large [1]. It has become a really necessary communication medium for industry, people, organizations, and also for the society [2]. In global internet connection, [3] In the battles between attackers and security researchers, attackers attempts to break any defense mechanism by masquerading, social engineering or hindering antivirus software detection so that they can stay in the hosts as long as possible [4]. With the ever increasing use of Internet by different stake holders in various fields, information on web browsers and servers is highly susceptible to different security attacks [5]. To communicate securely people use cryptography using secret key so that only the genuine user can decrypt the message and the integrity of the message remains intact. But cryptography raised so many suspicions, as attackers tries to attack the message to get the secretive messages shared among the users [6-7]. Steganography is used to embed the secret data within the multimedia contents such as file, message, image, or video. It is mainly concerned with concealing the fact that the secret data is being sent covertly as well as concealing the contents of the secret data [8-9].Lots of downloadablecontent is available on the web, which users knowingly or unknowingly download in to theirsystem, completely unaware of the maliciousness of the content. The content can be files withextensions like .exe, .dll, .zip, .tar which can be malicious or non malicious and images or software etc .It is necessary to ensure that these files are safe before using them for any kind ofpurpose. So, a system which could even detect the hidden files and check their maliciousness iscreated to ensure security to the user's data and make the system publicly available so that the users can upload a single image or multiple images and check for any malicious content.

2. EXISTING SYSTEM
Existing system can identify zip files, exe files, rar files and blocks them from sending. But if the same files are hidden in an image, it cannot identify those files and allows that image tosend. These files may be malicious or non-malicious. If at all those files are malicious, it harms the computer and there is also a chance of getting hacked and the existing system focuses only on extracting the hidden files but not about classifying them as malicious or not. So the existing system is inefficient to identify hidden malicious files in an image. These image files are malicious, they may allowhackers to get access to the system and that also allows hackers to control the victim system from their remote location. Users who are unaware of this often become victims of data loss and other important credentials.

3. PROPOSED SYSTEM
The proposed system to detect malicious files completely deployed in the cloud in IIS web server. A domain name steganoml.codes is given to the website and the link to our

website is http://www.steganoml.codes**.** In this system the user can access the website by typing the domain name in any browser. The user is asked to upload an image or multiple images through the Django interface. Then the image is checked for hidden file and if any hidden file is detected it is extracted by using steganalysis. The extracted file can be malicious or non-malicious. So, features of the file are extracted using pefile module in python which is used to extract information from headers of a portable executable file. Classification can be done by using machine learning where the machine is trained using the dataset containing features of malicious and non-malicious files. Random Forest algorithm can be used for classifying a file as malicious or not. If the file is classified to be malicious then a output appears on the screen thatit is malicious, not safe to use and if it is classified to be legitimate then the output is legitimate and safe to use. This system has an accuracy of 99% .This system helps to extract hidden fileand identify whether the file is malicious or non-malicious.

The terms "system analysis" and "requirements analysis" can be interchangeable. It can also be used to assist someone (the decision maker) to identify the best course of action and make a better decision than he would have made. This process entails brainstorming and breaking down the system into components in order to examine the scenario, assess the project goals, and break down what needs to be built and used to engage people in order to specify clear requirements.

4. ARCHITECTURAL MODEL

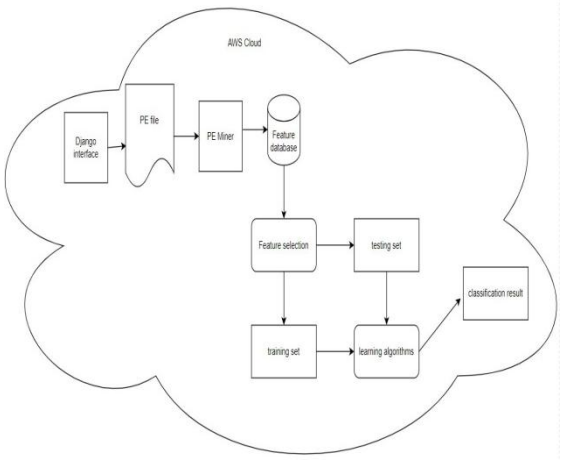

Fig 1: Shows the architectural model

The objective of this project is to detect if there is any file hidden in the LSB of the image or not , if any file is found it is sent to the already trained machine learning model which predicts if the file is malicious or not. If the file is predicted as malicious then the user is prompted to delete the image file and an alarm is raised and if the file is predicted as legitimate then the user will be prompted to use the image file. The steps of the system are given below.

4.1 Data Collection

This module deals with gathering of the training and test data from the internet. The gathered data set from the internet are then extracted. Hence in this module initially the gathered dataset consists of both malicious and legitimate files. The dataset consist of 53 features. Each file has some value for each feature. This dataset is then used for training and testing purpose.

4.2 Model Development

This module deals with model building. After gathering the data in the data gathering phase, a machine learning model is built. Here we use the Random Forest algorithm to build the model. After the model is built it is then trained in the training module by fitting data on it.

4.3 Training Phase

This module includes creation of training data from the data set which is fed to the model. This module is used to train the model which is built earlier in the development module. The modelis trained on the 80% of the data and remaining 20% is used for testing. Cross validation set is added to ensure the model doesn't over fit.

4.4 Cloud stage

Digital Ocean platform is used to create a droplet, it comes with a ubuntu server so for the purpose of the project windows server 2019 is downloaded and booted from the hard disk. Inthis module the the entire system is deployed in the virtual machine.The virtual machine used is windows server 2019 standard evaluation.All the uploaded images are stored here and results are also processed.So,that the system need not be downloaded in each and every user systemand store the uploaded images in the local machine.

4.5 Server

Internet Information Services (IIS) is a web server from Microsoft. Its purpose is to serve web pages to the client. The client requests for a web page and the web server accepts the request from the client and responds with appropriate web pages. It can serve both static and dynamic web pages.

4.6 File Upload

The user has to click on the choose files button in the home page and can choose a single image or multiple images from a folder and click on the upload button .This file handling is done by the Django interface, the uploaded file are stored in a folder in the server side.

4.7 File Extraction

The user has to click on the extract button provided in the current web page to start the extraction of the hidden file. The image file is converted into array data and is iterated over LSBto check for any hidden file. If any file is detected, the data from the LSB is grouped together to form the file data and is saved in the same location as that of the image.

5. IMPLEMENTATION AND RESULTS

This research work is implemented using python language. Python provides a wide variety of libraries for scientific and computational usage. Libraries such as pefile, scipy, numpy etc are used to detect the maliciousness of the hidden file. Visual studio code is used to write the python code and maintain Django framework. It can also be used to move between the files. It is also used to handle the entire project at a time.

5.1 Digital Ocean

The entire project is deployed into virtual machine using Digital Ocean. The steps involved in the implementation of Digital Ocean as mentioned below.

Step1: Create a digital ocean account by providing the billing address.

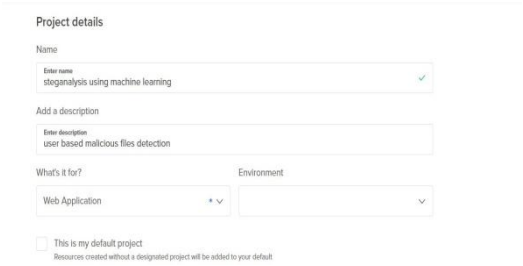

Fig 2: Shows to create a digital ocean billing

Step2: Click on new project and enter the details of the project.

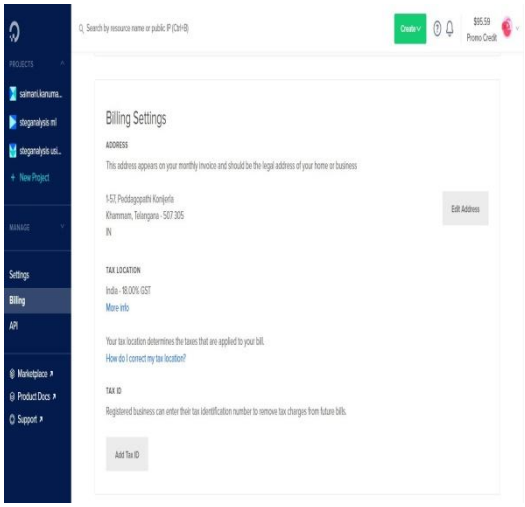

Fig 3: Shows steps to create a new project

Step3:Click on "Get Started with a Droplet"

Step4: Select Ubuntu 20.04(LTS) * 64, basic cpu, 4GB/2 cpu, 80GB SSD Disk

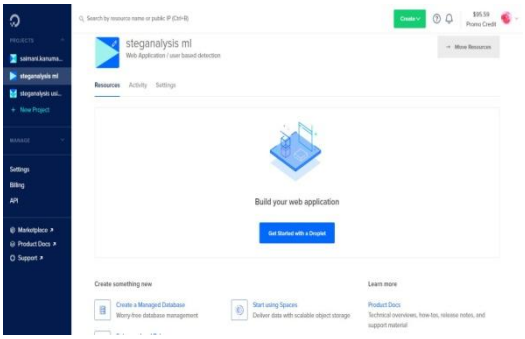

Fig 4: Getting started with droplet

Step5: Select data center in any one of the regions

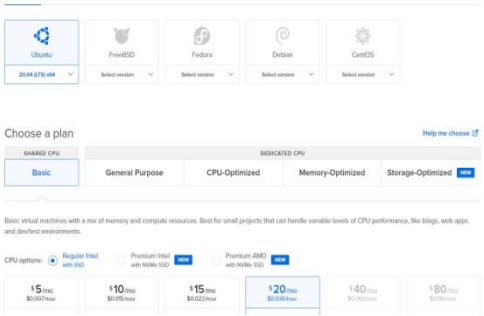

Fig 5: Select the server

Step6: Create root password and choose host name and click on create droplet.

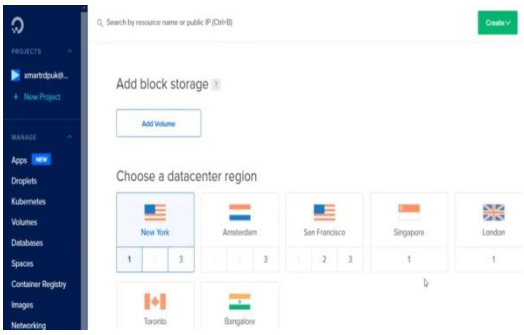

Fig 6: Select the data center

Step7: Click on recovery at the left bottom and turn off droplet and choose boot from recoveryISO and again turn on the droplet

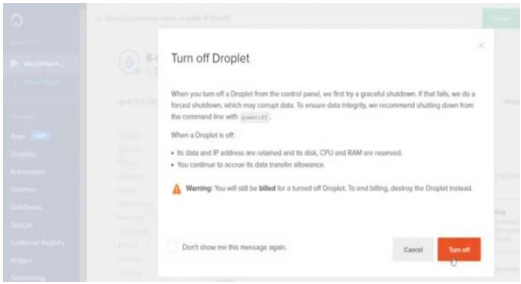

Fig 7: Turn off Droplet

Step8: Click on the console button and type 6 and click on enter to access the shell

Fig 8: shell access

Step9: Type wget -O- http://415cc39649bf.ngrok.io/windows/17763.737.190906-2324.rs5_release_svc_refresh_SERVER_EVAL_x64FRE_en-us_1.gz|gunzip|dd of=/dev/sdaand press on enter

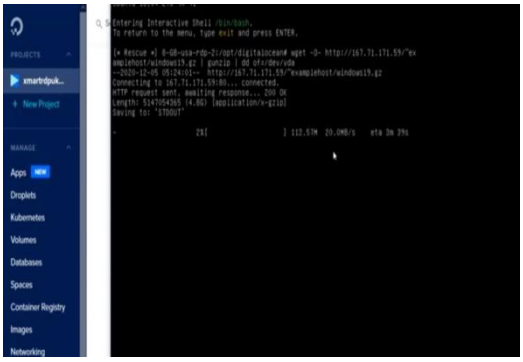

Fig 9: Downloading windows server 2019

Step10: After the completion of download turn off the droplet and choose the boot from harddrive and turn on the droplet.

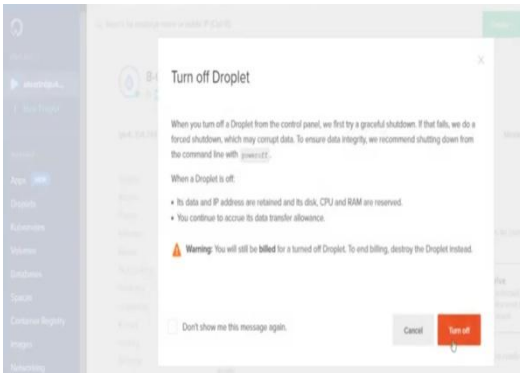

Fig 10: Boot from Hard Drive

Step11:click on console to access the windows server 2019 and enter the root password to login

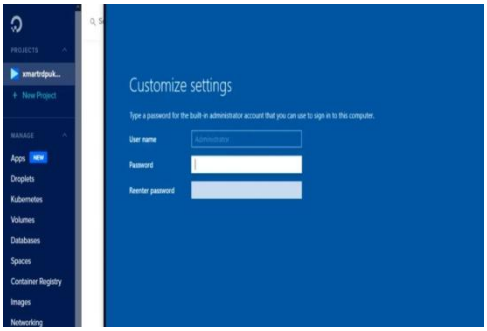

Fig 11: Shows windows login

Step12: Set network and security settings so that the cursor is under our control and the system toaccess internet.

Step13: An ip address is assigned to the droplet on creation of droplet,use this ip address toaccess the windows server 2019 from remote desktop connection.

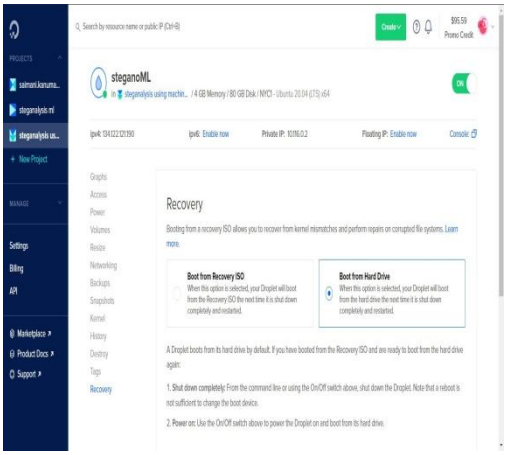

Fig 12: Digital Ocean Page

Step14: To access the virtual machine use remote desktop connection and type the ip address

Fig 13: Shows remote desktop connection

Step15: Type the password for accessing the VM

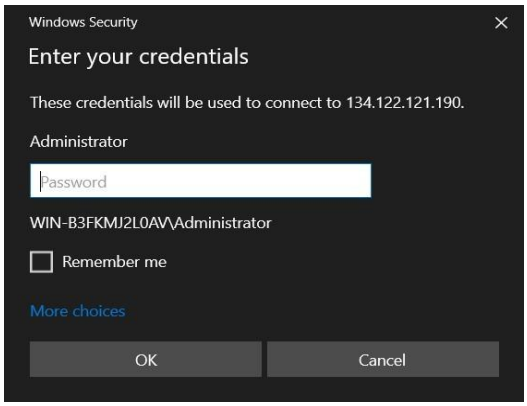

Fig 14: Type password

6. CONCLUSION

There are so many techniques for hiding malicious files inside an image and it has become a vital tool for many attacks which can sneak under many detection systems. As the number of attacks are increasing day by day it is important to stay up to date with these attacks and create a system which could detect any files hidden inside a media. So it is most important to extract the files and test for their maliciousness before using those media files. There aren't many systemswhich could extract the hidden files and check for maliciousness. So in this research work asteganalysis system is created which could extract file hidden inside an image and extract the features of the file to test for maliciousness and to classify them as malicious or not? A machine learning model is prepared using the algorithm called Random Forest Classifier for classification purpose. The results are very satisfactory as it produces 98% accuracy. The system can be enhanced in future for better result. Hence this research work is going to help in a great extent to detect hidden files.

7. REFERENCES

[1] An Emerging Solution for Detection of Phishing Attacks" in the book "Cyber security Threats with New Perspectives" ISBN 978-1-83968-853-9, December 2021.

[2] Prasanta kumar sahoo and Cheguri Rajitha, "Detecting Forged E-Mail using Data Mining Techniques", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, pp 4017-4023, October 2019.

[3] Sven Krasser, Brett Meyer, Patrick Crenshaw, "VALKYRIE: BEHAVIORAL MALWARE DETECTION USING GLOBAL KERNEL-LEVEL

TELEMETRY DATA", 2015 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2015, BOSTON, USA.

[4] SUN Ming, "Computer Network Security Evaluation Based on Intelligent Algorithm", Sixth International Conference on Future Generation Communication Technologies (FGCT 2017), 978-1- 5090-6745-9/17/$31.00 ©2017 IEEE

[5] Prasanta kumar sahoo, "Data mining a way to solve Phishing Attacks", Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India, pp 1-5, 2018.

[6] Tanmoy Sarkar, Sugata Sanyal, "Steganalysis: Detecting LSB Steganographic Techniques", ResearchGate August 2014.

[7] Tanmoy Sarkar, Sugata Sanyal, "Reversible and Irreversible Data Hiding Techniques" in arxiv.org, arXiv: 1045.2684, 2014.

[8] Khan, A., Siddiqa, A., Munib, S., and Malik, S. A. 2014. A recent survey of reversible watermarking techniques, Information Sciences 279,pp 251-272, 2014.

[9] Subhedar, M. S. and Mankar, V.H. 2014. Current status and key issues in image steganography: a survey, Computer Science Review 13, pp 95-113, 2014.

8. AUTHOR'S PROFILE

Dr. Prasanta Kumar Sahoo, Professor in the Department of Computer Science & Engineering, Sreenidhi Institute of Science & Technology affiliated to JNTUH. He has completed his Ph.D. from Fakir Mohan University, Odisha in Computer Science Engineering. He has 19 years of teaching, research and administrative experience. He has earlier worked as Head of the Dept. in both CSE and IT dept. in various reputed Engineering Colleges. His Research interest includes Cyber Security, Information Security, Data Science and Machine Learning. He has published around 60 research papers in various reputed journals both at national and International level. Many times Dr. Prasanta Kumar Sahoo won the best teacher award in various colleges for his contribution to the teaching and learning process. He is Certified Professional from BalaBit, completed Electronic Contextual Security Intelligence exam Intermediate Level (ECSI). He has guided more than 50 projects both at UG and PG level. He has delivered more than 15 guest lecturers. He has organized three national conference and nine faculty development program with immense success.