Detection And Classification of Malicious Software Using Machine Learning and Deep Learning

Mr. k. Balakrishna Maruthiram[1], Bushra Fatima[2]
[1]*Assistant Professor of CSE, Department of IT, University College of Engineering Science and Technology, JNTUH, Hyderabad, India*
[2]*Master student of CNIS, University College of Engineering Science and Technology Hyderabad JNTUH, Hyderabad, India*

***Abstract*—Malicious software is designed to intentionally disrupt computer systems. It can be analyzed using either static or dynamic techniques, which help to identify unique patterns essential for accurate malware detection. Over the past decade, numerous methods have been proposed to detect malware, often emphasizing threats that target networks. For example, DDoS attacks represent a prevalent risk in network security, overwhelming devices with excessive requests and thereby blocking legitimate access. Software vulnerabilities present another significant security concern, capable of compromising entire systems, stealing information, altering data, denying service, and damaging devices. This paper focuses on dynamic analysis to develop a malware detection system using machine learning techniques. It introduces a behaviorbased approach to malware detection. The methodology involves setting up a dynamic analysis environment and running malware samples through classification algorithms. Various behavioral indicators such as PSI.**

Index Terms*—API calls, registry modifications, and file operations are extracted and utilized within the malware detection system.

I. INTRODUCTION

Over the past decade, reliance on computer systems has surged dramatically, automating tasks from daily activities to business operations. Malicious software, or malware, poses a significant threat to these systems, with cybercriminals continually evolving their tactics to evade traditional security measures.

This highlights the critical need to identify and combat malware by analyzing distinct artifacts. Our report centers on utilizing dynamic analysis techniques to create a robust malware detection system. By scrutinizing software behavior in real-time, we can uncover distinctive patterns and signatures, facilitating precise identification of malicious software. Our approach integrates machine learning techniques to analyze the dynamic aspects of software execution, encompassing system calls, network traffic, and file interactions. Through training models on labeled datasets containing both benign and malicious software, our system acquires the capability to differentiate between normal and malicious behaviors, thereby elevating the accuracy of malware detection.

II. LITERATURESURVEY

This paper presents a novel approach to robust malware detection using Residual Attention Networks (RAN). Traditional methods for identifying malicious software often struggle with accuracy and adaptability due to the evolving nature of malware. To address these challenges, we leveraged the Residual Attention Network's ability to focus on pertinent features while maintaining high classification accuracy. Our approach involves training the RAN on a comprehensive dataset of malware and benign software samples, enabling it to learn and identify subtle differences between them. The experimental results demonstrate that the proposed model outperforms conventional machine learning and deep learning techniques, achieving superior detection rates and reduced false positives. This research highlights the potential of Residual Attention Networks in enhancing malware detection systems, providing a more resilient and accurate solution for cybersecurity applications. This paper explores the application of neural

networks and transfer learning for image-based malware classification. Traditional malware detection methods often fall short when addressing the sophisticated and evolving nature of modern threats. To overcome these limitations, we utilized convolutional neural networks (CNNs) alongside transfer learning techniques, leveraging pre-trained models to enhance classification accuracy. By converting malware binaries into grayscale images, we enabled the neural networks to analyze and classify the malware based on visual patterns. Our approach was validated on a comprehensive dataset, demonstrating that the combination of CNNs and transfer learning significantly improves detection rates and reduces false positives compared to conventional methods. The findings indicate that this technique offers a promising direction for robust and efficient malware detection, contributing to the advancement of cybersecurity measures.

III. PROPOSED SYSTEM

In our proposed system, we introduce novel advancements in the analysis of malware, departing from previous approaches by delving into string features at the word level through sophisticated text mining techniques. Our system's core involves the utilization of machine learning algorithms to scrutinize Malware API calls, a departure from conventional methodologies. Additionally, we introduce a unique aspect to malware detection by transforming malware files into image representations. These representations are then subjected to classification using a variety of algorithms, including

Support Vector Machine, Naïve Bayes, KNN, Random Forest, and Decision Tree. To facilitate training, we employ the APIMDS dataset for Malware API calls and the MALIMG Dataset for malware image representation. Through rigorous evaluation, we compare the performance of each algorithm in terms of accuracy scores, providing valuable insights into their effectiveness in identifying and mitigating malware threats. An intriguing aspect of your system is the transformation of malware files into image representations. This technique likely involves converting malware binaries or other data into visual representations (images). These images are then used as input for classification models. This method could potentially reveal patterns or features in malware that are not easily discernible in traditional feature sets. Each algorithm has its strengths and is suited for different types of data and classification tasks. By using multiple algorithms, you can compare their performance in terms of accuracy scores. This comparative evaluation helps in understanding which algorithms are most effective for the specific tasks of malware detection and classification. Overall, system appears to be at the forefront of integrating machine learning with malware detection, combining innovative feature engineering (API calls and image representations) with a diverse set of classification algorithms. This approach could potentially lead to more effective and robust malware detection systems compared to traditional methods.

IV. IMPLEMENTATION

Fig: 1 Home page

Creating a malware classification homepage framework that leverages machine learning (ML) and deep learning (DL) involves several steps. The homepage should be intuitive and user-friendly, providing clear navigation and easy access to various functionalities.

Fig :2 Admin Login

To create an admin login for a malware classification homepage, you need a secure authentication system. Log in with the correct user ID and password.

Fig:3 Selection of API Calls/image datasets

After login as admin the portal will have Malware API calls datasets and also Malware binary Image dataset. From this the selection of api calls and images are taking place as it is shown in the figure left side.

Fig:4 Uploading files

As shown in the above figure uplaoding the files as training and testing. Here in this only machine learning classification accuracies can be view. The classifiers are Support vector machine learning algorithm, Naïve bayes classifier, kNeighbors classifier, Artificial neural networks and Random forest classifier.

Fig :5 Training and Testing

Training and testing a malware detection dataset using machine learning (ML) algorithms involves.

Gather a comprehensive dataset that includes both malware and benign software samples. Extract relevant features from the dataset. These could include static features (e.g., file size, file type, strings) and dynamic features (e.g., API calls, network behavior). Divide the dataset into training and testing sets. A common split is 70-80% for training and 20-30% for testing. Train the chosen ML algorithms on the training dataset. During training, the model learns to differentiate between malware and benign samples by adjusting its parameters to minimize the error on the training set.

V. RESULTS

Fig: 6 Accuracies of ML algorithms

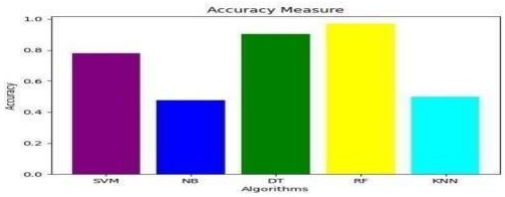

Fig :7 Graph of Accuracy

From the above figures The Random Forest algorithm still has the highest accuracy at 96%, followed by the Decision Tree at 90%. The SVM performs moderately well at 78%, while Naïve Bayes and K-Nearest Neighbors have lower accuracies at 47% and 49%, respectively.

Fig:8 Uploading files to view accuracy with DL Algorithm (CNN)

A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is particularly effective for image and pattern recognition tasks, including malware

detection. Training involves feeding labeled data into the CNN, allowing it to learn distinguishing features of malware.

Fig :9 Accuracy

Accuracy of 0.96: Indicates that 96% of the predictions made by the model are correct. This is a very high accuracy, showing that the model is reliable in identifying malware versus benign files. The loss of 0.04 is a very low loss, that the difference between the predicted outputs and the actual outputs is minimal. This means the model's predictions are very close to the actual labels, indicating good performance.

Fig :10 Malware detected in text based files

A heuristic (heiur) Trojan is a type of malware that can pose a significant threat to computer systems. It operates by disguising itself as legitimate software to trick users into executing it. Once activated, it can perform a variety of malicious actions, such as stealing personal information, corrupting files, or creating backdoors for other malware.

Fig:11 Malware detected in Image based file

Dialer's malware is a type of malicious software that primarily targets systems to connect them to premium rate telephone Numbers without the user's consent. A dialer malware is a significant threat, especially if it leads to unauthorized premium rate charges.

VI. CONCLUSION

In conclusion, this study demonstrates the efficacy of using machine learning and deep learning techniques to detect and classify malicious software. By analyzing text-based API requests and image files, we were able to distinguish between harmful and benign software with high accuracy. The Random Forest model achieved a remarkable accuracy rate of 96% for text-based malware detection, while the Convolutional Neural Network (CNN) model attained a 94% accuracy rate in identifying malicious images. These results underscore the potential of these models in enhancing cybersecurity measures.

REFERENCE

[1] Alaeiyan, Mohammadhadi, Saeed Parsa, and Mauro Conti. "Analysis and classification of context-based malware behavior." *Computer Communications* 136 (2019): 76-90.

[2] Yerima, Suleiman Y., Sakir Sezer, and Gavin McWilliams. "Analysis of Bayesian classification-based approaches for Android malware detection." *IET Information Security* 8.1 (2014): 25-36.

[3] Al-Janabi, Maryam, and Ahmad Mousa Altamimi. "A comparative analysis of machine learning techniques for classification and detection of malware." *2020 21st International Arab Conference on Information Technology (ACIT)*. IEEE, 2020.

[4] Sewak, Mohit, Sanjay K. Sahay, and Hemant Rathore. "Comparison of deep learning and the classical machine learning algorithm for the malware detection." *2018 19th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)*. IEEE, 2018.

[5] Rathore, Hemant, et al. "Malware detection using machine learning and deep learning." *Big Data Analytics: 6th International Conference, BDA 2018, Warangal, India, December 18–21, 2018, Proceedings 6*. Springer International Publishing, 2018.

[6] Bhodia, Niket, et al. "Transfer learning for image-based malware classification." *arXiv preprint arXiv:1903.11551* (2019).
[7] Pant, Dipendra, and Rabindra Bista. "Image-based malware classification using deep convolutional neural network and transfer learning." *Proceedings of the 3rd International Conference on Advanced Information Science and System*. 2021.
[8] Bidoki, Seyyed Mojtaba, Saeed Jalili, and Asghar Tajoddin. "PbMMD: A novel policy based multi-process malware detection." *Engineering Applications of Artificial Intelligence* 60 (2017): 57-70.
[9] Jeong, Ju Hyeon, Jong Hun Woo, and JungGoo Park. "Machine learning methodology for management of shipbuilding master data." *International Journal of Naval Architecture and Ocean Engineering* 12 (2020): 428-439.
[10] Hiran, Kamal Kant, et al. *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications, 2021.
[11] Bushby, Andrew. "How deception can change cyber security defences." *Computer Fraud & Security* 2019.1 (2019): 12-14.
[12] Tariang, Diangarti Bhalang, et al. "Malware classification through attention residual network based visualization." *2020 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*. IEEE, 2020.
[13] Shao, Yanli, et al. "Malicious code classification method based on deep residual network and hybrid attention mechanism for edge security." *Wireless Communications and Mobile Computing* 2022.1 (2022): 3301718.
[14] Cho, In Kyeom, et al. "Malware analysis and classification using sequence alignments." *Intelligent Automation & Soft Computing* 22.3 (2016): 371-377.
[15] Tariang, Diangarti Bhalang, et al. "Malware classification through attention residual network based visualization." *2020 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)*. IEEE, 2020.