Malware Classification Framework Based On Deep Learning

1st Abeera Biju
Computer Science and engineering
St. Joseph's College of Engineering
Palai, India
abeerabiju@gmail.com

2nd Aleena T James
Computer Science and engineering
St. Joseph's College of Engineering
Palai, India
aleenatj2501@gmail.com

3rd Diya Joji
Computer Science and engineering
St. Joseph's College of Engineering
Palai, India
diyajojixyx@gmail.com

4th Joanna Rachael Biju
Computer Science and engineering
St. Joseph's College of Engineering
Palai, India
joannabiju5050@gmail.com

5th Prof.Anu V Kottath
Computer Science and engineering
St. Joseph's College of Engineering
Palai, India
anuvkottath@sjcetpalai.ac.in

***Abstract*—Recent developments in technology and internet have sped up culprits' attention as they turn to virtual reality rather of the real world. The pitch of this change has grown after the Covid- 19 illness outbreak. culprits and other hackers have discovered that online crime is more simpler to commit than it's in the real world.Unwanted software known as vicious software(malware) is regularly used by online culprits to launch cyber attacks. Malware is a train or piece of law that can nearly perform any action an bushwhacker solicitations. It's frequently distributed via a network. Also, there are numerous ways to infect computers due to the wide variety of viruses.Advanced quilting and obfuscation styles are being used by malware strains to continue their elaboration. The discovery and categorization of these malwares have gotten harder as a result of the development of these concealing new tactics. To negotiate this, new approaches that are distinct from and different from the conventional ways must be devised.The current approaches to artificial intelligence(AI) and indeed machine literacy(ML) can not successfully address this issue. This concentrates our attention on the use of deep literacy(DL), a system that differs significantly from conventional machine literacy(ML) algorithms. This system's unique deep-literacy- grounded armature, which classifies malware variants using a mongrel model and the operation of the LSTM model, is proposed.The design of the deep neural network armature, the construction of the proposed deep neural network armature, training of the proposed deep neural network armature, and evaluation of the trained deep neural network comprise the four primary way of this armature. The suggested fashion was estimated using the Malevis, BIG 2015, and Malimg datasets. Using the Malimg dataset, the suggested system achieved 97.78 delicacy, outperforming the maturity of ML- grounded malware discovery styles.**

***Index Terms*—Malware,malware detection,deep neural network,malware variants,deep learning,**

I. Introduction

A person's life is made easier and more convenient by recent developments in computer and Internet technology. To launch cyber attacks, cyber criminals frequently utilise malware. Malware is any software that carries out unauthorised and suspicious actions on the computers of its victims.The different varieties of malware include viruses, worms, Trojan horses, rootkits, ransomware, etc. Malware variations have the ability to steal sensitive information, launch distributed denial of service (DDoS) assaults, and cause havoc to computer systems.In order to hide themselves in the victim's system, new malware varieties use techniques like encryption and packaging. Malware must be found as soon as it attacks the computer systems in order to defend them.The process of evaluating a suspicious file to determine its maliciousness or benign nature is known as malware detection. The classification of malware goes one step farther. Specifying the class or family of malware is known as malware classification and is done after the file has been classified as malware.Malware detection requires the following 3 steps:

1) Appropriate tools are used to examine malware files.
2) From the analysed files, static and dynamic features are extracted.
3) Features are organised in specific ways to distinguish malicious software from good software.

Starting to replace the inadequacies of current malware detection and classification technologies is a deep learning-based approach. Several fields, including image processing, computer vision, human action detection, driving safety, facial expression recognition, and natural language processing, have made substantial use of deep learning.Nevertheless, it hasn't been applied enough in the field of cyber security, particularly in malware detection. Artificial neural networks (ANNs) are the foundation of the artificial intelligence subfield known as deep learning. Deep learning learns from examples and employs numerous hidden layers.

II. Overview of Existing Malware Classification Systems

The objective of the system is to detect and classify the malwares present using deep learning methods efficiently.The purpose of malware analysis is usually to provide the

information we need to respond to a network intrusion. Our goals will typically be to determine exactly what happened, and to ensure that the system located all infected machines and files.By the use of Deep learning models of Resnet and Alexnet-50 it could potentially be able to achieve the objective of classifying malware.Here are some of the existing solutions present for malware classification.

A Comprehensive Review on Malware Detection Approaches[2]: The authors of this study offer a thorough analysis of the many methods for malware detection that have been put forth in the literature. The introduction of the paper discusses the many forms of malware and how they affect computer systems. The various methods for identifying malware are then covered, including signature-based detection, anomaly-based detection, and behavior-based detection. The limits of these methods and the difficulties in identifying malware are also covered in the study. The writers also discuss and evaluate the performance of the many malware detection systems and technologies that have been created. Overall, the report indicates topics for further research and offers a comprehensive review of the state of the art in malware detection.

Multi-scale Learning based Malware Variant Detection us- ing Spatial Pyramid Pooling Network[3]: In this paper, the author suggests a spatial pyramid pooling network-based multi-scale learning method for detecting malware variants. The proposed method is built on the notion of employing various feature scales to capture the various traits of malware strains. Using a dataset of malware and benign samples, the author assesses the suggested strategy and compares its performance to numerous cutting-edge methods. The outcomes demonstrate that the suggested method may identify malware variants with high accuracy. Overall, this study proposes a promising method for detecting malware variants that makes use of data at various dimensions. The performance of the suggested approach might be enhanced, and it could be used to combat various kinds of malware, through additional study.

A Dynamic Heuristic Method for Detecting Packed Malware Using Naive Bayes[4]: In this study, a dynamic heuristic technique and naive Bayes classifier are used to detect packed malware. Malware that has been compressed or encrypted in order to evade detection by security systems is referred to as packed malware. Malware that has been packed may not display the same characteristics as malware that has not been packed, making its detection difficult. The authors of this study suggest a technique that employs a dynamic heuristic approach to analyse the behaviour of the packed programme during execution in order to identify malware that has been compressed. Based on the behaviour seen, the packed software is classified by the approach as either malicious or benign using a naïve Bayes classifier. Using a dataset of packed malware, the authors assess the performance of the suggested technique.

A Survey on Malware Detection and Classification[5]: The authors present a survey of the various approaches proposed in the literature for detecting and classifying malware in this paper. The paper starts by going over the various types of malware and their effects on computer systems. It then goes over the various techniques for detecting and classifying malware that have been proposed, such as signature-based detection, anomaly-based detection, behavior-based detection, and machine learning-based approaches. The paper also discusses the limitations of these approaches as well as the difficulties associated with detecting and classifying malware. Furthermore, the authors review and compare the various tools and systems that have been developed for detecting and classifying malware. Overall, the paper provides an informative overview of the current state of malware detection technology.

Using a Subtractive Center Behavioral Model to Detect Malware[6]: This method models a program's behaviour as a series of events that take place as it runs. The software's most distinctive behaviours are discovered using the subtractive centre behavioural model, and these behaviours are then used to categorise the programme as dangerous or benign. The subtractive centre behavioural model is a machine learning-based method for spotting irregularities in a computer system's behaviour. The model operates by first learning the system's typical behaviour and then spotting deviations from it that might be caused by malware. The authors of this study assess the suggested method's performance against a dataset of malicious and benign programmes and contrast it with a number of state-of-the-art approaches.The outcomes demonstrate that the suggested approach can achieve high accuracy in finding malware. Overall, this study proposes a viable method for detecting malware that is based on the distinctive behaviour of programmes while they are being executed. Further study might be done to enhance the efficiency of the suggested technique and expand its applicability to various virus types.

Detection of Malicious Code Variants Based on Deep Learning[7]: One approach that has been shown to be effective for detecting malicious code is deep learning. Deep learning is similar as machine learning and it uses neural networks to find patterns. There have been several studies that have explored the use of deep learning for detect- ing malicious code variants. For example, in a study published in the journal Computer Science and Information Technology, the authors proposed a multi-scale learning approach for detecting malware variants using a spatial pyramid pooling network. The results of the study showed that the proposed approach was able to achieve high accuracy in detecting malware variants. In another study, published in the journal Computer Science Review, the authors pro- posed a hybrid malware classification method that combined segmentation-based fractal texture analysis with deep convolutional neural network features. The results of the study showed that the proposed method was able to achieve high accuracy in classifying mal- ware and benign samples.Overall, the use of deep learning for detecting malicious code variants has shown promising results in the literature. Further research could be done to improve the performance of these approaches and to

apply them to different types of malicious code.

How to Make Attention Mechanisms More Practical in Malware Classification[8]: Attention mechanisms have been widely used in natural language processing and image recognition tasks to improve the performance of machine learning models. In recent years, attention mechanisms have also been applied to the problem of malware classification. Malware classification is the task of identifying whether a given program or file is malicious or benign. It is a challenging problem because malware can take on various forms and can evolve over time, making it difficult to detect using traditional techniques such as signature-based detection. One of the main challenges in applying attention mechanisms to malware classification is the practicality of the approach. Attention mechanisms require large amounts of training data and can be computationally intensive, making them difficult to apply in real-world scenarios. To make attention mechanisms more practical in malware classification, several approaches have been proposed in the literature. For example, in a study published in the journal Computer Science Review, the authors proposed a hybrid malware classification method that combined segmentation-based fractal texture analysis with deep convolutional neural network features and an attention mechanism. The authors showed that the use of an attention mechanism improved the performance of the model on a dataset of malware and benign samples. Another approach that has been proposed to make attention mechanisms more practical in malware classification is the use of lightweight attention mechanisms. These mechanisms are designed to be more efficient and require less training data than traditional attention mechanisms.

Intelligent Vision-Based Malware Detection and Classification Using Deep Random Forest Paradigm[17]:This study project makes this endeavour and is discovered to be more successful in doing so. In this study, a vision-based analysis technique is used to examine binary file executables (also known as malware cleanware). To create new malware varieties, malware programmers recycle the same dangerous code portions. When we visualised the malware's binary code, this was simple to see. The Vision-based Analysis Method is the name of this method. Recently, malware visualisation has been employed as an effective alternate method for malware investigation. An exhaustive analysis is not necessary with this strategy.

A hybrid deep learning image-based analysis for effective malware detection[11]:The paper has two objectives: The first goal is to demonstrate how image-based methods can be used to identify systems that are acting suspiciously, and the second goal is to suggest and research the usage of hybrid image-based methods combined with deep learning architectures for an accurate malware classification. Using multiple malware behaviour pattern similarity metrics as well as cost-conscious deep learning architectures, performance is evaluated. Our suggested hybrid technique is tested for scalability using both publicly available and privately obtained big malware datasets, which demonstrate the excellent accuracy of our malware classifiers. In order to accurately identify and categorise obfuscated malware, this paper developed a new hybrid model for image-based analysis employing similarity mining and deep learning architectures.In order to deal with new malware in the future, the suggested cost-sensitive deep learning based model can be continuously taught in real-time. The accuracy of the SMO-Normalized Polynomial kernel used in the paper was approximately 99 percent, and the performance of our suggested cost-sensitive deep learning architectures was on par with some of the best designs previously described in the literature. In conclusion, we believe that our image-based methods have successfully distinguished between the behavioural patterns of several malware families.

Detection of Malicious Code Variants Based on Deep Learning [10]:In order to improve the identification of malware variants, this research developed a novel technique that made use of deep learning. Prior studies showed that deep learning performed quite well at image recognition. We turned the harmful code into grayscale photos to use with our suggested detection technique. A convolutional neural network (CNN) that could automatically extract the features of the malware images was then used to identify and categorise the photos. In order to overcome the data imbalance across various malware families, we also used a bat algorithm. We ran a number of tests on malware picture data from Vision Research Lab to validate our methodology. The testing findings showed that, in comparison to previous malware detection methods, our model obtained good accuracy and speed.

Intelligent Dynamic Malware Detection using Machine Learningin IP Reputation for Forensics Data Analytics[12]: IP reputation is anticipated in its pre-acceptance stage using the idea of big data forensics, and its associated zero-day assaults are characterised by behavioural analysis by using the Decision Tree technique. The suggested method calculates severity and risk score in addition to simultaneously evaluating confidence and lifetime. The number of false alarms when anticipating the type of zero-day attack and the condition of an IP is decreased without adding additional administration costs. The suggested study uses data forensics on the active cybercrimes to identify fresh assaults. We have also covered the problems and difficulties with big data forensics that emerged during the prediction of IP address reputation without taking big data forensics into account. The proposed system is assessed in two stages: first, the ML approaches are compared to get the best F-measure, precision, and recall scores; second, the overall reputation system is compared to the existing reputation systems. In addition to being cross-checked with other sources, our framework is able to lessen security concerns that were disregarded by out-of-date reputation engines that are now in use.

Classification of Malware Based on Integrated Static and Dynamic Features[13]: We introduce the first categorization approach that combines static and dynamic features into a single test in this article. Our method outperforms earlier

findings based on individual features and cuts the time required to examine such features separately in half. By contrasting the outcomes of two sets of malware, the first gathered between 2003 and 2007, and the second collected between 2009 and 2010, robustness to changes in malware production is assessed. Our combined test shows substantially greater robustness than the previous techniques when identifying the older set compared to the complete data set, losing accuracy of 2.7% as opposed to a decrease of 7%.We come to the conclusion that some earlier malware should be included in the set of data in order to attain acceptable accuracy in classifying the most recent malware.

MtNet: A Multi-Task Neural Network for Dynamic Malware Classification [15]:In this article, we suggest a brand-new multi-task, deep learning architecture for malware classification for the binary, i.e. malware versus benign malware classification task. Data gleaned from dynamic analysis of harmful and benign files is used to train all models. For the first time, we observe advancements in malware classification utilising a deep neural network design with numerous layers. The system is tested using a holdout test set of 2 million files after being trained on 4.5 million files, making this the largest study to date. The multi-task design combines the objective functions for the binary classification task and malware family classification task to produce a binary classification error rate of 0.358%. Also, we suggest a standard, single-task malware family classification architecture that also yields an error rate of 2.94%

DeepAM: a heterogeneous deep learning framework for intelligent malware detection[16]: In this research, we investigate how a deep learning architecture may be created for intelligent malware detection based on the Windows application programming interface calls retrieved from the portable executable files. To identify recently discovered malware, we provide a heterogeneous deep learning framework made up of an Auto Encoder stacked with multilayer limited Boltzmann machines and a layer of associative memory. The unsupervised feature learning operation used by the proposed deep learning model is greedy layer-wise training, which is followed by supervised parameter fine-tuning. We use both labelled and unlabeled file samples to pre-train many layers in the heterogeneous deep learning framework for feature learning, in contrast to previous efforts that only used files with class labels (either dangerous or benign) during the training phase.

CONCLUSION

Although there has been extensive study on malware detection and classification, the ability to accurately identify malware variants still poses a severe danger to cyber security.Malware identification is an extremely difficult operation due to code obfuscation and packaging tactics. The suggested solution demonstrates a cutting-edge deep learning architecture to accurately identify malware variants.Many thorough pretrained networks that use the transfer learning technique are included in this approach. The initial malware data collection was carried out using a number of extensive databases.The features are then retrieved using networks that have already been trained. The deep neural network architecture's training phase is then carried out using a supervised learning approach. For the Malimg, Microsoft BIG 2015, and Malevis datasets, the suggested deep learning approach is assessed. On a large size domain, it is seen that the suggested strategy is effective and decreases feature space. Since we train the data using an LSTM model, the findings we get are significantly more accurate.

REFERENCES

[1] Aslan, O., Yilmaz, A (2021) *A New Malware Classification Framework Based on Deep Learning Algorithms* ,IEEE Access, 9, 87936–87951.

[2] O. Aslan and R. Samet (2020) *A comprehensive review on malware detection approaches* ,IEEE Access, vol. 8, pp. 6249–6271.

[3] S, S., R, V., V, S., Alazab, M., KP (2020) *Multi-scale Learning based Malware Variant Detection using Spatial Pyramid Pooling Network*, IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops.

[4] Alkhateeb, E. M., Stamp, M (2019) *M. A Dynamic Heuristic Method for Detecting Packed Malware Using Naive Bayes* ,2019 International Conference on Electrical and Computing Technologies and Applications.

[5] Komatwar, R., Kokare (2020) *M. A Survey on Malware Detection and Classification.* ,Journal of Applied Security Research, 1–31.

[6]] Cui, Z., Xue, F., Cai, X., Cao, Y., Wang, G., Chen, J (2018) *Detection of Malicious Code Variants Based on Deep Learning* ,IEEE Transactions on Industrial Informatics, 14(7), 3187–3196.

[7] O. Aslan, R. Samet, and O. O. Tanrı over (2020) *"Using a subtractive center behavioral model to detect malware* , Secur. Commun. Netw., vol. 2020, pp. 1–17.

[8]] Ma, X., Guo, S., Li, H., Pan, Z., Qiu, J., Ding, Y., Chen, F (2019) *How to Make Attention Mechanisms More Practical in Malware Classification* ,IEEE Access, 7,155270–155280.

[9] D. Gibert (2016) *Convolutional neural networks for malware classification* , M.S. thesis, Univ. Rovira Virgili, Tarragona, Spain.

[10] Z. Cui, F. Xue, X. Cai, Y. Cao, G.-G. Wang, and J. Chen (2018) *""Detection of malicious code variants based on deep learning* ,IEEE Trans. Ind. Informat., vol. 14, no. 7,pp. 3187–3196.

[11] S. Venkatraman, M. Alazab, and R. Vinayakumar (2019) *A hybrid deep learning image based analysis for effective malware detection* ,J. Inf. Secur. Appl., vol. 47, pp.377–389.

[12] N. Usman, S. Usman, F. Khan, M. A. Jan, A. Sajid, M. Alazab, and P. Watters (2021) *"Intelligent dynamic malware detection using machine learning in IP reputation for forensics data analytics* ,Future Gener. Comput. Syst.,vol. 118, pp. 124–141.

[13] R. Islam, R. Tian, L. M. Batten, and S. Versteeg (2013) *Classification of malware based on integrated static and dynamic features* ,J. Netw. Comput. Appl.,vol. 36, no. 2, pp. 646–656.

[14] R. Gupta and S. P. Agarwal (2017) *A comparative study of cyber threats in emerging economies* , Globus Int. J. Manage. IT, vol. 8, no. 2, pp. 24-28.

[15] 88W. Huang and J. W. Stokes (2016) *MtNet: A multi-task neural network for dynamic malware classification* ,in Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment. Cham, Switzerland: Springer, pp. 399–418.

[16] Y. Ye, L. Chen, S. Hou, W. Hardy, and X. Li (2018) *DeepAM: A heterogeneous deep learning framework for intelligent malware detection* ,Knowl. Inf. Syst., vol. 54, no. 2, pp. 265–285.

[17] S. A. Roseline, S. Geetha, S. Kadry, and Y. Nam (2020) *Intelligent vision-based malware detection and classification using deep random forest paradigm* ,IEEE Access, vol.8, pp. 206303–206324y.