Proceedings of the Sixth International Conference on Trends in Electronics and Informatics (ICOEI 2022)
IEEE Xplore Part Number: CFP22J32-ART; ISBN: 978-1-6654-8328-5

A Study of Malware Datasets and Techniques to Detect the Malware using Deep Learning Approach

1st V. S. Jeyalakshmi
Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University,
Tirunelveli, India
vsjeyalakshmi@msuniv.ac.in

2nd J. Jayapriya
Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University,
Tirunelveli, India
jayapriya@msuniv.ac.in

3rd N. Krishnan
Centre for Information Technology and Engineering,
Manonmaniam Sundaranar University,
Tirunelveli, India
krishnan@msuniv.ac.in

2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-8328-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOEI53556.2022.9777220

***Abstract*—Cyber analytics play a vital role in solving the various domain problems in our day-to-day life. In this, Malware and web based attacks are most common types. Business organizations have their own apps to run their business. Malware captures the business information and corrupt the system. Malwares are developed based on financial gain. Security issues are now a big challenge with the ever increasing risk of malware attacks. Recently researchers are highly motivated to detect the malwares in the cyber field. Similarly some international highly trained programmer's community are also interested to detect the malwares for profit yielding purpose. The proposed study is based on the different malware datasets, deep learning techniques and its applications involved in malware analysis. The study infers the comprehensive comparison between different neural networks using Deep learning algorithm such as CNN, LSTM, RNN, GRU, GAN, Transfer learning, etc for Static malware analysis with different datasets.**

***Index Terms*—Malware Analysis, Static, Deep Learning Algorithms(DLA), Cyber Analytics**

I. Introduction

Over the years, the innovation of Malware has been increased in all sectors. Malware is a malicious software program code [51], that usually gives harmness to the entire network. Virus, keylogger, worms, Trojan horse, DDoS, Phishing, etc are the examples of malware. In 2021, it is estimated that every 11 seconds, ransomware attack occurs, cause a damage upto 20 billion dollars in business sector [1]. Around 2,00,000 malware samples are being detected every day. Microsoft Security Essentials is a free antivirus software that gives protection against different types of malicious software such as computer viruses, worm, spyware, rootkits and Trojan horses. It is used for performing manual scan of files and folders or an entire system for home use or small business purpose. Microsoft Security Essentials describe the security levels report to discover Malware. Low level indicates that the unwanted programs have some harmless functionality with malicious intentions. Medium level indicates the software that damages the user's privacy. High level indicates that the harmful programs that may misuse the unauthorized modification. Severe level indicates that the well known malware species [2]. If the computer is attacked by the virus, then it slow down slower than before, high abnormal CPU usage, browser popups fake updates. To counter these attacks, malware detection is necessary to safeguard the environment and protect the sensitive information [49] from the exposure of compromised data in the cyber world.

A. Motivation - Malware Analysis

Malware analysis is to identify the hidden malicious functionality in normal programs and to evaluate the characteristics of threat in a network. Figure 1 represents the types of Malware analysis. Malware classification, Malware detection and Malware Prediction are the three functionalities in Malware analysis. Malware classification is to classify the malware features according to their activity in the system [50]. Malware detection is to detect the malware's prevalence in the system. Malware prediction [51] is to predict the chances of getting attacked by any malware in the network. There are three forms of analysis - Static analysis, Dynamic analysis and Hybrid analysis which are used to extract the malwares. This paper gives a clear study about the Static malware analysis using Deep Learning Algorithms in Cyber Analytics.

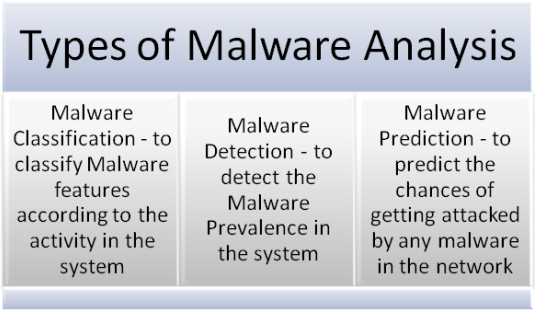

Fig. 1. Malware Analysis

The Contributions are, 1. This paper focuses, in understanding the dataset and techniques used for malware analysis from the contemporary works. 2. To understand the datasets and its features characteristics, how it collects from the repository. 3. Emphasis the deep learning algorithm that are used in recent research to detect the malwares for analysis purpose. 4. This illuminates the performance measure, accuracy of each model, drawbacks and efficacy in recent work.

Outline of this paper is as follows, the characterization of malware and its history introduced in Section I. Section II discusses the malware detection method in terms of dataset collection, preprocessing, modeling and feature extraction. Section III presents the related work of Static Malware Analysis using Deep Learning Algorithms and performance. Finally the conclusion is given in Section IV.

II. Background Study

A. Deep Learning Algorithm

Deep learning is a type of artificial intelligence, that imitates the human brain and the way humans gain certain types of knowledge. These algorithms are inspired by the structure and function of brain. The deep learning model makes data analyzing and data interpretation works very fast and easy. It is also used in the fields such as face recognition, natural language processing, speech recognition, image recognition, bioinformatics, malware feature extraction, etc. Some of the Deep learning algorithms used in this study are Artificial Neural Network (ANN), Deep Neural Network (DNN), Multi Layer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short Term Memory Network (LSTM), Gated Recurrent Unit (GRU), Recurrent Neural Network (RNN), Hidden Markov Model (HMM), K Nearest Neighbor (KNN), Support Vector Machine (SVM), Generative Adversarial Network (GAN), etc [16], [18].

B. Malware Datasets

Nowadays, malware samples are growing out in a vast manner. Security companies are eager to collect and analyze the extreme number of malware samples in the past decades. If the malware sample functionalities are known then the cyber attacks can easily be tackled. As a result, from different domains a security company can collect lot of files per day, stored and analyzed it by static, dynamic and hybrid analysis.

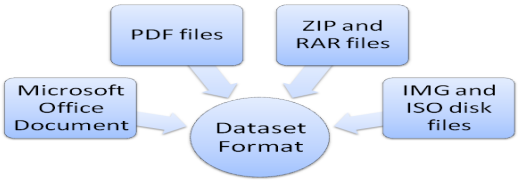

Fig. 2. Dataset Format

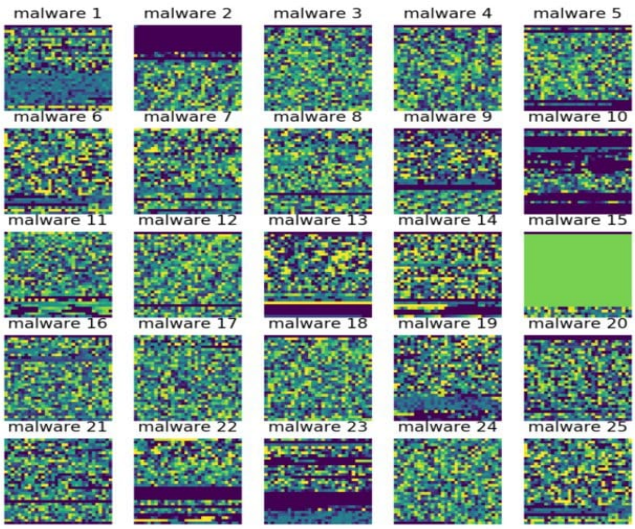

Fig. 3. Malware Image Format

Figure 2 represents different dataset format that can be used in malware analysis. The systematic dissection of the datasets are explained to know the real contents of the dataset [3], [8]. Intruders are easily hide the malware as one of the messages in any file. The malware file format may be of Zip and Rar files [49], Microsoft Office Documents, PDF files, IMG and ISO Disk files, etc [4].

Figure 3 represents the sample Malware image dataset format. In Malware samples collection, duplicates are removed by comparing MD5 hash of each file. To analyse malware samples, convert the binary file into a RGB malware image. Binary file is in zeros and ones, divide it into eight bits and map the upper and lower nibbles as a two dimensional colour map that gives the binary file's of each byte in to RGB pixel values [6]. The Malware image dataset is dissected by its textual similarity of the same as well as different malware classes. Here, the reuse of old malware codes to the new malware codes is not easily detected by signature matching techniques like obfuscation, packing or encryption but it is rectified in malware image dataset. So convert the binary file to image format and classify the malware.

C. Dataset Collection

The malware samples are collected from various malware repositories such as Malshare, Virus Share, Virus Total, etc [41]–[47]. Next, malware duplicates are removed in the collection by comparing MD5 hash of each file. Then the

malware samples are analyzed using Virus Total, which is not detected by Microsoft antivirus engine. Select the malicious one by Virus Total report and use the malware samples for classification using Deep Learning Model or convert it into image file. Then apply Deep Learning Algorithms for malware classification.

D. Dataset Preprocessing

Figure 4 represents the data preprocessing. The malware datasets are processed in four ways namely preprocessing, feature extraction, feature selection and feature detection [36]. In the data preprocessing step, the raw data can be changed to a suitable format to extract the features. Malware Predictive Analysis has four stages namely Data Exploration, Data Cleaning, Modeling and Performance Analysis. Data Exploration is the collection of dataset. Select the dataset, the recent issues and provides a solution for it to solve a big problem. Data Cleaning is to remove the noisy (redundant) data and changed it into consistent data (Preprocessing). In Data Model step, the deep learning or machine learning algorithms are applied in the dataset for feature extraction, feature selection and feature detection purposes. The processed features are given to the model in which the training and testing process can be done and the outputs are obtained. Performance Evaluation has done to evaluate the accuracy of the model. Finally, the intruder's attack or malware is detected [8]–[10], [16], [20], [40].

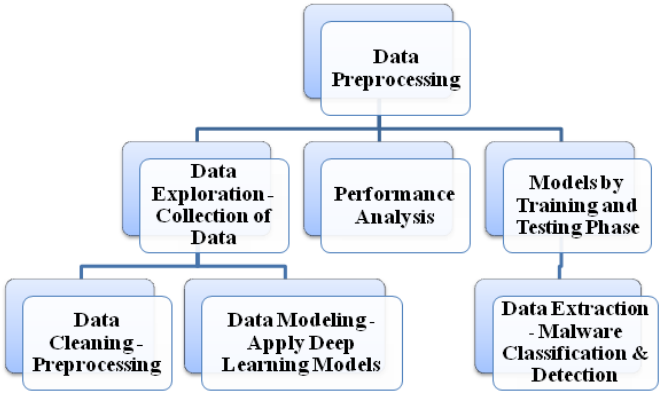

Fig. 4. Data Preprocessing

From the Background study, it is infered that different preprocessing tasks can be done in dataset to perform modeling and reporting.

III. Comparative Research of Deep Learning Algorithms for Static Malware Analysis

In this session, dataset collection and the application of Deep learning models in malware analysis are discussed. Table I indicates the information about some of the dataset and its collection released in a open data repository (some websites). Researchers use the public dataset from the simple data repository and perform their analysis work. Kaggle, Google Dataset Search, Microsoft research open data, University of California Irvine machine learning repository, etc are examples of dataset repositories. In the dataset, features are the elements of input vectors of the neural network for malware classification. Malicia dataset contains 11,688 binaries collected from 500 drive-by download servers. Driveby download are the chosen distribution vector for various malware families. Few features are used, dynamic features are not considered and more image features could not be tested is the drawback for this dataset. [16], [18]. Malimg dataset contains 9,339 malware files from 25 various malware families such as worm, Dialer, Backdoor, Trojan, Rogue, PWs, etc. The problem of this dataset is adversarial malware samples are not added, few hidden layers are used, obfuscation techniques are not used [14], [39].
Microsoft malware classification challenge was generated using the IDA disassemble tool. The raw data contains the hexadecimal representation of the file's binary content without the PE header. It consists of known malware files of 9 different families. Each malware file has an Id, 20 character hash value uniquely identifying the file, and a Class, an integer representing one of nine family names such as Ramnit, Lollipop, Vundo, Simda, Tracur, Gatak, etc. The weakness of this dataset is adversarial malware samples are not added to improve the system, few hidden layers are used, obfuscation techniques are not used [14], [40], [44].
IoT-23 contains 20 malware files captured in IoT devices and three harmless files captured in IoT device network traffic. The disadvantage of this dataset is simulation of the distance bound NFV patching model is not examined [46]. Derbin dataset contains 5,560 applications from 179 different malware families and collected from mobile sandbox project [45], malshare is a public repository gives free access service contains malicious samples, scanned by SandDroid [43] and AMD Android malware dataset contains 24,553 malware samples with 71 malware families.

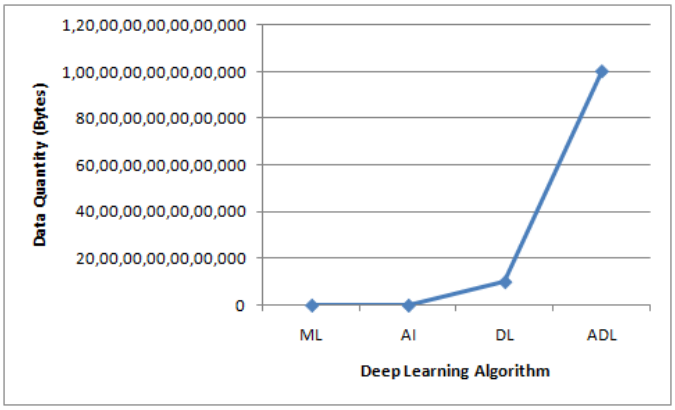

Fig. 5. Performance of DLA with Data Volume

Figure 5 shows the performance of deep learning algorithm when working with different volume of data. Nowadays, as there is a tremendous increase in the quantity of data in internet, there is a need for efficient learning algorithms. Machine Learning deals with megabytes of data, Artificial Intelligence works with gigabytes of data, Deep Learning is going with petabytes of data whereas Advanced Deep Learning

TABLE I
DATASET FOR MALWARE ANALYSIS

Sr. No.	Dataset	Description
1	Malicia dataset [41]	11,688 malware binaries were collected from 500 drive-by download servers.
2	Malimg dataset [42]	9,339 malware files belonging to 25 malware families and their variants.
3	Malshare [43]	Free malware repository to access the malicious feed samples.
4	Microsoft malware classification challenge [44]	Known malware files, representing a mix of 9 different malware families.
5	Derbin dataset [45]	It has 5,560 applications from 179 various malware families.
6	IoT-23 [46]	It has malicious and benign IoT network traffic.
7	AMD Android malware dataset [47]	It has 24,553 malware samples are integrated by 71 malware families.

Algorithm deals with millions of terabytes of data is infered from Figure 5.

Static analysis is to analyze the malware samples without running the program code. It is used to find the weakness in the source code that could lead to vulnerabilities [5]. There are a lot of automated tools that are much more effective to analyze the malware. Speed, accuracy and depth analysis are the advantages of static analysis. Signature based technique and Heuristic detection are the two methods used in static analysis. Signature based technique is used to evaluate the signatures in accordance with the known pattern matching. Heuristic method is not similar to signature based method, here malware detectors are used to identify the commands and instructions that are not present in the application program [6]. Machine learning and Deep Learning Algorithms are employed to train the features like extracting opcode sequences after disassembling the binary executable file. Then these models are used to predict the malwares with good accuracy [51].

The existing Methodologies in Machine Learning and Deep Learning Algorithms are Support Vector Machine, Decision Tree, K Means clustering, Multi Layer Perceptron, Convolutional Neural Network, Recurrent Neural Network, Long Short Term Memory, Gated Recurrent Unit, Generative Adversarial Network and Transfer Learning. SVM assigns data to various classes based on its features [48], [49]. Decision tree makes the subclasses of each leaf related to the probability distribution [48]. K means clustering groups a classes with similar features [48]. MLP is a back propogation algorithm which generates a set of inputs from a set of outputs. CNN has multilayers such as convolution, pooling and hidden layers neural network for processing structured array image data. RNN takes the last step is given as input to the current step. LSTM is a type of RNN achieves the time dependency [50]. GRU is the update version of LSTM, use connections through a sequence of nodes with memory. GRU achieves the result faster than RNN and LSTM. GAN contains generator and discriminator, generator creates fake data whereas the discriminator learns to identify the real data which is used in image, voice and video generation. Transfer learning need not spend time for training, the previous training task is recycled here to do a new task.

The following Table II explains the various algorithms that are used for malware detection. From the study, it is infered that various Deep Learning algorithms such as CNN, Random forest, Decision tree, Bi-LSTM, HMM, Word2vec, RNN, SVM and GRU are used to detect the malware. Amongst these algorithms, here the performance and accuracy result is very well. The study identifies some issues such as time complexity, low dimensional data usage and the no use of adversarial malware samples that should be solved.Usage of simple techniques, few features in particular only dynamic features and few hidden layers in the model are used and more expensive are the major drawbacks infered from Table II.

The outcoming Table III describes the various algorithms that are used for malware classification.From the Table III, it realizes that different Deep Learning algorithms such as CNN, One hot encoding, Word2vec, LSTM with attention, RNN, KNN, MLP and GRU are used for malware features classification or clustering. The performance and accuracy of the above algorithms are good. Increase of Time complexity, adversary attacks are not tested and use of low dimensional data are the major drawbacks that should be resolved. Obfuscated IoT malware samples are not used, More features like API call sequence, opcode are not added in the dataset are the drawbacks infered from Table III.

Table IV represents the various modern algorithms that are used in Deep learning for malware analysis. From the study, it is understood Transfer learning Deep learning Algorithm uses the pre-trained model to get more accuracy with less time and large data. Generative Adversarial Networks take the random noise as input and generate output. GAN algorithm confuse the model and check the probability result whether the model is strength or not. Both the results, are more accurate and effective model. Few hidden layers are used, more image features, parameters could not be tested, increase of time complexity and adversary samples are not used are the drawbacks infered from Table IV.

Figure 6 indicates working performance of various data with Deep Learning Algorithms such as Convolutional Neural Network, Recurrent Neural Network, Gated Recurrent Unit, Long Short Term Memory, Support Vector Machine, etc for malware feature classification. These algorithms are used in both image and text dataset. Each algorithm has specialized in text or image classification. Support Vector Machine, Machine Learning linear regression algorithm is mainly used for text classification. Convolutional Neural Network has kernels, used to filter all the relevant features of the image perfectly using

TABLE II
MALWARE DETECTION USING DEEP LEARNING ALGORITHM

Sr. No.	DLA	Dataset	Research Contribution	Drawback
1	CNN (LeNet-5, ResNet-101, ResNet-152, DenseNet-169, DenseNet-201)	Malimg dataset	To detect the malware within 5 seconds of execution [13]	-
2	HMM, Random Forest, CNN, LSTM (standard, Bagging, Boosting, Voting)	Real world dataset	To detect the malware with high precision [15]	Complex classes of stacking techniques are not used
3	Word2Vec + HMM	Malicia dataset	Performs better than any other approaches [16]	Few features used, dynamic features are not used
4	KNN, Random Forest, SVM, Decision Tree, Naïve Bayes, Feature selection –Information Gain	Real world HDLSS dataset (download - Virus Share)	The ML model with IS feature selection method works faster than IG method with accuracy of 99% [19]	-
5	ML-NB, RF, DT, SVM DL-CNN, LSTM, Ensemble of LSTM , CNN	Kaggle Competition	Performs better than all other algorithms with 98.46% accuracy [20]	Few features, DLM, parameters are considered
6	LSTM, CNN, BiRNN, CNN+LSTM, Stacked CNN	Sophos research and Alexa.com	It achieves around 93-98% malicious url detection rate with a false positive rate of 0.001 [21]	File paths, registry keys and network reputations are not used to detect Url
7	HMM, Word2Vec, NB,XGBoost,SVM	Sighan 2005 competition, Weibo Comment data	Identify the troll comments on Sina weibo platform with high accuracy of 89% [22]	-
8	Bi-LSTM, CNN(ResNet-50, RsNETXt-50, EfficientNet-B3)	KISA-datachallenge-2019	Dynamic and Static analysis with an accuracy of 94.98% and 94.46% [24]	Performance is very high, feature selection techniques are not used
9	LSTM+RNN, Bi-RNN	113 Linux malware executables from Virustotal	Achieves the performance of over 99% accuracy [25]	It is not trained by multiple malware infection
10	CNN, LSTM, GRU, Bi-LSTM, Bi-GRU, Stacked LSTM, Stacked GRU	Collect from Google Playstore	It provide a detection service less than 3 seconds on average, good accuracy and fast responsive [32]	Few features are used, high cost and Adversarial samples are not added
11	RNN-LSTM	BoT - IoT data set	The RNN-LSTM shows 93% accuracy [34]	Simulation of the distance bound NFV patching model is not tested
12	Tree Based Ensemble Models AdaBoost, XGBoost, Random forest, AUC-PRC Evaluation	Window Entropy Map image, API Binary/disassembly files	Model performance, time requirement is good [37]	-
13	RNN, CNN Sigmal- dynamic analysis, Hybrid Analysis – Window Static Brain Droid Model	Malimg dataset	Model detect the zero day malwares with good performance [39]	Adversarial malware samples are not added to improve the system

TABLE III
MALWARE CLASSIFICATION USING DEEP LEARNING ALGORITHM

Sr. No.	DLA	Dataset	Research Contribution	Drawback
1	IMG-CNN, STR LSTM	Mirai and Gafgyt	10,234 IOT malware samples are classified with 99.78% accuracy [11]	Obfuscated Iot malware samples are not used
2	CNN, LSTM	CICAndMal2017 (real world dataset by CIC Meter tool)	In binary category malware classification achieves 97.29% accuracy [12]	-
3	Multi Layer Perceptron, Word2Vec	Microsoft Malware Classification Challenge Big 2015	Model classify the samples with a high accuracy of 99.54% [35]	-
4	CNN, RNN	Kaggle dataset	Classify the features with an accuracy of 92% [36]	Feature extraction is not used
5	LSTM, GRU, 1DCNN	Microsoft Malware Classification Challenge	Model gives the low bias and an acceptable variance with good performance [40]	Adversarial samples are not added to improve the system
6	One hot encoding, HMM+LSTM	Real world data extracted by Open Source Tools	Model classify the malware with good accuracy [38]	-
7	Word2Vec, HMM2Vec and PCA2Vec	Collected samples in the dataset	Word embedding techniques generate features with good accuracy [29]	Simple, low dimensional data is used
8	LSTM with Attention	Malware sample	Model achieves the accuracy of 94.25% and f1 score of 0.95 [33]	-
9	CNN	Malimg dataset	The model achieves an accuracy of 97.537% [28]	Increase in hidden layers increase the time complexity
10	KNN, Random Forest	Real world dataset	K - means clustering technique is used to cluster seven families [17]	Opcode, API call sequence features is not used

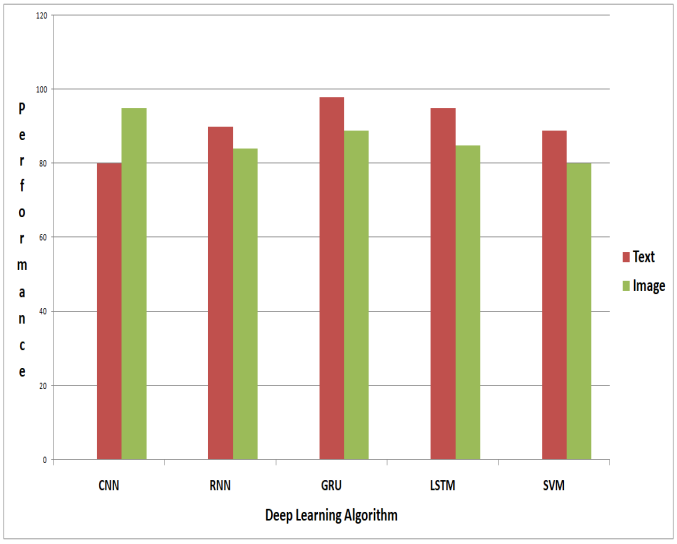

Fig. 6. Performance of Variant Data Using DLA

TABLE IV
MODERN DEEP LEARNING APPROACH

Sr. No.	DLA	Dataset	Research Contribution	Drawback
1	Alex-Net, ResNet-50, ReLU, Softmax, Transfer Learning	Malimg, Microsoft BIG 2015 and Malevis datasets	Features are more similar to misclassify the malware with 97.78% accuracy [14]	Adversary's attacks, tested three datasets with few hidden layers
2	MLP, CNN, GRU, RNN (BPTT), LSTM, Transfer Learning	Malicia dataset	Image based transfer learning models with an accuracy 92% [18]	More parameters , image features could not be tested
3	ANN, SVM, DNN, CNN, Transfer Learning	Dredze Dataset, Combined Dataset	Achieves 98.2% accuracy got in different combination of datasets [23]	Feature selection is not used. So it needs time to get result
4	Adversarial ML	Malware sample	To reduce the detection accuracy from 82% to 81.2% [26]	-
5	MLP, CNN, GAN	Public dataset USTC-TFC2016	Model detects the malware in a network traffic with a result [27]	-

convolution and pooling operation. CNN gives best classification in image dataset. Recurrent Neural Network are not possible to handle too long sequences and takes lot of time for training. As there is no memory for storage many features are not considered. This problem is overcomed by LSTM algorithm. Long Short Term Memory uses back propogation, avoid the gradient descent vanishing problem in classification and prediction based on time series data. Gated Recurrent Unit are a feed forward neural network has taken the output as input to the next step mainly used in time series data for classification to solve the gradient vanishing problem. GRU don't need memory units for processing and therefore training given to the model is easy. So it gives best performance than LSTM and RNN in text classification.

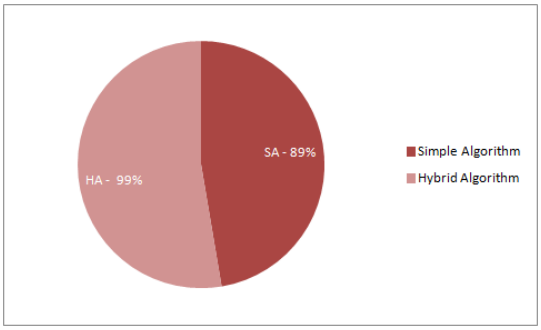

Fig. 7. DLA Accuracy

Figure 7 gives the performance accuracy of simple and hybrid Deep Learning Algorithms. Simple Deep learning algorithms classify the malware well but still hybrid algorithms are needed to give good feature extraction. Hybrid algorithm is the utilization of two or three specialized Deep learning algorithms or Machine learning algorithms to classify the malware features better than simple algorithm. When hybridizing two algorithms, the unique layers involved in each algorithm is utilized to achieve the efficient result.

RNN and GRU remembers the work more simple because of feeding the previous output as input in each step. LSTM is the updated version of RNN, to handle the problem with no time duration. So the art of LSTM, RNN and GRU gives proper classification with time complexity. CNN need not have spend time for feature extraction, simply compares the handcrafted features and classify the different features of image with better accuracy. GAN simulate the lensing for dark area. The art of LSTM, GAN and CNN performs well for image detection with time complexity. SVM is a linear seperable classifier used in classification of dog and cat images. The art of SVM and CNN gives a good image detection. HMM is used to observe the evolution of new events with their internal factors to detect the molecular sequence analysis of biological images. The art of HMM and SVM detect the molecular images well.

In Malware Analysis survey, researchers focuses in its architectures MLP, CNN, RNN, LSTM, ResNet,GAN, ELM, HMM2Vec, PCA2Vec and Word2Vec are selectively surveyed but still zero vulnerabilities are not discovered yet [30]. Researchers have explained all the malware functionalities, problems, classification and approaches like Characteristic based detection for multinomial classification [31].

From the above study infers MLP, CNN, LSTM with attention, RNN, GRU performs well in malware prediction, detection and classification. Hybrid algorithm gives good accuracy than single deep learning algorithm implementation in malware analysis. Transfer learning and Generative Adversarial Network functionalities are the recent deep learning algorithms which play a vital role for extracting the malware features in static analysis.

In malware recognition and classification field, Deep learning algorithms solve all the complex problems to be more effective. It is able to create transferable solution with the support of neurons leads to the most accurate final classification. Deep learning Algorithm works best with the vast amount of unstructured data in a powerful manner [31]. Deep learning algorithms are used to achieve less time complexity. From the whole, deep learning algorithms are more efficient than all the other algorithms in malware classification.

Deep Learning Algorithm enable the neurons and process them in each learning step one by one with their weights during their execution to get more prediction or classification result than any other traditional algorithms. Hybrid algorithm outperforms the art of machine learning and the existing deep learning techniques or the combination of existing deep learning techniques to rectify all the major problems of researchers to do malware prediction in the cyber security field. The study concludes that there is a need for efficient model in malware analysis

IV. CONCLUSION

Need of malware analysis is to understand the type and functionality of the malware such as how the system was infected with the malware, identifying the malware and how

it communicates with the attacker. Deep learning algorithms have lot of models that will easily classify the malware in a better way than any other techniques. The feedbacks are maintained so that the model predicts malware signatures rather than good signatures. This study presents the static malware analysis using various Deep learning algorithms. The main advantage of deep learning is using neural networks [10] like CNN, RNN, the best classifiers to learn the model and classify the malware as harmless or malicious with the fast and increase of accuracy of the model is obtained.

The drawbacks from the study are the usage of only few features, parmeters and hidden layers in the model. Sometimes feature selection is not enforced in the model. More expensive and the increase of time complexity. Only dynamic features are implemented in the model. Adversarial malware samples, obfuscated IoT samples are not included in the model. As there is an increase in the volume of dataset, new hybrid algorithm models are needed to protect data from malware.

REFERENCES

[1] Dynamic Malware Analysis in the Modern Era - A State of the Art Survey, Orior-Meir, Nir Nissim, Yuval elovici and Lior rokach. https://doi.org/10.1145/3329786.
[2] Review of the Malware Categorization in the Era of Changing Cyberthreats Landscape: Common Approaches, Challenges and Future Needs, Andrii Shalaginov, Geir Olav Dyrkolbotn.
[3] The Android Malware Static Analysis: Techniques, Limitations and Open Challenges, Khaled Bakour, H. Murat Unver, Razan Ghanem.
[4] Available:[https://www.perallis.com/blog/what-are-the-four-most-dangerous-file-types].
[5] Available:[https://www.malwarefox.com/malware -analysis-tools].
[6] Available:[https://security.cse.iitk.ac.in/sites/default/files/15111005.pdf].
[7] Malware Detection with Sequence-Based Machine Learning and Deep Learning, William B. Andreopoulos.
[8] A Close Look at a Daily Dataset of Malware Samples, Xabier Ugarte-Pedrero, Mariano Graziano, Davide Balzarotti.
[9] Available:[www.crowdstrike.com/cybersecurity-101/malware/malware-analysis/]
[10] Available:[http://sersc.org/journals/index.php/IJAST/article/view/27206]
[11] A Multidimensional Deep Learning Framework for IOT Malware classification and Family Attribution, Mirabelle Dib-2021.
[12] Android Malware Detection and Classification based on Network Traffic Using Deep Learning, Mahshid Gohari, Sattar Hashemi, Lida Abdi – 2021.
[13] Deep Learning Techniques fo Behavioral Malware Analysis in Cloud IaaS, Andrew Mc Dole, Maanak Gupta, Mahmoud Abdelsalam, Sudip Mittal and Mamoun Alazab -2021.
[14] A New Malware Classification Framework Based on Deep Learning Algorithms, Omer Aslan and Abdullah Asim Yilmaz.
[15] On Ensemble Learning, Mark Stamp, Aniket Chandak, Gavin wong and Allen Ye -2021.
[16] Word Embedding Techniques for Malware Evolution Detection, Sunhera Paul and Mark Stamp -2021.
[17] Cluster Analysis of Malware Family Relationships, Samanvitha Basole and Markstamp -2021.
[18] An Empirical Analysis of Image based Learning Techniques for Malware Classification, Prattikkumar Prajapati and Mark stamp -2021.
[19] Fast and Straight forward Feature Selection Method-A case of High Dimensional Low sample size dataset in Malware Analysis, Sergii Banin -2021.
[20] Abusive Comments using Ensemble Deep Learning algorithms, R. Ahiya, A. Banga, S C Sharma -2021.
[21] DURLD:Malicious URL Detection using Deep Learning-Based Character Level Representations, Sriram srinivasan, R. Vinayakumar, Ajay Arunachalam, Mamoun Alazab and K P Soman -2021 .
[22] Sentiment Analysis for Troll detection on Weibo, Zidong Jiang, Fabio Di Troia, Mark Stamp -2021.
[23] Image Spam Classification with Deep Neural Networks, Ajay Pal singh and Katerina Potika -2021.
[24] Two stage Hybrid Malware Detection using Deep Learning, Seungyeon Back, Jueun Jeon, ByeonghuiJeong and Young-Sik Jeong -2021.
[25] Neural Networks Based Online Behavioural Malware Detection Technique for Cloud Infrastructure, Jeffrey C. Kimmel, Andrew D. MCdole, Mahmoud Abdelsalam -2021.
[26] HMD-Hardener: Adversarially Robust and Efficient Hardware-Assisted Runtime Malware Detection, Abhijitt Dhavlle, Sanket Shukla, Setareh Refatrad, Houman Homayoun, Sai Manoj Pudukottai Diakarrao-2021.
[27] Efficient Malware Originated Traffic Classification by Using Generative Adversarial Networks, Zhicheng Liu, Shuhao Li, Yongzheng Zhang, Xiaochun Yun, Zhenyu Cheng -2020.
[28] Deep Learning in Malware Identification and Classification, Balram Yadav and Sanjiv Tokekar -2021.
[29] Comparison of Word2Vec, HMM2Vec and PCA2Vec for Malware Classification, Aniket Chandak, Wendey Lee and Mark Stamp -2021.
[30] A Selective Survey of Deep Learning Techniques and Their Application to Malware Analysis, Mark Stamp -2021.
[31] A Performance-Sensitive Malware Detection System Using Deep Learning on Mobile Devices, Ruitao Feng, Sen Chen, Xiaofei Xie, Guozhu Meng, Shang-Wei Lin and Yang Liu.
[32] Malware Family Classification using LSTM with Attention, Qi Xie, Yongjun Wang, Zhiquan Qin.
[33] A network function virtualization system for detecting malware in large IoT based networks, Nadra Guizani, Arif Ghafoor.
[34] Malware Classification Based on Multilayer Perceptron and Word2Vec for IoT Security, Yanchen Qjao,Weizhe Zhang, Xiaojiang Du, Mohsen Guizani -2021.
[35] Malware Classifier for dynamic deep learning algorithm, Young-bok Cho - 2021.
[36] Comparative analysis of Low Dimensional features and Tree based Ensembles for Malware detection System, Seoungyul Euh, Hyunjong Lee, Donghoonkim and Doosung Hwang 2020.
[37] Detection with sequence based Machine Learning and Deep Learning, William B. Andreopoulos -2020.
[38] Intelligent Malware detection using deep learning, Vinayakumar R, Mamoun Alazab, Soman KP, Prabaharan Poornachandran and Sitalakshmi Venkatraman -2019.
[39] Benchmarking Convolutional and Recurrent Neural Networks for Malware Classification, Haidar Safa, Mohamed Nassar, Wael Al Rahal Al Orabi -2018.
[40] Available:[http://www.malicia-project.com/ dataset.html].
[41] Available:[https://github.com/Hydrogenion /Malimg].
[42] Available:[https://www.malshare.com/ index. php].
[43] Available:[https://github.com/naritapandhe/ Microsoft Malware Classification Challenge readme].
[44] Available: [https://paperswithcode.com/dataset /drebin-1].
[45] Available: [https://www.stratosphereips.org/ datasets-iot23].
[46] Available:[https://www.impactcybertrust.org/ datasetview?idDataset=1275].
[47] Sathesh, A. "Enhanced soft computing approaches for intrusion detection schemes in social media networks." journal of soft computing Paradigm(JSCP) 1, no. 02 (2019):69-79
[48] Vivekanandam, B. "Design and Adaptive hybrid approach for genetic algorithm to detect effective malware detection in android division" Journal of ubiquitous computing and communication technologies 3, no. 2 (2021):135-149.
[49] Soni, Jayesh, Suresh K. Peddoju, Nagarajan Prabhakar and Himansu Upadhyay."Comparative analysis of LSTM, one-class SVM, and PCA to monitor Real-Time Malware Threats Using System Call Sequences and virtual machine introspection." In International Conference on Communication, Computing and Electronics Systems : Proceedings of ICCCES 2020, vol. 733, pp. 113. Springer Nature, 2021.
[50] Kumar, Ashwin A., G.P. Anoosh, M. S. Abhishek and C. Shraddha. "An effective machine learning based file malware detection- A survey." In International Conference on Communication, Computing and Electronics Systems, pp. 355-360. Springer, Singapore, 2020.
[51] Agarwal, Prerna, and Bhusan Trivedi. "AndroHealthCheck: A Malware Detection System for Android Using Machine Learning. "In Computer Networks, Big Data and IoT, pp. 35-41. Springer, Singapore, 2021.