Synopsis on

Developing a Systematic Machine Learning Approach to Generate Malware Detection And Classification Technique.

July-2020

Submitted for registration in the degree of

Doctor of Philosophy

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CHITKARA UNIVERSITY

HIMACHAL PRADESH

Submitted by

Gaurav Mehta

PHDENG18056

UNDER THE JOINT SUPERVISION OF

Supervisor

Dr Prasenjit Das
Associate Professor
Department of Computer Application
Chitkara University, Himachal Pradesh

Co-Supervisor

Dr Vikas Tripathi
Associate Professor
Department of Computer Science & Engineering
Graphic Era University, Dehradun

Table of Contents

List of Figuers

List of Tables

List of Abbreviations

PE…….....	Portable Executable files
CB-MMIDE….....	Consortium Blockchain for Malware Detection and Evidence Extraction
PUDROID……...	and Unlabeled learning-based malware detection for Android
MAS…….....	Sequence Alignment Algorithm
PCA…….	Principal Components Analysis
DBN…….....	Deep Belief Network
PSO……	Particle Swarm Optimization
ROC…….....	Receiver Operating Characteristic
CFG…….	Control Flow Graph
WFCM-AANN.....	Weighted Fuzzy K-means Clustering Algo-rithm with Auto Associative Neural-Network
GLMC…….	Gray Level Co-occurence Matrix
HMM……	Hidden Markov Models

1. Introduction

Threat of.malicious code or malware is increasing at high rate due to growth of internet and open source platform like android. Currently scope of malware is mot only restricted to machines(desktop or laptops or mobile phones) but its existence can also be seen in IoT and cloud. The growth of IoT devices and cloud architecture had given big platform to malware detector to proceed for security breach and get personal information without the knowledge of host.[1]

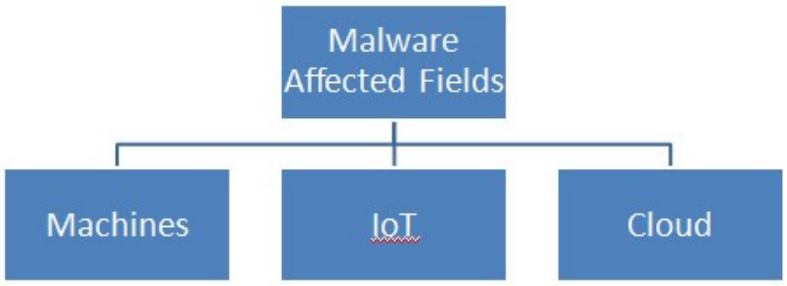

Fig 1 : Malware attack in different fiels.

Android is most widely used operating system in mobile all over the world, occupying total market share of 82.8% [2]. Millions of apps are available on Google play store and with millions of download count of thousands of apps which shows the polularity of android platform all over the worls. In comparision to iOS platform android also allows the users to download apps from unsecured or unverified platforms which also adds in the easy use of platform on user point of view. The huge amount of amdroid devices aloows the attackers to target this platform and 97% of attackers have the target field i.e android. [3].

Each type of attacking malicious code had 50 different varients that makes it more difficult to be identified by malware detection community. [4]. Malicious apps like Trojan are created

for different types-of-attacks, worms, exploits and viruses. Different approaches for static and dynamic analyses are used by researche to address security-concerns.

Malware detection is never ending process it's a never ending chase between malware detector and malware creator.[5]. Malicious code is not an emerging or new trend, it's from ages, since the start of computer machine. Large numbers of malware gets introduced in one or other fields which increase the demand to detect malware in all the areas that are attacked by malware. There exists large number of attacks which affects the host machine or data or security settings in one or the other way. Following table discuss some of the attacks and their types.

Table 1 Types of attack and description

Categories of attack	Description	Types of attack	Description of attack
Bot	Botnet is collection of connected devices through internet	Zombies	Daemon computer(under control of malicious hacker)placed in network to capture information.
		Spamming	Repeatedly sending messages to website
		Web injection	Changes the content of web page at client side by adding malicious code in browser
		URL spoofing	Changing address of site in address bar.
		DNS spoofing	Online traffic gets redirected by altered DNS records.
Dos / DDos	Single / Multiple computer(s) to flood target machine on	DoS GoldenEye	Denial of service attck kind
		DoS Hulk	Attack web server with huge amount of obfuscated traffic.

	network	DoS Slow httptest	Denial-of-service-(DoS) attacks in this attacker slowly sends HTTP requests to wer server due to which server resources remain busy for long time.
		DoS slowloris	Type of Denial-of-service-(DoS) attacks in which one machine take down another web server machine with minimum bandwidth consumption
		Heartbleed	Leaks the infprmation stored in memory to hacker by getting access to private encription key.
Infiltration	Unauthorized access of host machine to steal information		
PortScan	Identify weak points of system to attack based of varying destination port - Open, Closed, or Filtered.		
Web Attack	Malicious attack on web server machines on applications, Database, Operating system or in the network	Brute Force	Attempt to crack password or username or encryption key.
		Sql Injection	Retriving hidden data by modifying SQL query like UNION attack that retrives data from different databases.
		XSS	Cross site scripting is attack vector that inject malicious code in web application.
Brute Force	Illegal "Black Hat" attempt to obtain password or PIN.	FTP Patator	Password dictionary is accessed, thousands of username and password
		SSH-Patator	

			combination is tried on server.
Ransomware	Malware application that demands money to unlock owners data	GoldenEye	Combination of Petya and MISCHA ransomware.Spreads through massive social engineering campaign which launches macro that encrypt few or many files on victim's compute
		WannaCry	Most widely known ransomware also known as WCry or WanaCrypt0r
		Bad Rabbit	Spread through fake adobe flash updates on compromised websites
		Cerber	Target cloud based offices 365 users. ransomware as service - RaaS model is used.
		Crysis	Encryt files on fixed drives, removable drives and network drives, Spreads through malicious attachments in email hvinf twofold record extension.
		CryptoWall	Advanced form of CryptoLocker came in existence in 2014 with varients like CryptoBit, CryptoWall and Crypto - Defense.
		Jigsaw	Most ruinous kinds of ransomware which encodes and logically erases the scrambled documents until a payment is paid

		Locky	Designed to lock victim machine to prevent using it until ransom is paid.. Multiple files gets encrypted using AES encryption.
Worm	Malicious computer program that copies itself in entire network and damage all machines.		
Trojan Horse	A trap to user that install malware in computer to extract all information.	Remote – Access - Trojans(RAT's)	Backdoor malware to get administrator control
		Data – Sending - Trojans	Retrives senstive information(Password, emails, payment cards etc) and sends to malware qwner.
		Destruc-tive Trojans	Undetected(by antivirus) malicius program that deletes all files from machine.
		Proxy - Trojans	Victim computer behave like proxy server and hacker can perform any opetation on behalf of victim.
		FTP Trojans	Opens the port number 21 and activate FTP to transfer data.
		Security software disabler Trojans	Stops all security applications(anti virus,firewall)
		Denial-of-service attack (DoS) Trojans	Flooding technique is used to damage network with useless traffic.
Spyware	infiltrates device to steal internet usage data, senstive information and system details.		

Rootkits	Take privelaged access of machine by hiding its presence.		

1.1. Malware detection approach

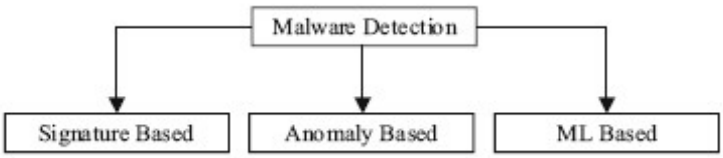

Fig 2 Malware detection approach.[6]

- **Signature based malware detection:** Database of known malwares is updated by malware detector when any new malicious code is identified. The signature of malicious code is added in this database to refer for malware detection. The new identifier is established for known threats to be identified in future. Signature based technique have two major disadvantage: firstly: malware detection product/tool need to look into big database to identify attack, secondly: newly developed malware cant be detect by this approach.[7]

- **Anomaly based malware detec-tion:** is used to detect malicious activity both in network and computer. Its a process to detect malicious activity by comparing defination of code. The classification of malicious code in anomaly based detection is as per heuristic or rules based rather than signature or pattern. Major disadvantage of this technique is that littele deviation from normal traffic or pattern gives alarm to security administrator to check and validate accordingly.

- **Machine learning based malware detection:** 0A data analytics tool to effectively perform specific task. Machine learning technique to detect malicious code is followed by many researchers. The power of machine learning tools helps to differentiate malware from bening by using different classification and clustering algorithm.

Table 2 : ML learning approach

Supervised learning (classification)	**Unsupervised learning (clustering)**
--Learning the model where input variable (say, x) and an output variable (say, Y) and algorithm that map the input with output. $Y = f(X)$ --Aim is to approximate the mapping function that: when there is new input say (x) then corresponding output variable say (y)can be predicted.	--Only the input data (say, X) is present and no corresponding output variable is there. --Main aim of Unsuper-vised lear-ning is to model distribution in data so as to order it to learn more about the data.

2. Literature Review

Lichao et al. [2017]: New malicious application is create every 4 second which increases the number of such apps on different download platforms like playstore etc. which makes it difficult to differentite between benign and malicious app. The task of analyst becomes more critical as the probability of downloading malicious app becomes too high which may leads to biased results for any framework to detect malware. PUDROID (Positive and Unlabeled learning-based malware detection for Android) is proposed to remove contaminants, [8].

Ding et al. [2017]: All the malware family have common behavior that makes them different from bemimg apps. Common behavior graph is created to show the behavior of malware as dependency graph. Taint tags of system call were marked with technique known as dynamic taint analysis. By tracking the propagation of taint data, dependency graph was created on the basis of which algorithm is proposed to create common graph. Finally the code is categorized as malicious as per the maximum weight of graph, [9].

Pektaş et al. [2017]: Building and maintaining effective security directly depends on classification techniques of malware. Model to classify malware in scaleable and distributed environment is proposed which is validted on 17900 malign codes which gives the accuracy upto 94%, [10].

Mirza et al. [2017]: Large amount of resources of host machine gets waisted during the process of malware detection. Author had used bespoke feure delection tool to apply ML technique on rich feature extracted. A cloud based architecture CloudIntell was proposed to detect malwre effectively. Relevant fetures gets extracted by removing obfuscated part with the proposed feature selection tool, [11].

Jingjing et al. [2017]: Blockchain technologie is used to detect a mobile based android malware for which framework CB-MMIDE(‘Consortium Blockchain for Malware Detection and Evidence Extraction’) was proposed. Consortium chain by test members is compared with public chain by users in the consortium blockchain framework. Two features i.e permission information and signature are important features to be considered for malware detection, [12].

Kim et al. [2017]: New malware gets introduced and that too in large number which makes the malware detection process to be more effective. Things gets more critical when malware creators wraps the malware with techniques such as anti-emulation, packing, anti-virtualization, obfuscation etc. Behavioral sequence chain is generated to collect malware followed by the process of clustering, pre-processing to crete input sequence of MAS('sequence alignment algorithm') which generate behavi-oural sequence chain of malware, [13].

Chowdhury et al. [2017]: Its not only the individuals that gets affected by serious cyber threat, malware also affects different fields like businesses, national intelligence, research organizations. Malware can breach, damage or modify crucial data.Need of an hour is to build more effective detection technique both for signature - based and anomaly - based with the help of machine learning and data-mining techniques. Author had used Principal Compo-nents Analysis (PCA) to delect features.The PCA has important feature of dimensionality reduction to enhance the computational speed. An ensembling of the N - Gram and API-call features increase the effectiveness of malware detection, [14].

Yuxin et al. [2017]: Author claims that DBN-(Deep Belief Network) perform better as compared to support vector machines, decision trees, and the k-nearest neighbor classifier algorithm. The machine language-opcode describes the behaviour of code/program.The

opcode n-gram is used to describe the behavioural feature of malware as malware is represented as sequence of opcode. The model consist of PE parser, feature-extractor and detection module for malware. [15]

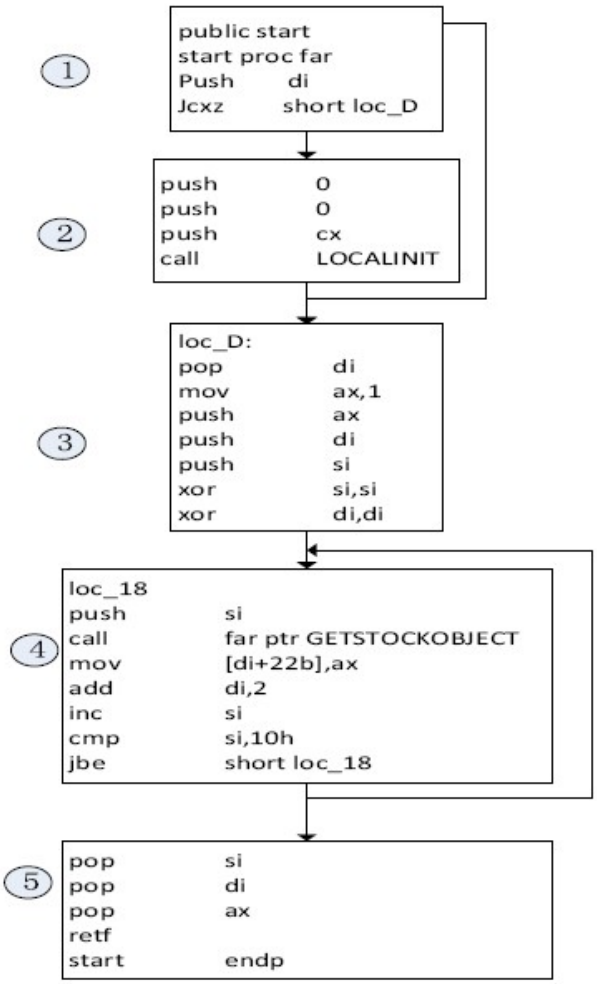

Fig. 3 Windows program-control flow graph [15]

Gamal et al. [2017]: Machine learning is approach to signatureless malicious code detection as it can gernalize to never before seen malware family. Obfuscations makes it difficult for malware detector to identify malicious code or bening applications. Things get more critical for metamorphic malware as its difficult to detect by regular string-signature. Author had proposed novel approach of behavioor and signature technique to improve detection of

metamorphic malware. A hybrid framework by combining string – based and behavior – based technique has been proposed,[16].

Table 3 : Malware detection techniques [17]

Detection-technique	**Definition and nature**	**Contribution**	**Limitations**
Signature – based detection	Most widely used anti virus technique is detection on the basis of signature. Specific malware is detected based on sequence of bytes.	This technique is fast and more effective against common type of malware in existence..	1 Malware not present in database will not be detected so anti malware team needs to update database frequently. 2 Signature based detection can be easily bleached with obfuscation techniques 3 Require time and space to maintain a repository of signatures of known malware. 4 As new threats are discovered daily so the repository is to be updated frequently.
Behavior – based detection	This approach monitors the malware action during their execution.	Behavior study of malicious code.	-During training phase malware and bening both are analyzed. -Classification is done at execution phase.
Specification – based - detection	Properties got from program includes as in the HMM, utilized in their exploration paper, are utilized toclassify malware.	Is used in other studies as benchmark.	− Is notices while using HMM approach for malware detection which is considered in research.

Rajesh et al.[2018]: Main disadvantage of machine learning approach to detect malware is its manual steps which overcome by malware detection technique with the help of image processing. Malware classification and detection technique by visualizing gray scale image is big achievement of researchers. Along with many advantages the visualization technique suffers from big disadvantage as a minor change in image pixel can leads to miss-classification of file. Manual steps of feature extraction can be skiped with the help of deep learning approach. Images are directly taken as input for deep learning model which predict the object and removes manual process of feature selection [18]

Shanshan et al. [2018]: The flexibility of android OS makes it most populat operating system and also create big platform for malware attackesrs to work upon. Author had proposed network traffic analysis on multiple levels to identify features and combines it with machine learning algorithm. In this approach HTTP and TCP network flow is monitored to determine the malicious activity. Data is collected under traffic collection module followed by feature extraction, lear-ning based detection to give final results. [19]

TaeGuen et al. [2018]: High popularity of android makes it big opourtunity for malware attacker. In proposed technique features are refined by existance and similarity based feature extraction to effectively extract features for malware detection. The proposed framework is based on features like method op-code features, string and shared library func-tion op-code feature along with API and component feature, permission feature and environmental feature. Taking these features feature vector is generated with parameters like permission/component/predefined settings and get merged into one feature vector. [20]

Li et al. [2018]: A rampant android malware has reached to alarmig scale and millions of malware samples are added in application market every year. In proposed approach multilevel fingureprint is extracted from application by n-gram analysis and feature hashing.

These fingureprint features acts as input to online clasifier. The final decision on application to decide its bening or malware is based on confidence scores of classifier and devices combination function.[21].

Wenjia et al. [2018]: The static malware analysis is concidered as most cost effective and lightweight process to obtain features for malware detection. Proposed technique had used two features - permissions and API function calls which is used as input for the deep learning algorithm. The risky permissions and malicious API calls are combined to make feature set for weight adjusted Droid-Deep_learning approach to distinguish bening from malware. [22]

Mohd et al. [2018]: Malware applications are one of the most widely used tools to pull off cyber security. Author had proposed bio-inspired algorithm approach to select permission features that are reliable and able to identify malicious code. Comparison of bio-inspired-algorith particle swarm optimization-PSO and evolutionary computation is done with information gain to get best features. ROC-curve is used to visualize performance and gives reliable information of per-formance. [23]

Sang et al. [2018]: Converting malicious code to image and visualizing it to identify malware family is effective technique to detect malware. Malware Classification using SimHash and CNN-MCSC approach is used to convert malware code to gray scale images using SimHash function to identify malware family by CNN. [24].

Sitalakshmi et al. [2019]: Increasing treands in auto-mation had also increased the use of internet and devices. This high growth had also opened big platform for malware attackers. Visualizing malware had effectively increased the detection rate of malicious code in this paper authors had proposed hybrid approach of deep learning and visualization. Authors had proposed hybrid model based on similarity and deep learning for analysing images to detect

obfuscated malicious code. The model is cost effective and can be continuously trained in real time environment to detect new malicious code also, [25]

Ahmed et al. [2019]: Off the shelf model use different conflicting method to force misclassification the GEA approach-graph embedding and augmen-tation preserve the functionality of samples generated by off the shelf model and embed bening sample in malicious ones. The approach gives misclassification rate of 100% which shows that full-bodied tool is required to detect malware in IoT by Considering the feature that are not easily manipulated like CFG-control flow graph based features [26]

Ram et al. [2019]: Data on the cloud is fetched through internet that makes it mor prone to malware attack. Authors had mentioned that detection rate by machine learning techique is too less. To overcome the drawback the proposed technique consolidated weigh-ted FuzzyKmeans clustering algo-rithm with Auto Associative Neural-Network(WFCM-AANN) is increased by significant value of precission 92.45%, Recall 75.48%, and F-measure of 58.47%. [27]

Jelena et al. [2019]: Critical problem as compared to malware detection is to optimize the tradeoff between precision, time and powerutilization. during malware detection. Authors had addressed these issues in detection of malware in real-time. Classification is divided into two half process classification on record level followed by classification on application level. Different sliding window alogorithm has been applied along with different monitoring period ranging from 2s to 16s and analyzed that performance gets degraded when nonsuitable parameters are used for certain time-frame. [28]

Hashem et al. [2019]: Unknown malware detection is one of the most challenging task, authors had proposed method to detect unknown malware based on micro patterns existing in executable files for which machine-vission field is used. Firstly executable files gets

converted in images and visual features gets extracted from these images and finally malware is detected by machine learning method. The proposed method focus on difference between behaviour and functionality of malware and bening files. The features gets selected from images of executable files by LBA the most famous texture extraction method. [29]

Evanson et al. [2020]: Evolving and complex phenomena now a days is malicious code threats in IoT. Malware analysis is more complex in IoT as compared to conventional networks due to unique attributes like high scalability, diverse architecture and hetrogenity of devices. Authors had proposed haralick image texture feature along with machine learning to analyse and classify malware in IoT. GLMC- gray level co-occurence matrix is computed on extracted image. Five features are extracted from GLCM namely entropy, angular second and inverse different moment, contrast and correla-tion are considered to classify malware, [30].

Daniel et al. [2020]: TrustSign- an approach malware signature generation based on deep learning VGG19 neural network trained on ImageNet dataset. Approach is to produce signatues on the basis of malicious process present in volatile memory and this approach removes the limitation associated with static and dynamic analysis of malware as per conventional approaches. TrustSign analyze malware in cloud virtualization in trusted manner. Author also claims that the approach followed removes the dependency on executable files as this approach is capable of signing fileless malware also. TrustSign is done in unsupervised manner hence removes the need of human dependency whic add the cost effective parametes in this approach,[31].

Zhongru et al. [2020]: The force behind the growth of IoT devices is open source android platform and also a big platform for malware attackers. The conventional static or dynamic malware detection approach is time consuming due to which an human independent end-to-

end approach is required for malware detection. The proposed method had removed bytecodes of the classes.dex file and make it input of deep learning model, [32].

Danish et al. [2020]: This paper proposed ensemble convolution neural network-CNN based architecture to detect packed and unpacked malware. Different sementic image representation is provided by different CNN architecture. CNN approach had reduced the time that gets spend in gathering features in machine learning approach directly by looking the raw bytes of portable executable files. Author had proposed IMCEC and ensemble techinique to detect malware even under obfuscation,[33].

Preeti et al. [2020]: The new era of computing is using and dependent on cloud, that increased the demand of cloud security from malware attack.KVMInspector a dynamic approach to detect malicious code in KVM cloud environ-ment. Author claims that the proposed approach is more robust as KVM-Inspector is deployed at KVM layer. Raw VM memory and KVM provides both advanced security and basic security by combining VMI and ML techniques at virtualization layer. ,[34].

Andrea et al. [2020]: Proposed VizMal tool visualize execution traces of application to highlight which portion of traces is behaving maliciously.VizMal takes the executaion trace of app as input and outputs the sequence of colored boxes image. Box color represents the degree of maliciousness of application behavious during specific interval. Activeness of application in perticular time duration is represented by size of the box.VizMal is two fold process in which image builder buids the image and trace classifier with maliciousness and activity level. [35]

2.1 Tools and Technologies

- MATLAB:(MATrix LABoratory) is a computing environment having multi-paradigm that supports 4th generation PL (programming language) developed by Math-Works[36] .It is mostly used in applied mathematics and engineering. It has also properties that make it useful in network analysis too. It implements matrix manipulations, data and functions plot, algorithm[37] implementation, user interfaces formation and interaction with programs that are written in other languages like C, C++, C#, Java, Python and Fortran.

- PYTHON: Python is a widely used programming language for general-purpose programming. It allows the programmer to express concepts in fewer lines of code that might be used in C++ or Java. Its interpreter is available for many operating systems which can be run on a wide category of systems. It supports automatic memory management and multiple programming paradigms.

- WEKA: Weka is a assortment of ML algorithms to perform data mining tasks. The algorithms can be applied to a dataset directly or called from own Java code. Weka contains tools[38] for pre-processing, classifica - tion, regression and clustering along with different association rules and visualization.

- JAVA: Java is widely programming language made by James Gosling (Sun Microsystems) in year 1991.The purpose of Java is to compose program once and afterwards run this program im numerous working frameworks. The latest version of Java-SE 17 (LTS) launched is september,2021. Java is characterized by a particular and comprises of a programming language,a compiler, center libraries and a runtime virtual machine. Java runtime allows developers to write code in different languages than Java programming language that can still runs on Java virtual machine.

- RpidMiner: Is an data(information) science platform that utilizes a client-server architecture with server offered on premises basis or in cloud infrastructure. RapidMiner is AI-artificial intelligence for any enterprise through open and extensible platform(data science). This tool is built for analytics teams, RapidMiner unify the entire data science cycle from data preperation to machine learning and also gives predictive model deployment.

- IBM_SPSS: The world's leading statistical software, is intended to tackle business and research issues through adhocand geospatial analysis, analytics and hypothesis testing. SPSS is software bundle used for batched or interactive statistical analysis. Features of SPSS are accessible by drop-down menus or it can also be programmed with 4GL command syntax language having the benefit of reproducing output, simplifying the repetitive task and hand-ling complex data and itsanalyses.

3. Justification for Research

In this section, the motivation behind the work and research gaps which laid the foundation for the problem formulation is discussed. The malware detection techbique ranges from portable executable files to visual approach to classify malware[39] with CNN[51] and deep learning[50] process.[40,41,42] and classification and clustering[43,44] techniques along with ensemble approach.[45,46]

3.1. Motivation

The growth rate of malware is in big volume every day.[47,48] Image processing[52] plays an important role in malware detection and even better if zero day attack can be identified[53]. Analyzing malware in gray scale image is basic technique[54] of image processing to detect malware. Manual extraction of features using machine learning approach is its

biggest disadvantage and author claims that deep learning approach is far better then machine learning approach [55]

3.2. Research Gaps

Malware detection is never ending process Need to develop technique(deep learning, visualization etc.)[56] to accurately identify malware both for signature baed and behaviour based. Constant race between cyber-attackers and antimalware software motivates researchers to develop effective, efficient and economical approach to detect malicious activities.

1) It is a challenging task to accurately identify malware before it breach the users data as malware hides the malicious behaviour and change frequently due to obfuscation approach followed by malware, [19].
2) As IoT is evolving field same is the malicious threats in IoT. The attributes like device hetroginity, diverse architecture and scalability leads to complex analysis of malware. Integration of mobile devices with IoT exposes IoT devices to malware threats for personal information, corporate and financial information.[57]
3) There is a need to generate operating system specific malware detection technique as large number of organizational servers are on linux. The volatile memory-RAM consist of many data structures and hierarchical objects due to which structural feature extraction method can contribute to improved detection technique, [31].
4) The ensemble CNN architecture –IMCEC is capable of learning rich feature but it is not time effective and has not been tested on large dataset.]33]
5) The cloud security is outmost important in todays era. The KVMinspector approach can be further extended to monitor the malicious behaviour in network also so as to analyse both network and program behaviour. [34]

4. Problem Statement

Malicious code or malware attacks are growing daily and in different ares like computer, mobile phones, cloud and IoT to breach the security and extract personal or sensitive data from host machine. Different varients of malware makes it difficult to be identified by single detection technique. Things becomes more critical when malware developers apply wrapping technique to hide it from detection technique both signature based and anomaly based. So there arises the need of technique that helps to identify malicious code to save the data from attackers.

4.1. Research Objectives

The overall objective of this research is to enrich the malware detection with the help of machine learning approach.

1) To study existing state-of-art techniques for detection of malware and Identify the limitations of the existing systems
2) To extract and build set of features from images to classify malware.
3) To design and develop malware detection model based on supervised machine learning techniques
4) To validate proposed model on family of malware dataset and compare with existing models in literature .

4.2. Research Methodology

The methodology of the proposed work is primarily divided into a literature review followed by the development of the proposed approach and comparing it with the existing models.

- To study existing techniques and identifies the challenges in them, study solutions available so far and identifying the parameters that contributes most in malware detection.
- Gather the malware datasets of different families and apply feature selection and feature selection.
- Identification of tools and technologies towards the formation of a solution. Some of the potential tools are Weka, Matlab,IBM-SPSS, Rapid Minner etc.
- Develop the classification model based on selected features. Train the model on training dataset to classify malware.
- Apply test dataset on model to design the confusion matrix and calculate the accuracy and other parameters.
- To compare accuracy and other parameters with existing models/techniques

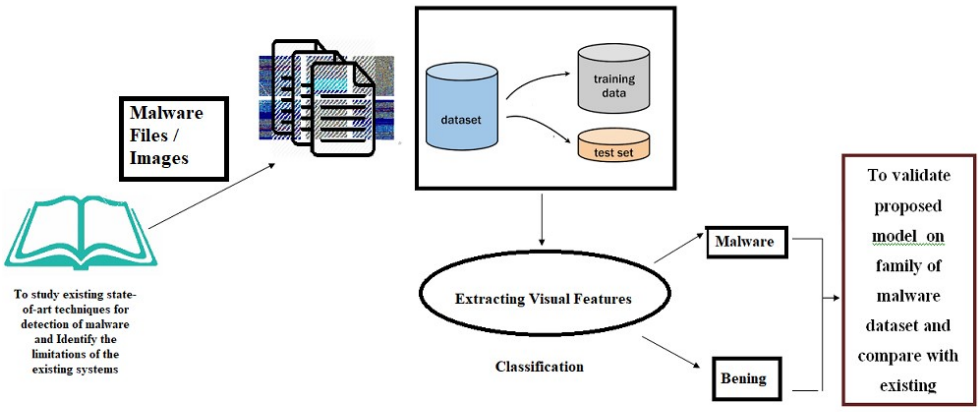

Fig 4. Research Methodology

4.3. Work Plan

Table 4. Work Plan for the Research

Tasks	(July 2018 – June 2019)				(July 2019 – June 2020)				(July 2020 – June 2021)				(June 2021 - July 2022)			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Course work, Submission of Synopsis.																
Identify and study the existing literature of malware detection																
To study and identify the different data sets and parameters.																
Design & Development of proposed model																
Comparison ofresults with existing literature																
Documentation of Thesis.																

Work Completed Work In Progress/future

5. Expected Outcomes

Is model based on supervised machine learning approach to classify malware family. Model is expected to detect malicious code or malware in different areas like machine, cloud or IoT.

Contributions for the research community:0

This research will focus on detecting malware of different families so as to protect personal data from getting breached by malware attackers.

6. References

[1] Saif, D., El-Gokhy, S.M. and Sallam, E., 2018. Deep Belief Networks-based framework for malware detection in Android systems. Alexandria engineering journal, 57(4), pp.4049-4057.

[2] I. IDC Research, "Smartphone os market share, 2015 q2," in IDC Research Report, 2015.

[3] G. Kelly, "Report: 97% of mobile malware is on android. this is the easy way you stay safe," in Forbes Tech, 2014.

[4] Symantec, "Latest intelligence for march 2016," in Symantec Official Blog, 2016.

[5] Gibert, D., Mateu, C. and Planes, J., 2020. The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. Journal of Network and Computer Applications, p.102526.

[6] Taheri, R., Ghahramani, M., Javidan, R., Shojafar, M., Pooranian, Z. and Conti, M., 2020. Similarity-based Android malware detection using Hamming distance of static binary features. Future Generation Computer Systems, 105, pp.230-247.

[7] Amin, M., Tanveer, T.A., Tehseen, M., Khan, M., Khan, F.A. and Anwar, S., 2020. Static malware detection and attribution in android byte-code through an end-to-end deep system. Future Generation Computer Systems, 102, pp.112-126.

[8] Sun, L., Wei, X., Zhang, J., He, L., Philip, S.Y. and Srisa-an, W., 2017, December. Contaminant removal for android malware detection systems. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 1053-1062). IEEE.

[9] Ding, Y., Xia, X., Chen, S. and Li, Y., 2018. A malware detection method based on family behavior graph. Computers & Security, 73, pp.73-86.

[10] Pektaş, A. and Acarman, T., 2017. Classification of malware families based on runtime behaviors. Journal of information security and applications, 37, pp.91-100.

[11] Mirza, Q.K.A., Awan, I. and Younas, M., 2018. CloudIntell: An intelligent malware detection system. Future Generation Computer Systems, 86, pp.1042-1053.

[12] Gu, J., Sun, B., Du, X., Wang, J., Zhuang, Y. and Wang, Z., 2018. Consortium blockchain-based malware detection in mobile devices. IEEE Access, 6, pp.12118-12128.

[13] Kim, H., Kim, J., Kim, Y., Kim, I., Kim, K.J. and Kim, H., 2019. Improvement of malware detection and classification using API call sequence alignment and visualization. Cluster Computing, 22(1), pp.921-929.

[14] Chowdhury, M., Rahman, A. and Islam, R., 2017, June. Malware analysis and detection using data mining and machine learning classification. In International Conference on Applications and Techniques in Cyber Security and Intelligence (pp. 266-274). Edizioni della Normale, Cham.

[15] Yuxin, D. and Siyi, Z., 2019. Malware detection based on deep learning algorithm. Neural Computing and Applications, 31(2), pp.461-472.

[16] Anderson, H.S., Kharkar, A., Filar, B. and Roth, P., 2017. Evading machine learning malware detection. black Hat.

[17] Mohamed, G.A. and Ithnin, N.B., 2017, April. SBRT: API signature behaviour based representation technique for improving metamorphic malware detection. In International Conference of Reliable Information and Communication Technology (pp. 767-777). Springer, Cham.

[18] Kumar, R., Xiaosong, Z., Khan, R.U., Ahad, I. and Kumar, J., 2018, March. Malicious code detection based on image processing using deep learning. In Proceedings of the 2018 International Conference on Computing and Artificial Intelligence (pp. 81-85).

[19] Wang, S., Chen, Z., Yan, Q., Yang, B., Peng, L. and Jia, Z., 2019. A mobile malware detection method using behavior features in network traffic. Journal of Network and Computer Applications, 133, pp.15-25.

[20] Kim, T., Kang, B., Rho, M., Sezer, S. and Im, E.G., 2018. A multimodal deep learning method for android malware detection using various features. IEEE Transactions on Information Forensics and Security, 14(3), pp.773-788.

[21] Zhang, L., Thing, V.L. and Cheng, Y., 2019. A scalable and extensible framework for android malware detection and family attribution. Computers & Security, 80, pp.120-133.

[22] Li, W., Wang, Z., Cai, J. and Cheng, S., 2018, March. An Android malware detection approach using weight-adjusted deep learning. In 2018 International Conference on Computing, Networking and Communications (ICNC) (pp. 437-441). IEEE.

[23] Ab Razak, M.F., Anuar, N.B., Othman, F., Firdaus, A., Afifi, F. and Salleh, R., 2018. Bio-inspired for features optimization and malware detection. Arabian Journal for Science and Engineering, 43(12), pp.6963-6979.

[24] Ni, S., Qian, Q. and Zhang, R., 2018. Malware identification using visualization images and deep learning. Computers & Security, 77, pp.871-885.

[25] Venkatraman, S., Alazab, M. and Vinayakumar, R., 2019. A hybrid deep learning image-based analysis for effective malware detection. Journal of Information Security and Applications, 47, pp.377-389.

[26] Abusnaina, A., Khormali, A., Alasmary, H., Park, J., Anwar, A. and Mohaisen, A., 2019, July. Adversarial learning attacks on graph-based IoT malware detection systems. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS) (pp. 1296-1305). IEEE.

[27] Yadav, R.M., 2019. Effective analysis of malware detection in cloud computing. Computers & Security, 83, pp.14-21.

[28] Milosevic, J., Malek, M. and Ferrante, A., 2019. Time, accuracy and power consumption tradeoff in mobile malware detection systems. Computers & Security, 82, pp.314-328.

[29] Hashemi, H. and Hamzeh, A., 2019. Visual malware detection using local malicious pattern. Journal of Computer Virology and Hacking Techniques, 15(1), pp.1-14.

[30] Karanja, E.M., Masupe, S. and Jeffrey, M.G., 2020. Analysis of internet of things malware using image texture features and machine learning techniques. Internet of Things, 9, p.100153.

[31] Nahmias, D., Cohen, A., Nissim, N. and Elovici, Y., 2020. Deep feature transfer learning for trusted and automated malware signature generation in private cloud environments. Neural Networks, 124, pp.243-257.

[32] Ren, Z., Wu, H., Ning, Q., Hussain, I. and Chen, B., 2020. End-to-end malware detection for android IoT devices using deep learning. Ad Hoc Networks, 101, p.102098.

[33] Vasan, D., Alazab, M., Wassan, S., Safaei, B. and Zheng, Q., 2020. Image-Based malware classification using ensemble of CNN architectures (IMCEC). Computers & Security, p.101748.

[34] Mishra, P., Verma, I. and Gupta, S., 2020. KVMInspector: KVM Based introspection approach to detect malware in cloud environment. Journal of Information Security and Applications, 51, p.102460.

[35] De Lorenzo, A., Martinelli, F., Medvet, E., Mercaldo, F. and Santone, A., 2020. Visualizing the outcome of dynamic analysis of Android malware with VizMal. Journal of Information Security and Applications, 50, p.102423.

[36] Yan, P. and Yan, Z., 2018. A survey on dynamic mobile malware detection. Software Quality Journal, 26(3), pp.891-919.

[37] Sharafaldin, I., Lashkari, A.H. and Ghorbani, A.A., 2018, January. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In ICISSP (pp. 108-116).

[38] Lashkari, A.H., Kadir, A.F.A., Taheri, L. and Ghorbani, A.A., 2018, October. Toward developing a systematic approach to generate benchmark android malware datasets and classification. In 2018 International Carnahan Conference on Security Technology (ICCST) (pp. 1-7). IEEE.

[39] Vidal, J.M., Monge, M.A.S. and Villalba, L.J.G., 2018. A novel pattern recognition system for detecting Android malware by analyzing suspicious boot sequences. Knowledge-Based Systems, 150, pp.198-217.

[40] Al-Dujaili, A., Huang, A., Hemberg, E. and O'Reilly, U.M., 2018, May. Adversarial deep learning for robust detection of binary encoded malware. In 2018 IEEE Security and Privacy Workshops (SPW) (pp. 76-82). IEEE.

[41] Sewak, M., Sahay, S.K. and Rathore, H., 2018, August. An investigation of a deep learning based malware detection system. In Proceedings of the 13th International Conference on Availability, Reliability and Security (pp. 1-5).

[42] Kakisim, A.G., Nar, M., Carkaci, N. and Sogukpinar, I., 2018, November. Analysis and

evaluation of dynamic feature-based malware detection methods. In International Conference on Security for Information Technology and Communications (pp. 247-258). Springer, Cham.

[43] Martín, A., Lara-Cabrera, R. and Camacho, D., 2019. Android malware detection through hybrid features fusion and ensemble classifiers: the AndroPyTool framework and the OmniDroid dataset. Information Fusion, 52, pp.128-142.

[44] Kumara, A. and Jaidhar, C.D., 2018. Automated multi-level malware detection system based on reconstructed semantic view of executables using machine learning techniques at VMM. Future Generation Computer Systems, 79, pp.431-446.

[45] Noor, M., Abbas, H. and Shahid, W.B., 2018. Countering cyber threats for industrial applications: An automated approach for malware evasion detection and analysis. Journal of Network and Computer Applications, 103, pp.249-261.

[46] Ye, Y., Chen, L., Hou, S., Hardy, W. and Li, X., 2018. DeepAM: a heterogeneous deep learning framework for intelligent malware detection. Knowledge and Information Systems, 54(2), pp.265-285.

[47] Cai, H., Meng, N., Ryder, B. and Yao, D., 2018. Droidcat: Effective android malware detection and categorization via app-level profiling. IEEE Transactions on Information Forensics and Security, 14(6), pp.1455-1470.

[48] Zhu, H.J., You, Z.H., Zhu, Z.X., Shi, W.L., Chen, X. and Cheng, L., 2018. DroidDet: effective and robust detection of android malware using static analysis along with rotation forest model. Neurocomputing, 272, pp.638-646.

[49] Lin, C.H., Pao, H.K. and Liao, J.W., 2018. Efficient dynamic malware analysis using virtual time control mechanics. Computers & Security, 73, pp.359-373.

[50] Li, D., Wang, Z. and Xue, Y., 2018, May. Fine-grained android malware detection based on deep learning. In 2018 IEEE Conference on Communications and Network Security (CNS) (pp. 1-2). IEEE.

[51] Abdelsalam, M., Krishnan, R., Huang, Y. and Sandhu, R., 2018, July. Malware detection in cloud infrastructures using convolutional neural networks. In 2018 IEEE 11th International Conference on Cloud Computing (CLOUD) (pp. 162-169). IEEE.

[52] Venkatraman, S. and Alazab, M., 2018. Use of data visualisation for zero-day Malware detection. Security and Communication Networks, 2018.

[53] Kim, J.Y., Bu, S.J. and Cho, S.B., 2018. Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders. Information Sciences, 460, pp.83-102.

[54] Liu, X., Zhang, J., Lin, Y. and Li, H., 2019, June. ATMPA: attacking machine learning-based malware visualization detection methods via adversarial examples. In Proceedings of the International Symposium on Quality of Service (pp. 1-10).

[55] Souri, A. and Hosseini, R., 2018. A state-of-the-art survey of malware detection approaches

using data mining techniques. Human-centric Computing and Information Sciences, 8(1), p.3.

[56] Yen, Y.S. and Sun, H.M., 2019. An android mutation malware detection based on deep learning using visualization of importance from codes. Microelectronics Reliability, 93, pp.109-114.

[57] Sharmeen, S., Huda, S., Abawajy, J.H., Ismail, W.N. and Hassan, M.M., 2018. Malware threats and detection for industrial mobile-IoT networks. IEEE access, 6, pp.15941-15957.