iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35426

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Malware Detection using Deep Learning

T. Shiva Rama Krishna[1], CH. Akhil[2], S. Shivanandha[3], Dr. G. Somasekhar[4], Dr. B. Ramji[5]
[1, 2, 3]*B. Tech Student,* [4] *Associate Professor,* [5]*Assistant Professor, Dept of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India*

Abstract: ***Malicious software or malware continues to pose a major security concern in this digital age as computer users, corporations, and governments witness an exponential growth in malware attacks. Current malware detection solutions adopt Static and Dynamic analysis of malware signatures and behaviour patterns that are time consuming and ineffective in identifying unknown malwares. Recent malwares use polymorphic, metamorphic and other evasive techniques to change the malware behaviour's quickly and to generate large number of malwares. Since new malwares are predominantly variants of existing malwares, machine learning algorithms are being employed recently to conduct an effective malware analysis. This requires extensive feature engineering, feature learning and feature representation. By using the advanced MLAs such as deep learning, the feature engineering phase can be completely avoided. Though some recent research studies exist in this direction, the performance of the algorithms is biased with the training data. There is a need to mitigate bias and evaluate these methods independently in order to arrive at new enhanced methods for effective zero-day malware detection. To fill the gap in literature, this work evaluates classical MLAs and deep learning architectures for malware detection, classification and categorization with both public and private datasets. The train and test splits of public and private datasets used in the experimental analysis are disjoint to each other's and collected in different timescales. In addition, we propose a novel image processing technique with optimal parameters for MLAs and deep learning architectures. A comprehensive experimental evaluation of these methods indicate that deep learning architectures outperform classical MLAs. Overall, this work proposes an effective visual detection of malware using a scalable and hybrid deep learning framework for real-time deployments. The visualization and deep learning architectures for static, dynamic and image processing-based hybrid approach in a big data environment is a new enhanced method for effective zero-day malware detection.***

Keywords: Cybersecurity, Cybercrime, Malware detection, Static and Dynamic analysis, Machine Learning, Deep Learning, Image processing.

I. INTRODUCTION

In this computerized universe of Industry 4.0, the fast progression of advances has influenced the day-by-day exercises in organizations just as in close to home lives. Web of Things (IoT) and applications have prompted the advancement of the cutting-edge idea of the data society. Notwithstanding, security concerns represent a significant test in understanding the advantages of this modern unrest as digital hoodlums assault singular PC's and organizations for taking secret information for monetary profits and making disavowal of administration frameworks. Such aggressors utilize pernicious programming or malware to cause genuine dangers and weakness of frameworks . A malware is a PC program determined to make hurt the working framework (OS). A malware gets various names, for example, adware, spyware, infection, worm, trojan, rootkit, indirect access, ransomware and order and control (C&C) bot, in view of its motivation and conduct. Location and alleviation of malware is an advancing issue in the network protection field. As scientists foster new procedures, malware creators improve their capacity to dodge recognition. Late malwares utilize polymorphic, transformative and other shifty procedures to change the malware practices rapidly and to produce enormous number of malwares. Since new malwares are overwhelmingly variations of existing malwares, AI calculations (MLAs) are being utilized as of late to direct a powerful malware examination. This requires broad component designing, highlight learning and highlight portrayal. By utilizing the high-level MLAs like profound learning, the component designing stage can be totally stayed away from. In spite of the fact that some new exploration examines exist toward this path, the presentation of the calculations is one-sided with the preparation information.

II. LITERATURE SURVEY

Zero-day or obscure malware are made utilizing code muddling strategies that can alter the parent code to deliver posterity duplicates which have a similar usefulness yet with various marks. Current procedures announced in writing do not have the capacity of recognizing zero-day malware with the necessary precision and effectiveness. In this paper, we have proposed and assessed a novel strategy for utilizing a few information mining methods to distinguish and arrange zero-day malware with undeniable degrees of exactness and effectiveness dependent on the recurrence of Windows API calls.

This paper portrays the strategy utilized for the assortment of enormous informational indexes to prepare the classifiers, and examinations the presentation aftereffects of the different information digging calculations embraced for the examination utilizing a completely mechanized instrument created in this exploration to direct the different trial examinations and assessment. Through the presentation aftereffects of these calculations from our exploratory examination, we can assess and talk about the benefits of one information mining calculation over the other for precisely distinguishing zero-day malware effectively. The information mining system utilized in this examination learns through investigating the conduct of existing malignant and kind codes in enormous datasets. We have utilized vigorous classifiers, in particular Naïve Bayes (NB) Algorithm, k- - Nearest Neighbor (kNN) Algorithm, Sequential Minimal Optimization (SMO) Algorithm with 4 distinct pieces (SMO - Normalized PolyKernel, SMO - PolyKernel, SMO - Puk, and SMO-Radial Basis Function (RBF)), Back spread Neural Networks Algorithm, and J48 choice tree and have assessed their presentation. Generally, the robotized information digging framework executed for this investigation has accomplished high evident positive (TP) pace of over 98.5%, and low bogus positive (FP) pace of under 0.025, which has not been accomplished in writing up until now. This is a lot higher than the necessary business acknowledgment level demonstrating that our novel strategy is a significant jump forward in recognizing zero-day malware. This paper additionally offers future bearings for analysts in investigating various parts of confusions that are influencing the IT world today.

III. PROPOSED SYSTEM AND ARCHITECTURE

Now-a-days to detect cyber-attack we are using static and dynamic analysis of request data. Static analysis is based on signature which we will match existing attack signature with new request packet data to identify packet is normal or contains attack signature. Dynamic analysis will use dynamic execution of program to detect malware/attack but dynamic analysis is time consuming. To overcome from this problem and to increase detection accuracy with old and new malware attacks author is using machine learning algorithms and evaluating prediction performance of various machine learning algorithms such as Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, k- Nearest Neighbour's and Deep Learning Algorithms such as Convolution Neural Networks (CNN) and LSTM (Long Short-Term Memory). To implement this and to evaluate machine learning algorithms performance we are using binary malware dataset called 'MALIMG'. This dataset contains 25 families of malware and application will convert this binary dataset into Gray images to generate train and test models for machine learning algorithms. This algorithm converting binary data to images and then generating model so they are called as MalConv CNN and MalConv LSTM and another algorithm refers as EMBER. Application convert dataset into binary images and then used 80% dataset for training model and 20% dataset for testing. Whenever we upload new test malware binary data then application will apply new test data on train model to predict malware class. In dataset total 25 families of malware.

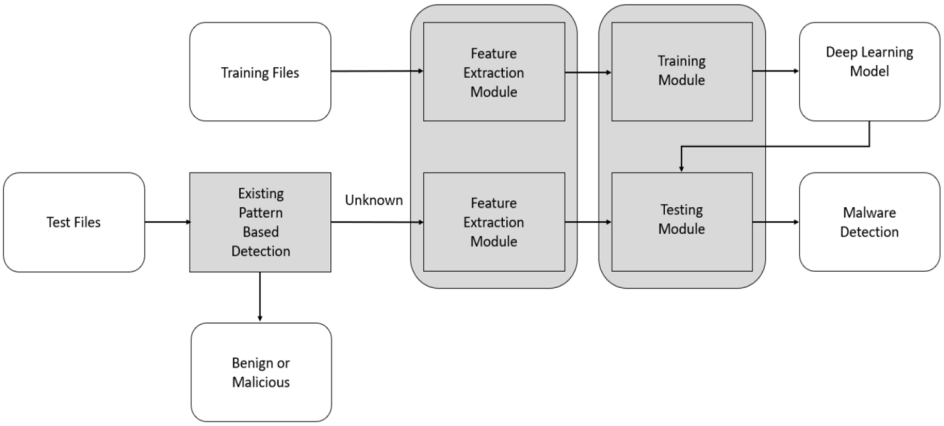


Figure 1: System Architecture

A system architecture is very important to understand the flow of the project. It describes the step-by-step procedure of the complete project. Here, the user gives the dataset which consists of malware families to train and test the models. Then the system will apply the image processing techniques on the input image and classifies the image on the basis of the classification values which were already trained to the system. The output produced will be in accuracy rate of each and every step of the detection process. By this, the user can get the accurate results and can easily analyse the data and compare each of them so that we will know which will be the fastest to detect with a good accuracy.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

IV. IMPLEMENTATION

A. *Module and its Description*

1) *Malware Classification using Image Processing Techniques:* Malware assaults are on the ascent and as of late, new malwares are handily produced as variations of existing malware from a known malware family. To beat this issue, it is essential to gain proficiency with the comparative qualities of malware that can assist with characterizing it into its family. A few investigations directed in this enjoy taken benefit of the way that most malware variations are comparable in structure, with advanced sign and picture handling methods utilized for malware arrangement. They have changed the malware pairs into dim scale pictures and report that malwares from the equivalent malware family appear to be very comparable in design and surface. Since picture preparing procedures require neither dismantling nor code execution, it is quicker in contrast with the Static and Dynamic investigation. The principal benefit of such a methodology is that it can deal with pressed malware, and can chip away at different malwares independent of the working framework. Trial results have shown 98% arrangement exactness on an enormous malware information base and it is additionally strong to well-known jumbling methods to be specific, encryption. They have made benchmarked information, Malimg as open for additional examination. They likewise introduced Search and Retrieval of Malware (SARVAM), an online malware search and recovery framework where parallel executable can be dissected by using closeness measurements. They likewise introduced Sigma, a malware likeness recognition system which depended on signal preparing. Heuristics dependent on the data about the PE structure were utilized to increase the exactness of the sign handling-based highlights. Test results show that Sigma's presentation out wings any remaining static malware discovery strategies as far as precision.

2) *Malware Detection using Deep Learning based on Static Analysis:* We receive an assessment sub module to benchmark the profound learning models dependent on Static investigation. The presentation of different old-style AI and profound learning for static versatile executable (PE) malware location and arrangement are assessed on freely accessible dataset called Ember alongside secretly gathered examples of kind and malwares. The variations of profound learning models are proposed via cautiously following a hyper boundary tuning approach. Investigations identified with profound learning engineering are run till 1,000 ages with fluctuated learning rate [0.01-0.5]. The entirety of the models of old-style AI and profound learning have minimal contrast in their exhibitions. Accordingly, the exhibition of the malware discovery can be upgraded by fusing a cross breed framework pipeline normally called as Windows-Static-Brain Droid (WSBD), which is made out of both traditional AI and profound learning models. WSBD can be conveyed at an association level to identify malware successfully progressively. Ash is utilized with a subset containing 70,140 considerate and 69,860 vindictive records. This dataset is arbitrarily separated into 60% preparing and 40% testing utilizing Scikit-learn. The preparation dataset contains 42,140 considerate documents and 41,860 vindictive records. The testing dataset contains 28,000 considerate documents and 28,000 noxious records. These examples were gotten from VirusTotal3, VirusShare4 and secretly gathered examples of amiable and malware tests. We set up the datasets for directing the trial examination utilizing the accompanying pre-handling stages: 1) Ember, 2) MalConv, 3) Variants of MalConv, 4) Other variations of MalConv. We present the information investigation and results acquired from different tests led on the variations of the current profound learning engineering .In request to assess the exhibition of different traditional AI classifiers like Logistic Regression (LR), Naive Bayes (NB), K-Nearest Neighbor (KNN), Decision Tree (DT), Ada Boost (AB), Random Forest (RF) and Support Vector Machine (SVM) and profound neural organization (DNN) on the area level highlights, we led different analyses utilizing the Ember dataset. All old-style MLAs utilized the default boundaries gave by scikit-learn AI library. At first, two paths of examinations were run for the DNN to discover the ideal boundaries for the quantity of units till 200 epochs.

3) *Malware Detection using Deep Learning based on Dynamic Analysis:* We present an evaluation sub module to compare classical machine learning algorithms and deep learning architectures based on Dynamic analysis for windows malware detection. All the models are examined on the behavioral data that are collected via Dynamic analysis. The parameters for deep networks are selected by following a hyper parameter selection approach with various trials of experiments conducted up to 1,000 epochs with varied learning rate [0.01- 0.5]. Deep learning architectures outperformed the classical MLAs in all types of experiments. This is due to the fact that those deep models are able to learn the optimal, high level and abstract feature representations by passing them into more than one hidden layer. The result of best performed model is not directly comparable, due to the splitting methodology used for training and testing which is entirely different. Within the first 5 seconds of execution, both classical MLAs and deep learning architectures have the capability to detect whether the executable file is benign or malicious.

B. Description Of Dataset

We have employed two types of datasets from previous re- search works. Dataset 1 was collected using VirtualBox[5] virtual machine using Cuckoo Sandbox[6] with a custom pack- age written in the Java library, Sigar[7] to collect the machine activity data. The virtual machine has the capacity of 2GB RAM, 25GB storage, and a single CPU core running 64-bit Windows 7. Dataset 2 was collected in a VirtualBox virtual machine using Cuckoo Sandbox with a custom package written in the Python library, Psutil[8] to collect the machine activity data. The virtual machine has the capacity of 8GB RAM, 25 GB storage, and a single CPU core running 64-bit Windows operating system. The detailed statistics of Dataset 1 and Dataset 2 is reported in Table 1.

Data set	Benign	Malicious	Total
Data set 1	1,21,701	1,18,717	2,40,418
Data set 2	52,245	50,792	1,03,037

Table 1: Statistics of datasets

C. Data Analysis And Results

We embrace a hyper boundary method to distinguish the ideal boundaries for profound learning models so that the malware identification rate is improved. At first, the preparation dataset is haphazardly parted into 70% preparing and 30% approval. The approval information assisted with noticing the preparation precision across various ages. At last, the exhibition of the prepared model is assessed on the test dataset. For network boundaries, three preliminaries of trials are run for the secret units to improve the learning rate with the fundamental CNN and DNN model. Both the CNN and DNN models tests have utilized Adam as analyser and paired cross entropy as misfortune work. Both the models are made out of 3 layers like info layer, covered up layer and a yield layer. In input layer, the two models contain 10 neurons for 10 distinct highlights and the yield layer contains 1 neuron with sigmoid initiation work. To discover the secret units for DNN, different tests are run for the neurons in the reach [4-128]. In the investigations with 64 neurons, DNN performed well in contrast with different neurons. To discover the quantity of channels in CNN, 3 preliminaries of trials are run for the channels in the reach [4-64]. CNN network with channels 32 performed well in contrast with different channels. These boundaries are set for the remainder of the analyses were directed to recognize the ideal boundary for learning rate and different setups of trials for network boundaries were made with learning rate inside the cut-off [0.01-0.5]. In the greater part of the cases, execution of investigations related with lower learning rate was discovered to be acceptable in recognizing the executable as either generous or malware. By auditing the preparation time and the malware recognition rate, the learning rate [0.01-0.5] is utilized for the remainder of the experiments.

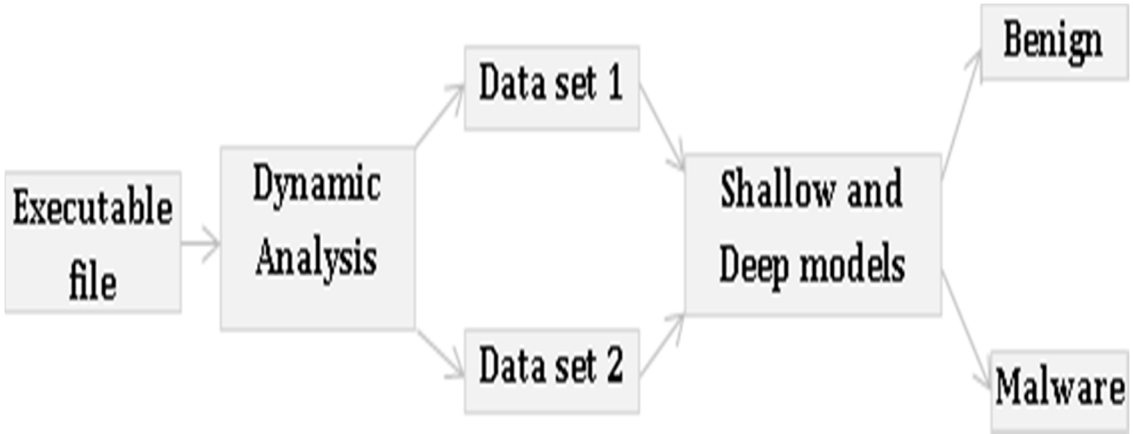


Figure 2: Proposed Deep learning architecture based on Dynamic Analysis

V. RESULTS AND ANALYSIS

The project is based on Machine learning and Deep Learning algorithms in which we be having a dataset and it is sent to two phases.i.e Training data and Test data. Here we will be getting the Prediction results of each and every algorithm after running and with the help of those we can predict which type of malware we are dealing. This is totally based on the Machine Learning.

A. Output Screens

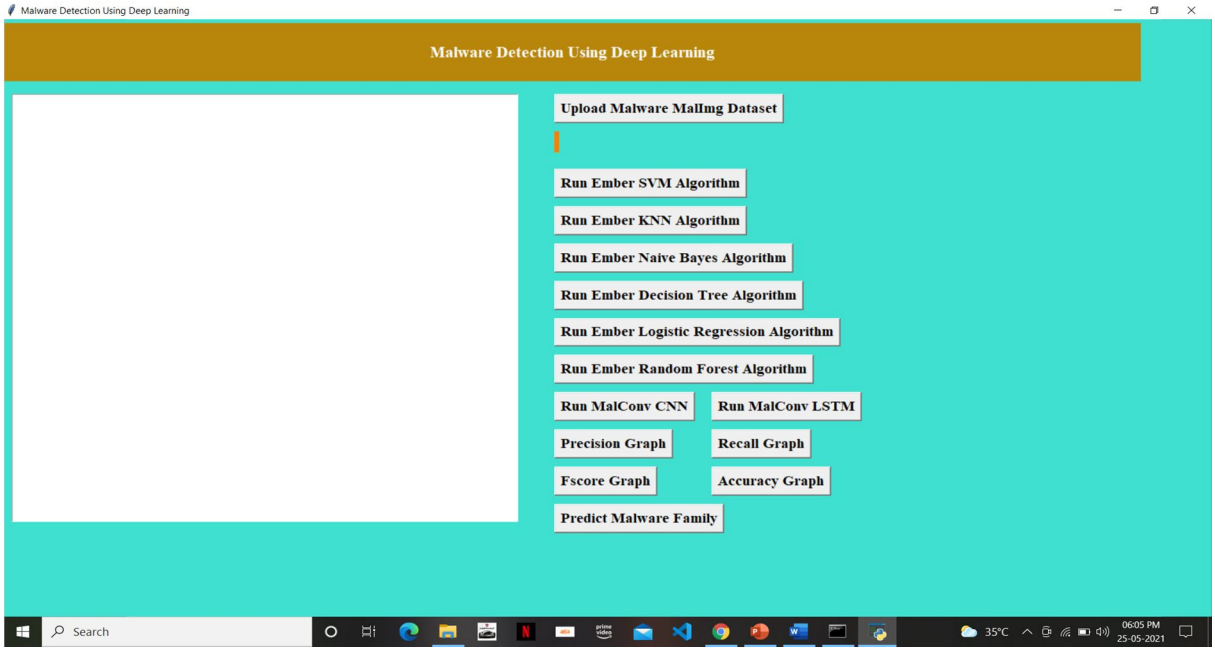


Figure 3 : Main Page

In the above screen we have to upload the Malimg dataset by clicking “Upload Malware Malimg Dataset” button .
After that one by one we have run the algorithms by clicking on the corresponding buttons so that we will be getting the accuracy of the malware dataset. In this we have included total of eight algorithms in which 6 are from Machine Learning and the rest of two are from Deep learning. We can compare the algorithms so that we can conclude which of the algorithms are accurate.

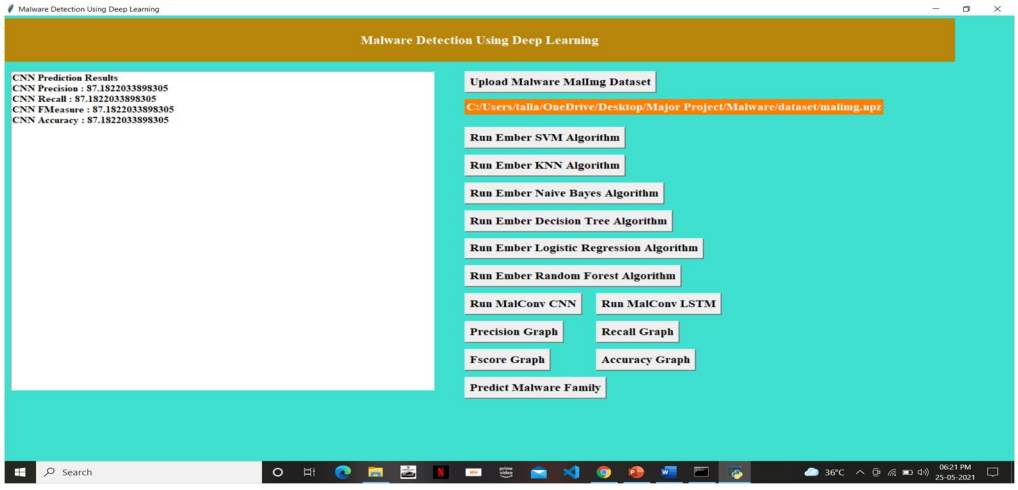


Figure 4: Running Deep learning Algorithms

In the above screen we can see that after running all the algorithms we will be getting Prediction Results of each and every algorithm.

Figure 5 : Accuracy Graph

In the above screen we can see the Accuracy of the algorithms in which there a huge difference between the machine learning and deep leaning algorithms. The accuracy of Deep Learning algorithms is slightly more than that of the Machine Learning algorithms.

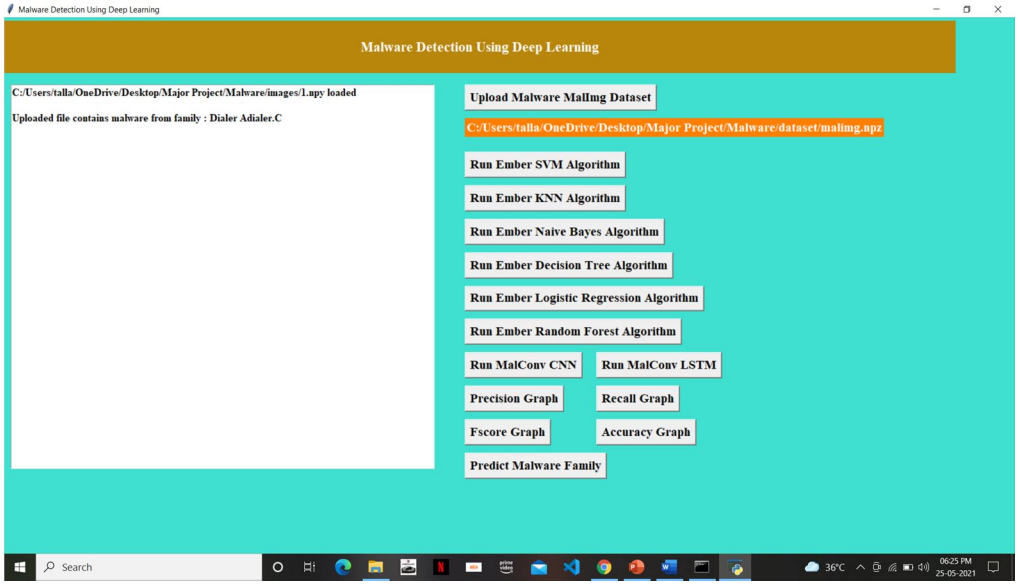


Figure 6 : Malware Prediction

In the above screen if we click on the “Predict Malware Family ” a window shows up with images containing malware families and if we upload any of the image it predicts to which malware family it belongs to. So if we want to predict another malware family then we have upload another dataset ,images and then we will be able to predict to which family it belongs.

1) Accuracy = (TP + TN)/ (TP + TN + FP + FN)
2) Precision =TP/(TP + FP)
3) Recall =TP/(TP + FN)
4) F1 - score = 2 X ((Precision x Recall)/(Precision + Recall)).
5) So, by using above formulas we can calculate the accuracy , precision, recall and F-score of all the algorithms whether they may be machine learning algorithms and deep learning algorithms.

VI. CONCLUSION

This undertaking assesses traditional AI calculations (MLAs) and profound learning structures dependent on Static investigation, Dynamic examination and picture handling strategies for malware location and planned an exceptionally versatile system called ScaleMalNet to identify, arrange and order zero-day malwares. This system applies profound taking in on the gathered malwares from end client has and follows a two-stage measure for malware investigation. In the principal stage, a crossover of Static and Dynamic examination was applied for malware characterization. In the subsequent stage, malwares were assembled into comparing malware classifications utilizing picture preparing approaches. Different trial investigation led by applying varieties in the models on both the openly benefit capable benchmark datasets and secretly gathered datasets in this examination demonstrated that profound learning-based systems outflanked old style MLAs. The created system is equipped for examining huge number of malwares continuously, and scaled out to investigate much bigger number of malwares by stacking a couple of more layers to the current structures. Future examination involves investigation of these varieties with new highlights that could be added to the current information.

VII. FUTURE ENHANCEMENTS

Future enhancement and scope related to the Malware detection are :

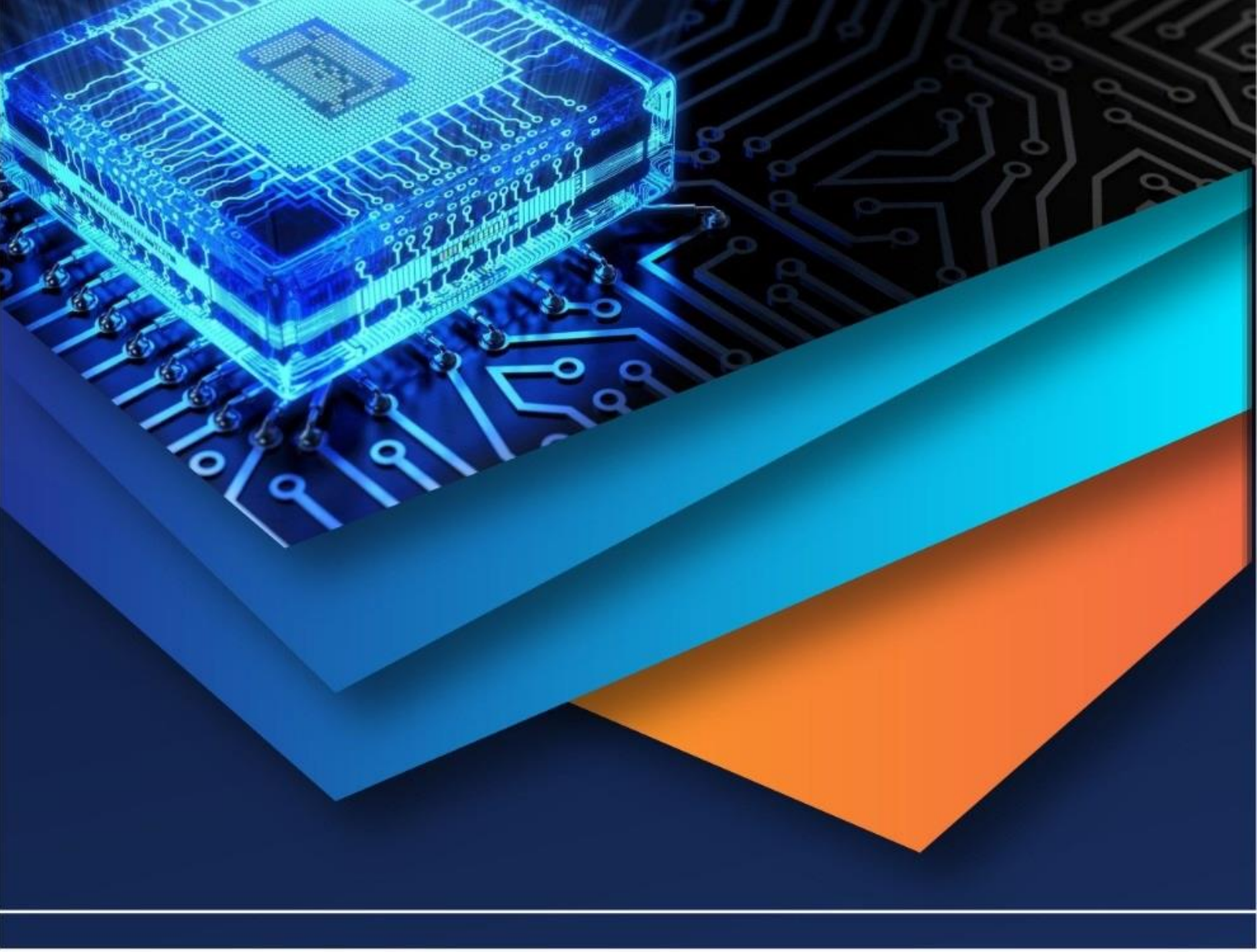
In future work, the spatial pyramid pooling (SPP) layer can be utilized to permit pictures of any size to be utilized as info. This learns highlights at variable scales and it tends to be placed in the middle of the sub inspecting layer and the completely associated layer to improve our model's adaptability. The malware families in Malimg dataset are exceptionally imbalanced. To deal with the multiclass malware families imbalanced issue, cost touchy methodology can be followed. This works with to bring the expense things into the backpropagation learning system of profound learning designs. Fundamentally the expense thing addresses the characterization significance which gives lower worth to the classes that has huge number of tests and higher incentive for the classes that has more modest number of tests. • The profound learning designs are powerless in an ill-disposed climate. The strategy generative antagonistic organization can be utilized to create tests during testing or arrangement stage can without much of a stretch the profound learning structures can trick. In the proposed work, the vigor of the profound learning models isn't examined. This is one of the huge headings towards future work since the malware absconding is a significant application in wellbeing basic climate. A solitary misclassification can make a few harms to the association.

VIII. ACKNOWLEDGEMENT

We take this opportunity to express our gratitude to PRC Coordinator Dr. B. Ramji, and Dr. G. Somasekhar, Department of CSE, CMR Technical Campus, for their guidance and support at every stage of the project. We would like to extend our gratitude to HOD Dr. K. Srujan Raju, Department of CSE, CMR Technical Campus for his support. We also take this opportunity to thank Dr. A. Raji Reddy, Director CMR Technical Campus, for providing us with all the facility that was required.

REFERENCES

[1] Anderson, R., Barton, C., Bohme, R., Clayton, R., Van Eaten, M. J., Levi, M., ... & Savage, S. (2013). Measuring the cost of cybercrime. In The economics of information security and privacy (pp. 265-300). Springer, Berlin, Heidelberg.
[2] Li, B., Roundy, K., Gates, C., & Vorobeychik, Y. (2017, March). Large-scale Identification of Malicious Singleton Files. In Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy (pp. 227-238). ACM.
[3] Alazab, M., Venkataraman, S., & Watters, P. (2010, July). Towards understanding malware behaviour by the extraction of API calls. In 2010 Second Cybercrime and Trustworthy Computing Workshop (pp. 52-59). IEEE.
[4] Tang, M., Alazab, M., & Luo, Y. (2017). Big data for cybersecurity: vulnerability disclosure trends and dependencies. IEEE Transactions on Big Data.
[5] Alazab, M., Venkatraman, S., Watters, P., & Alazab, M. (2011, December). Zero-day malware detection based on supervised learning algorithms of API call signatures. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 171-182). Australian Computer Society, Inc.
[6] Alazab, M., Venkatraman, S., Watters, P., Alazab, M., & Alazab, A. (2011, January). Cybercrime: the case of obfuscated malware. In 7th ICGS3/4th e-Democracy Joint Conferences 2011: Proceedings of the International Conference in Global Security, Safety and Sustainability/International Conference on e-Democracy (pp. 1-8
[7] Alazab, M. (2015). Profiling and classifying the behaviour of malicious codes. Journal of Systems and Software, 100, 91-102.
[8] Huda, S., Abawajy, J., Alazab, M., Abdullahi, M., Islam, R., & Yearwood, J. (2016). Hybrids of Support vector machine wrapper and filter-based framework for malware detection. Future Generation Computer Systems, 55, 376-390.
[9] Raff, E., Sylvester, J., & Nicholas, C. (2017, November). Learning the PE Header, Malware Detection with Minimal Domain Knowledge. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 121-132). ACM.
[10] Rossow, C., Dietrich, C. J., Grier, C., Kresbach, C., Paxson, V., Pohlman, N., ... & Van Steen, M. (2012, May). Prudent practices for designing malware experiments: Status quo and outlook. In Security and Privacy (SP), 2012 IEEE Symposium on (pp. 65-79).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY