An Analysis of Malware Classification Technique by using Machine Learning

P. S. S. Siva Krishna
M.Tech (CST with BDA)
Adikavi Nannaya University
Department of CSE

P. Venkateswara Rao
Associate Professor
Adikavi Nannaya University
Department of CSE

ABSTRACT
Development of the internet causes a major problem to the privacy and security of an organization and to personal systems. Security communities receive the huge number of malware every day, Categorization of malware to their corresponding families based on their behaviour is a complex task is to the computer security community. Traditional anti-virus systems based on the signature extraction procedures fail to classify the new malware. Therefore we propose a machine learning model to classify the malware to their corresponding families using the properties of the malware.

In this paper, we present a Review of Mansour Ahmadi et al.'s Feature fusion for effective Malware Family Classification system, Liu et al.'s Automatic Malware classification and detection system, Bashari et al.'s Malware classification and detection system using ANN. Ashu Sharma et al.'s Classification of advanced Malware system. Finally, we have done a comparative analysis of all the above-mentioned methods.

General Terms
Machine Learning, Analysis, Malware, Security, Algorithms.

Keywords
Windows Malware, Computer Security, Machine Learning, Static Analysis, Malware Classification, Microsoft Malware Data.

1. INTRODUCTION
The Advancement of internet and technology playing a vital role in communication and other applications. Similarly, dark net or unindexed internet also developing, with this development there is a problem to the communication [7]. By using the vulnerabilities of internet security and browsers privacy policies malware developers penetrate the malicious programs across the internet. Malware is a software or a program which is a contraction of malicious software, which causes the problem to the computer systems and other electronic devices which are connected to the internet. Malware properties vary from one another, the type of attack on systems is based on its properties. From that malware were differentiated as Virus, Trojans, Bot, Bug, Adware, Ransomware and etc [7].

Traditional Signature-based detection techniques are the basis of anti-malware vendors to detect the malware [5]. This technique identifies the presence of a malware bytecode in a software with the scanning of malware databases. Malware developers can easily evade from this by catching the bytecode patterns and change the sequence of code in software. Signature-based attacks can't able to detect Zero-Day attacks, this technique can only detect the known malware [1]. Thus, the behavioral-based approach is developed to detect the malware but the malware programmers used packed bytes mechanism to escape from it [7]. By using the mutation techniques modern malware can change their properties to evade from the counter mechanisms, Polymorphic and metamorphic layer techniques also there to avoid the anti-malware techniques. Whatever the technique that should be developed by the anti-malware developers, malware programmers utilize the vulnerabilities of counter mechanisms and develop another program to escape from it[1].

Malware detection and classification is a major problem to the anti-malware industries. This is a global problem, to provide protection to the system most of the researchers conducting researches on it to provide an accurate solution. Anti-malware companies perform analysis on its customer systems to gather information about malware. Microsoft is also one of the malware research organization, it has collected the data in bytecode and hexadecimal code. Basically, malware can be identified in these two forms whenever we are trying to match the bytecodes in it [1]. The data which is produced by Microsoft performs its analysis on over multiple systems across the world. In that byte and hex codes are used to classify the malware into their respected families. Half Terabytes of data generated from Microsoft to classify the malware [11].

The structure of the rest of this paper is as follows Section 2describes the Mansour Ahmadi et al's Novel Feature Extraction, Selection, and Fusion for Effective Malware Family Classification method. Section 3 describes Liu et al's. Automatic malware classification using and new malware detection using machine learning method. Section4 describes Bashari Rad et al's. Malware classification and detection using artificial neural network method. Section 5 describes the Ashu Sharma et al's. An effective approach for classification of advanced malware with high accuracy. Finally, we did a comparative analysis on all the above-mentioned methods and in Section 8 concludes the paper.

2. ANALYSIS OF MANSOUR AHMADI ET AL.'S METHOD
This malware classification method focused on a learning-based system which uses different malware characteristics to effectively assign the malware samples to their corresponding families. These Malware samples having the assembly data and byte data to classify the malware. Extraction and evaluation of features from the malware samples based on the content and structure of a malware that is directly performed on the packed executable file. N-gram, Metadata, Entropy, Image representation, String length features were the Hex dump-based features and Metadata, Symbol, Operation Code, Register, Application Program Interface, Section, Data Define and Miscellaneous were the features from the disassembled

files. They took only a Limited number of features compared to the other state-of-the-art systems so that the method is applicable to be used in large-scale malware categorization tasks. They used a feature fusion mechanism to accumulate the most effective features thus avoid the possible features and this was to maintain a trade-off between the features and accuracy. The high performance and effectiveness of XGBoost classification algorithm is the main motivating reason to used it as to classify the malware. Additionally, they used a bagging concept to boost the single model and yet an efficient method to improve the classification quality. To measure the classification performance they are assessed by using two measures namely the accuracy and logarithmic loss.

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

Eq 1: Log-loss equation for the classification model evaluation.

Where N is the number of observations, M is the number of class labels, the log is the natural logarithm, and y_ij indicates whether sample I belongs to class j or not, p_ij indicates the probability of sample I belonging to class j.

That proposed methodology for the classification of malware by Mansour Ahmadi et al. achieved a promising accuracy on the training set of 99.77%, as well as a very low log-loss of 0.0096 on the combination of all categories, and 99.76% accuracy and 0.0094 log-loss on the combination of the best feature categories.

This method has not yet been tested for robustness against evasion attacks or poisoning attacks. The use of a reduced set of features may ease the task of for an analyst to understand the classification results from the set of features related to a given sample but they haven't addressed this issue in their paper.

3. ANALYSIS OF LIU ET AL.'S METHOD

Liu et al.'s propose an incremental malware detection system to classify the malware families and to detect the new malware. This system divided into three main parts: Feature Extraction and selection, Decision Making and New Malware Detection

They had made the following contributions:

- They propose a feature extraction method based on gray-scale images, Opcode n-gram and other important features and mapped them into the feature vector for machine learning.
- They use improved information gain to reduce the high-dimensional features
- Decision-making system to assign the unknown malware to a corresponding family and to screen out suspicious software.
- They apply SNN to find new malware.

As shown in Fig. 1, The Proposed system can be divided into three parts:

- Data Processing
- Decision Making
- Clustering

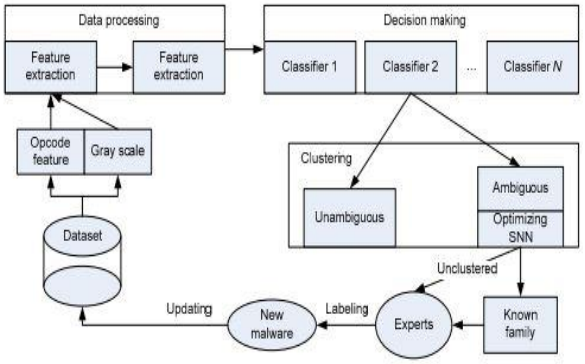

Fig 1. System Architecture

Data Processing handle the feature extraction and selection, this system uses Opcode, n-gram, gray-scale images, and other important functions as features by performing statistical analysis on the data to extract features of malware. Selection procedure reduces the dimension of the feature to improve the classifier performance. Several classifiers are trained by a large number of instances to predict the unknown malware samples by Decision-making system. Suspicious samples are assigned to their family by the clustering process.

IDA Pro transforms each malware sample into a binary file and assembly file. In this experiment, they use Random Forest, K-Nearest Neighbour, Gradient Boost, Naive Bayes, Logistic Regression, SP and Decision Trees classification algorithms are used to classify the malware. Each algorithm is used in every feature that was extracted from the malware data. An n-gram is an efficient method for text feature extraction, the length determines the performance of the algorithm. Nine-gram lengths were applied on seven algorithms, the performance of the algorithms decreased when n>4 and after that they combine all the features applied it on seven algorithms the RF and GB made up with 98.9% and 97.5% accuracy.

To evaluate this method effectively they use GIST & LMgist module based on Mat lab R2012b to extract the features from the images and use seven classifiers to test its accuracy for malware classification.

This system uses three parts: data processing, decision making, and clustering, by using these three parts this proposed method not only classify the malware but also identifies unknown malware. But this model didn't mention anything about the Zero-day attacks and poison attacks.

4. ANALYSIS OF BASHARI ET AL.'S METHOD

This Bashari et al.'s classification method focuses on the implementation of a neural network classifier that can classify an unseen PE file from windows library functions calls as malicious or benign. ANN had been used as a classification model for classifying the malware. It takes x_1, x_2…, x_j as inputs and w_1 , w_2 , … . , w_j as weights and the weighted sum $\sum jwj*xj + b$ is a function to define a hyper plane to be

divided as Positive or negative else 0 or 1. To train a model they collected malware samples from the malware repositories Zoo's that are shared by researchers or organizers. Clean PE files collected by traversing through a clean Windows system directory.

After extracting the samples they placed them in separate folders such as malicious files and benign files. Feature extraction should be done for the each PE sample. Function calls need to extract from the samples and unique function calls are getting separated from it. When the function calls are comparing with the list and if a function call exists within the sample, the respective index marked as 1 and every function call had a binary vector space, with the size equal to the total number of unique function calls across all the sample files. The proposed neural network classifier indicating whether the sample is malicious (0, 1) or benign (1, 0). For every unique function call, the input layer consists of equal inputs. Back Propagation is done every time a forward pass is done, GD is used update weights and reduce the cost for the next epoch.

The output neuron will use a soft max function as it is the most common activation function used in the output layer alongside a cross-entropy cost-function.

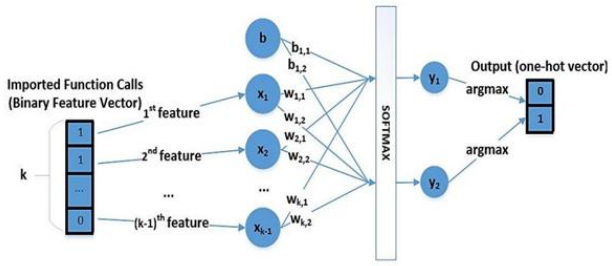

Fig. 2. Neural Network with Soft max Function Architecture in this proposed system.

10-fold cross-validation is used to evaluate the predicted model. Accuracy, recall and precision metrics were used for the evaluating the model. They perform 10 iterations on train and test data for every iteration they get different values for accuracy, recall, and precision.

$$\textbf{\textit{Precision}} = TP / (TP + FP)$$

$$\textbf{\textit{Recall}} = TP / (TP + FN)$$

$$\textbf{\textit{Accuracy}} = \text{Total correct predictions / Number of test instances}$$

Eq 2: Formulaic equations for the classifier evaluation metrics

In this proposed model, which can classify an unseen PE (PE) file as benign or malicious based on its loaded library function calls? The implemented model achieved an average accuracy of 97.8%, with 97.6% precision, and 96.6% recall over a dataset size of 4,000 PE files.This proposed system only deals with the few features and less data, to maintain a trade-off between the complexity of data and the accuracy, scalable approach is required. But they didn't mention any of these points in the prescribed paper. The huge generation of malware from different sources degrades the performance of this model and can't accurately classify the malware when new malware is raised.

5. ANALYSIS OF ASHU SHARMA ET AL.'S METHOD

Ashu Sharma et al.'s proposed a malware classification method which is collected malware data from malacia project and benign data collected from uninfected computer systems to classify the given sample malware or benign. These samples are applied to the 13 machine learning classifiers for training the models and test the samples with Top 5 classifiers based on accuracy, the below figure describes the proposed method process. By performing static analysis on data they extract the features, those features are opcodes of the executables obtained by objdump utility available in the Linux system.To find the best classifiers for detection of unknown malware they test thirteen tree-based classifiers viz. Random forest, J48, REPTREE, LMT, Decision stump, ADT, NBT, FT, LAD, Random Tree, Simple CART, BFT and J48 Graft available in the popular and widely used suite of machine learning software known as WEKA. Then with the obtained features, we run the WEKA n-fold cross-validation to train all the selected classifiers. Figure 3 shows the accuracy obtained by all classifiers for n = 2,4,6...,16 folds. We observed that Random forest is the best classifier and its accuracy is almost flat after n = 2.

The effectiveness of the top five classifiers viz. Random forest Tree, LMT, NBT, J48, and FT has been studied with the randomly selected 750 malware and 610 benign programs. The analysis is done in WEKA with ten-fold cross-validation, in terms of True Positive Ratio (TPR), True Negative Ratio (TNR), False Positive Ratio (FPR), False Negative Ratio (FNR) and accuracy. From the analysis, it is clear that Random forest (97.95%) is the best classifier for identification of unknown malware. Nevertheless, the other classifiers are also reasonably good (> 96.2%) for the detection of unknown malware**.**

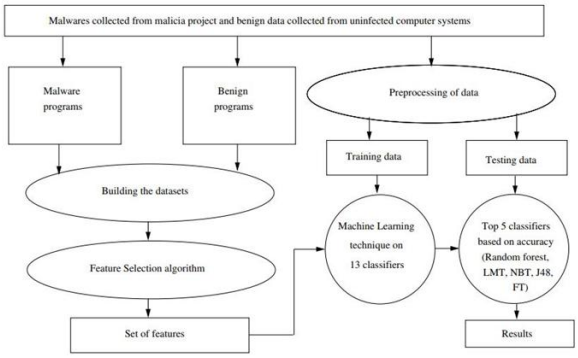

Fig 3. System Architecture

This proposed model only performs with the fewer data, but in the real time a scalable approach is required to classify the huge malware in a fast rate but with the usage more number of classifiers to train and test the model this method leads to a high complexity structure and time-consuming process.

6. COMPARATIVE STUDY

The following table shows the comparison among the Mansour Ahmadi et al.'s Malware Classification System, Liu et al.'s Automatic Malware Classification and Detection System, Bashari et al.'s Malware Classification System using ANN, and Ashu Sharma et al.'s Malware Classification System. Mansour Ahmadi et al.'s Malware Classification System derives classification of malware on static properties using XGBoost method with 99.77% accuracy. Liu et al.'s Automatic Malware classification and Detection system used

Random Forest (RF), Gradient-Boosting (GB), Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine-poly (SP), K-Nearest Neighbour (KNN), Decision Trees (DT) methods for classification and detection of malware and get 98.9% accuracy on static properties of malware. Bashari et al.'s system shows that classification and detection of static malware using Artificial Neural Networks (ANN) with 97.8% accuracy and Ashu Sharma et al.'s Malware classification system achieves 97.95% accuracy on Static Properties of malware by using Random Forest (RF).

Table 1: Comparative Study of Various Malware Classifier Systems.

Author	**Type**	**Analysis**	**Methods**	**Result**
Mansour Ahmadi et al's Malware Classification System	Classification	**Static**	XGBoost	99.77%
Liu et al.'s Automatic Malware Classification and Detection System	Classification & Detection	Static	RF,GB,LR,SP, K-NN,NB,DT	98.9%
Bashari et al.'s Malware Classification System using ANN	Classification & Detection	Static	Artificial Neural Networks	97.8%
Ashu Sharma et al.'s Malware Classification System	Classification	Static	Random Forest	97.95%

7. CONCLUSION

Malware Classification and Detection has become a very popular field of research. There are many issues as it processes executables data. In this paper, we have analyzed some of the approaches of Malware classification methods at their accuracy level and with statistical analysis. Ensemble classification algorithms take less time to classify the malware than Classification algorithms. It can be concluded that extensive and combination approaches will be done on real-world data sets, with an expectation to achieve comparable or greater accuracy than the existing techniques.In the future, we can plan to extend and improve this by implementing a novel method for classifying the malware with less complexity and with the usage of very important features. And then compare the performance with present approaches.

8. REFERENCES

[1] Mansour Ahmadi, Dmitry Ulyanov, Stanislav Semenov, Mikhail Trofimov, Giorgio Giacinto: Novel Feature Extraction, Selection, and Fusion for Effective Malware Family Classification.CODASPY 2016: 183-194

[2] Liu. L, Wang, B, Yu, B. et al. Frontiers Inf Technol Electronic Eng (2017) 18: 1336.

[3] Bashari Rad, Babak & Shahpasand, Maryam & Kazem Hassan Nejad, Mohammad. (2018). Malware classification and detection using the artificial neural network. Journal of Engineering Science and Technology. 14-23.

[4] Sahay, Sanjay. (2016). An effective approach for classification of advanced malware with high accuracy. International Journal of Security and its Applications. 10. 249-266.

[5] Mansour Ahmadi, Ashkan Sami, Hossein Rahimi, Babak Yadegari, Malware detection by behavioral sequential patterns, Computer Fraud & Security, Volume 2013, Issue 8,2013, Pages 11-19, ISSN 1361-3723.

[6] AnimeshPatcha, Jung-Min Park, An overview of anomaly detection techniques: Existing solutions and latest technological trends, Computer Networks, Volume 51, Issue 12, 2007, Pages 3448-3470, ISSN 1389-1286.

[7] Nir Nissim, Robert Moskovitch, LiorRokach, Yuval Elovici, Novel active learning methods for enhanced PC malware detection in windows OS, Expert Systems with Applications, Volume 41, Issue 13,2014, Pages 5843-5857, ISSN 0957-4174.

[8] Yujie Fan, Yanfang Ye, Lifei Chen, Malicious sequential pattern mining for automatic malware detection, Expert Systems with Applications, Volume 52,2016, Pages 16-25, ISSN 0957-4174.

[9] D. Gavriluţ, M. Cimpoeşu, D. Anton and L. Ciortuz, "Malware detection using machine learning," 2009 International Multiconference on Computer Science and Information Technology, Mragowo, 2009, pp. 735-741.

[10] R. Tian, R. Islam, L. Batten, and S. Versteeg, "Differentiating malware from clean ware using behavioral analysis," 2010 5th International Conference on Malicious and Unwanted Software, Nancy, Lorraine, 2010, pp. 23-30.

[11] Ronen, Royi & Radu, Marian & Feuerstein, Corina & Yom-Tov, Elad & Ahmadi, Mansour. (2018). Microsoft Malware Classification Challenge. 10.13140/RG.2.2.34695.91045.Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[12] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[13] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[14] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[15] Spector, A. Z. 1989. Achieving application requirements. 8. Yujie Fan, Yanfang Ye, Lifei Chen, Malicious sequential pattern mining for automatic malware detection, Expert Systems with Applications, Volume 52,2016, Pages 16-25, ISSN 0957-4174.