IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEEP LEARNING ALOGRITHM BASED ON A NEW MALWARE CLASSIFICATION FRAMEWORK

R.Krishnapriya[1],Mrs.P.Jasmine Lois Ebenezar[2],Mrs.EJulie Ruth[4] ,Mrs.J.Steffi[3]and Mrs.E.Julie Ruth[4]

Department of Computer Applications, Sarah Tucker College, Tirunelveli-7.

Abstract:

Recent technological developments in computer systems transfer human life from real to virtual environments. Covid-19 disease has accelerated this process. Cyber criminals' interest has shifted from real to virtual life as well. This is because it is easier to commit a crime in cyberspace rather than in regular life. Malicious software (malware) is unwanted software which is frequently used by cyber criminals to launch cyber-attacks. Malware variants are continuing to evolve by using advanced obfuscation and packing techniques. These concealing techniques make malware detection and classification significantly challenging. Novel methods which are quite different from traditional methods must be used to effectively combat new malware variants. Traditional artificial intelligence (AI), specifically machine learning (ML) algorithms, are no longer effective in detecting all new and complex malware variants. A deep learning (DL) approach, which is quite different from traditional ML algorithms, can be a promising solution to the problem of detecting all variants of malware. In this study, a novel deep-learning-based architecture is proposed which can classify malware variants based on a hybrid model. The main contribution of the study is to propose a new hybrid architecture which integrates two wide-ranging pre-trained network models in an optimized manner. This architecture consists of four main stages, namely: data acquisition, the design of deep neural network architecture, training of the proposed deep neural network architecture, and evaluation of the trained deep neural network. The proposed method was tested on Malimg, Microsoft BIG 2015, and Malevis datasets. The experimental results show that the suggested method can effectively classify malware with high accuracy, which outperforms the state of the art methods in the literature. When the proposed method was tested on the Malimg dataset, 97.78% accuracy was obtained, which outperformed most of the ML-based malware detection methods.

1. INTRODUCTION

Human existence has become easier and more convenient as a result of recent technical advancements in computer systems and the Internet. Everything these days can be done on the Internet, including social interaction, monetary transactions, and the assessment of human bodily changes, among other things. All of these improvements entice cyber criminals to conduct crimes online rather than in the actual world. Cyber attacks cost the global economy trillions of dollars, according to recent scientific and industry research . Malware is frequently used by cyber thieves to launch cyber attacks. Malware is any software that performs undesired and suspicious actions on the computers of its victims. Viruses, worms, Trojan horses, rootkits, and ransomware are just a few examples of malware. . Malware variations can steal sensitive information, launch distributed denial of service (DDoS) assaults, and cause computer systems to malfunction. New malware strains employ strategies such as obfuscation.To stay undetectable in the victim systems, encryption and packaging are used . Human trust was used as an infectious vector in the transmission of these novel variations. Opening email attachments, downloading bogus software, and viewing and downloading files from phoney websites, for example, are all well-known malware distribution routes.

We must detect malware as soon as it attacks computer systems in order to secure them. The process of evaluating a suspicious file and determining whether it is malware or benign is known as malware detection. The taxonomy of malware is taken a step further. After a file has been recognised as malware, malware categorization is used to specify the category or family of malware. Malware detection is a three-step process:

1. Malware files are examined using proper software.
2. From the studied files, static and dynamic features are extracted.
3. To distinguish dangerous software from benign software, features are arranged in specific ways.

Different disciplines and techniques, including as data science, machine learning, and heuristics, as well as technology like as cloud computing, big data, and block chain, are utilised in these procedures to boost the detection rate. Using the tools and technology mentioned above, there are several malware detection approaches. Signature-, behavior-, model-, and heuristic-based detection are the most common approaches. The names of these methods differ depending on the techniques and technology employed. For known and comparable versions of the same malware, a signature-based method is useful. However, it is unable to detect malware that has never been seen before. Although behavior-based, heuristic-based, and model-checking-based detection techniques are efficient in detecting some aspects of unknown malware, they do not perform as well when it comes to detecting sophisticated malware variants.

Deep learning-based approaches are now being employed as a new paradigm to address the flaws in traditional malware detection and classification methods. Image processing, computer vision, human action recognition , driving safety ,face expression detection , and natural language processing are just a few of the areas where deep learning has been employed widely. However, it has not been utilised to its full potential in the field of cyber security, particularly in the identification of malware. Artificial neural networks are used in deep learning, which is a subset of artificial intelligence (ANN). Deep learning learns from examples and employs numerous hidden layers. Deep neural networks (DNN), deep belief networks (DBN), recurrent neural networks (RNN), and convolutional neural networks are some of the deep learning designs that have lately been employed to improve model performance.

1. The DL model can produce high-level features from existing characteristics automatically.

2. The necessity for feature engineering is reduced with DL.

3. DL is capable of efficiently handling unstructured data.

4. DL can handle extremely huge datasets.

5. DL lowers the amount of space available for features.

6. DL is capable of unsupervised, semi-supervised, and supervised learning.

7. DL saves money and improves accuracy.

This paper presents a revolutionary hybrid deep-learning-based malware classification architecture. Malware data from the Microsoft BIG 2015, Malimg, and Malevis datasets is first transformed into grayscale images and then delivered to the system in the suggested manner. After the picture acquisition stage is completed, the proposed technique extracts high-level malware features from malware images utilising the suggested hybrid architecture's convolution layers. Finally, the system is put through its paces in a controlled environment manner. In the proposed model, three comprehensive deep-learning models that rely on a transfer-learning method are integrated to generate a hybrid model. Several hidden layers and the Rectified Linear Unit (ReLU) function are used in the aforementioned operations. According to the test results, the suggested method can efficiently extract different properties for each malware kind and family for categorization. Experiment findings further shown that the proposed DL technique accurately classifies different malware variants, outperforming state-of-the-art methods in the literature. The following are the paper's significant contributions:

1. A novel malware classification approach based on hybrid deep learning is proposed.

2. The proposed strategy presents a new hybrid layer that consists of two pre-trained models rather than one.

3. For diverse forms of malware, distinct properties are derived from the data.

4. The proposed technique considerably lowers feature spaces.

5. Three well-known malware datasets are used to test the proposed technique.

6. Accuracy rates measured are higher than recognised approaches.

The rest of this paper is laid out as follows. Section II discusses the malware analysis, feature extraction, and detection processes, as well as a review of the literature's existing malware detection and classification approaches. A proposed hybrid deep learning architecture framework is described in Section III. The experimental data and comments are presented in Section IV. The conclusion and future study directions are presented in Section V.

2. RELATED WORKS

In [1], the authors have presented a detailed review on malware detection approaches and recent detection methods which use these approaches. Paper goal is to help researchers to have a general idea of the malware detection approaches, pros and cons of each detection approach, and methods that are used in these approaches.

In [2], the authors have proposed a dynamic analysis for IoT malware detection (DAIMD) to reduce damage to IoT devices by detecting both well-known IoT malware and new and variant IoT malware evolved intelligently. The DAIMD scheme learns IoT malware using the convolution neural network (CNN) model and analyzes IoT malware dynamically in nested cloud environment. DAIMD performs dynamic analysis on IoT malware in a nested cloud environment to extract behaviors related to memory, network, virtual file system, process, and system call. By converting the extracted and analyzed behavior data into images, the behavior images of IoT malware are classified and trained in the Convolution Neural Network (CNN). DAIMD can minimize the infection damage of IoT devices from malware by visualizing and learning the vast amount of behavior data generated through dynamic analysis.

In [3], the authors have considered a heuristic malware detection method based on dynamic analysis of API calls. We utilize a naïve Bayes classifier to distinguish between benign and malware samples, and Levenshte in distance is shown to increase the effectiveness of the technique. For comparison, we consider commercial anti-malware products. We provide experimental evidence of the strength of our technique based on a substantial malware dataset. We show that our approach achieves particularly impressive results in detecting packed malware samples.

In [4], the authors have proposed a novel deep-learning-based architecture for the recognition and prediction of human actions based on a hybrid model. The main contribution of this study is to propose a new hybrid architecture, integrating four wide-ranging pre-trained network models in an optimized manner, using a metaheuristic algorithm. This architecture consists of four main stages: namely, the creation of the data set, the design of deep neural network (DNN) architecture, training and optimization of the proposed DNN architecture, and evaluation of the trained DNN. By adapting the aforementioned architecture, reliable features are obtained for the training procedure. In order to validate the superiority of the proposed architecture over other state-of-the-art studies, a performance evaluation between these architectures is presented using benchmark datasets.

The existing system is based on a layered ensemble approach that mimics the key characteristics of deep learning techniques but performs better than the latter. The proposed system does not require hyperparameter tuning or backpropagation and works with reduced model complexity in [5].

In [6], the authors have detected malware, we have employed various contemporary convolutional neural networks (Resnet, Inception, DenseNet, VGG, AlexNet) that have proven success in image classification problem and compared their predictive performance along with duration of model production and inference.

In [7], the authors have analyzed the correspondence between bytecode and disassembly of malware, and propose a new feature extraction method based on multi-dimensional sequence. Also, we construct a new classification framework based on attention mechanism and Convolutional Neural Networks mechanism. Furthermore, we also compare the different architectures based on the attention mechanisms. Experiments on open datasets show that our feature extraction method and our framework have a good classification effect, and the accuracy rate is 0.9609.

In [8], the authors have evaluated the classical MLAs and deep learning architectures for malware detection, classification, and categorization using different public and private datasets. Second, we remove all the dataset bias removed in the experimental analysis by having different splits of the public and private datasets to train and test the model in a disjoint way using different timescales. Third, our major contribution is in proposing a novel image processing technique with optimal parameters...

In [9], a malware classification methodology based on its binary image and extracting local binary pattern (LBP) features is proposed. First, malware images are reorganized into 3 by 3 grids which is mainly used to extract LBP feature. Second, the LBP is implemented on the malware images to extract features in

that it is useful in pattern or texture classification. Finally, Tensorflow, a library for machine learning, is applied to classify malware images with the LBP feature. Performance comparison results among different classifiers with different image descriptors such as GIST, a spatial envelop, and the LBP demonstrate that our proposed approach outperforms others.

3. PROPOSED MODEL

This section outlines our proposed malware categorization framework, which is based on deep learning. For malware categorization, this system provides a hybrid deep neural network architecture. The suggested system's technique, as shown in Figure 1, consists of three basic components. To begin, the malware data is gathered through the use of various comprehensive datasets. Second, pre-trained networks are used to extract the low and high level malware features. Finally, the supervised learning approach is used to execute the training phase of our deep neural network architecture.

Malware visualization and model overview are the two primary sub-sections of this section. Malware variations in binary file form are displayed as grey scale images in the malware visualization section. The proposed model is described in the model overview section.

A.MALWARE VISUALIZATION AS AN IMAGE FRAME

There are various methods for converting binary code to pictures. We used visualisation of executable malware binary files in our research . Our goal is to create a grayscale image from binary files. The technique of converting malware binary data to grayscale photographs is depicted in Figure 3. Figure 3 shows how the first malware binary file is read into an 8-bit unsigned integer vector. After that, using equation, the binary value of each component A $=(a_7+a_6+a_5+a_4+a_3+a_2+a_1+a_0)$ is translated to its decimal equivalent. Finally, the decimal vector is moulded into a 2D matrix, which is then translated into a grayscale image. Figure 3 shows samples of malware visualization image frames from several malware families as a consequence of the analysis.

$$A=(a*2^0+a*2^1+a*2^2+a*2^3+a*2^4+a*2^5+a*2^6+a*2^7) \quad (1)$$

B.PROPOSEDMODELOVERVIEWFORMALWARECLASSIFICATION

The suggested methodology presents a framework for malware categorization that is more efficient. The architecture of this framework is a hybrid deep neural network architecture. The suggested framework's technique, as shown in Figure 3, is divided into four phases: malware data collecting, deep neural network architecture design, training phase, and evaluation stage. In addition, the system flowchart in Figure 4 depicts a more complete characterization of those stages. Four steps are defined in three parts here. In Figure 4, the feature extractors are the pretrained networks from the pre-training section. In addition, the first three layers in the training portion show fully connected layers for the learning process, while the last layer exhibits a softmax classifier for the classification process

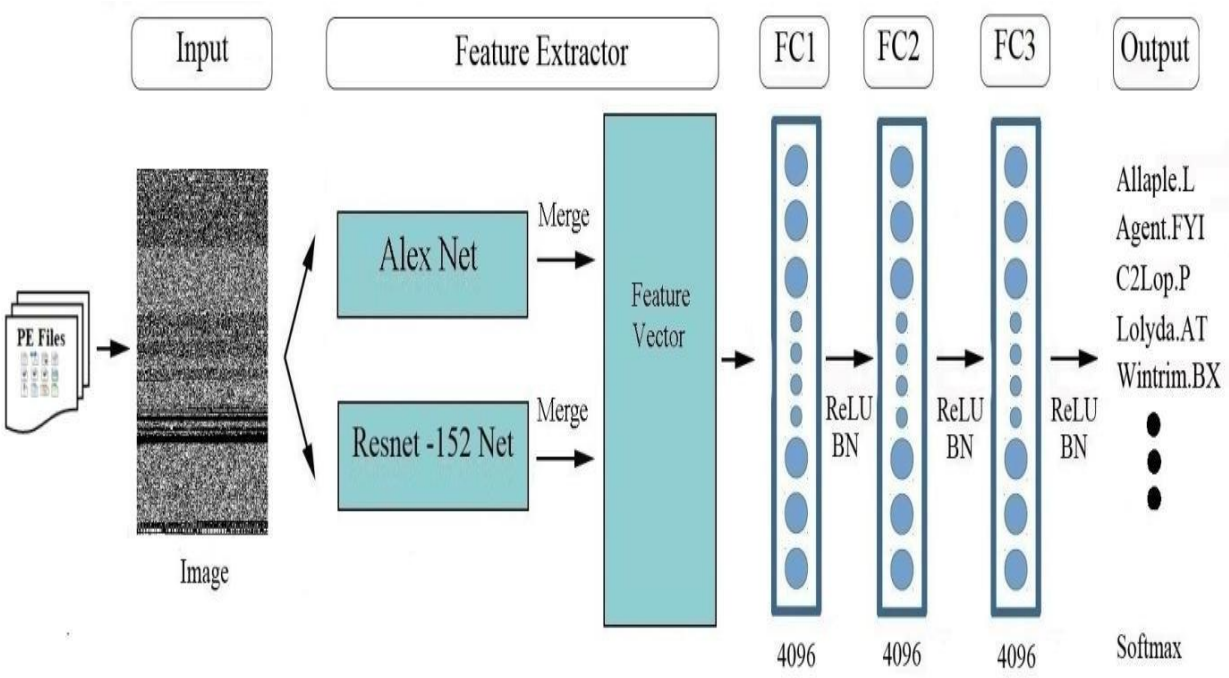

FIGURE 1.Proposed malware classification methodology.

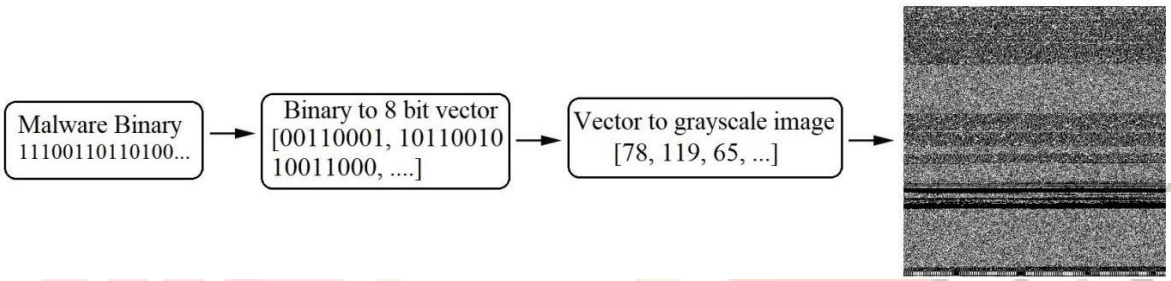

FIGURE 2:Overview of malware visualization process.

Malware data is initially gathered from a variety of sources, including Malimg , Microsoft BIG 2015 , and Malevis. The next section goes into the specifics of these malware classification datasets. The proposed deep neural network architecture is then constructed. Two stages of pre-processing are carried out here: To begin, the process of projecting a suitable DL architecture for use in malware classification processes was completed. In this paper, a hybrid module containing ResNet-50 and AlexNet architectures was developed using pre-trained architectures, since it was discovered in pre-experiments that a hybrid module can provide preferred overall precision.

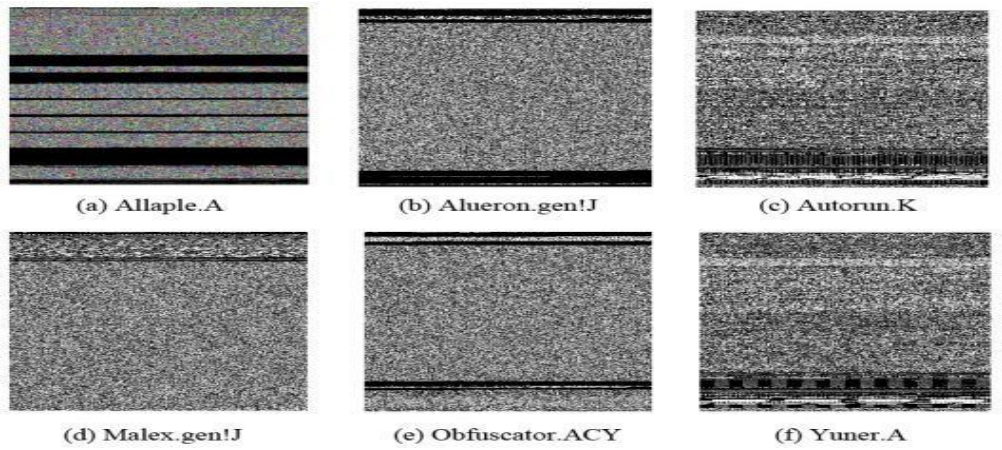

FIGURE 3.Malwaregrayscaleimagesfromdifferentmalwarefamilies.(These images are obtained from Malimg, Microsoft BIG 2015 and Malevis Datasets.

Figure 6 shows the ResNet-50 architecture, which was a winning model in the ILSVRC 2015 and COCO 2015 competitions. It is a 50-layer convolutional neural network. Five convolutional blocks are employed in this network model, consisting of 1 1, 3 3, and 1 1 convolution layers.

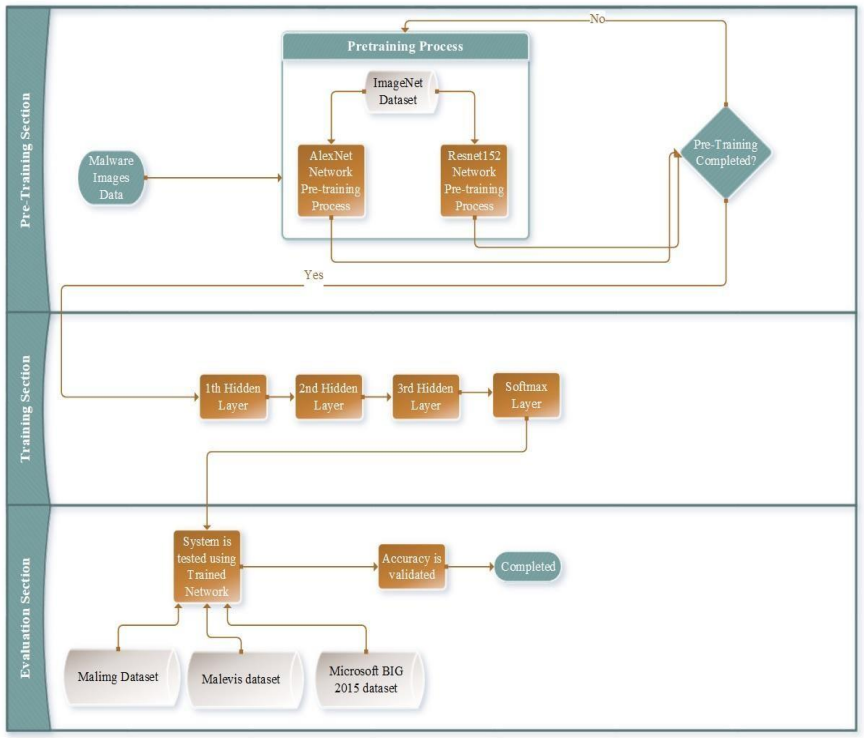

FIGURE 4.Flow chart of proposed deep learning architecture for malware classification.

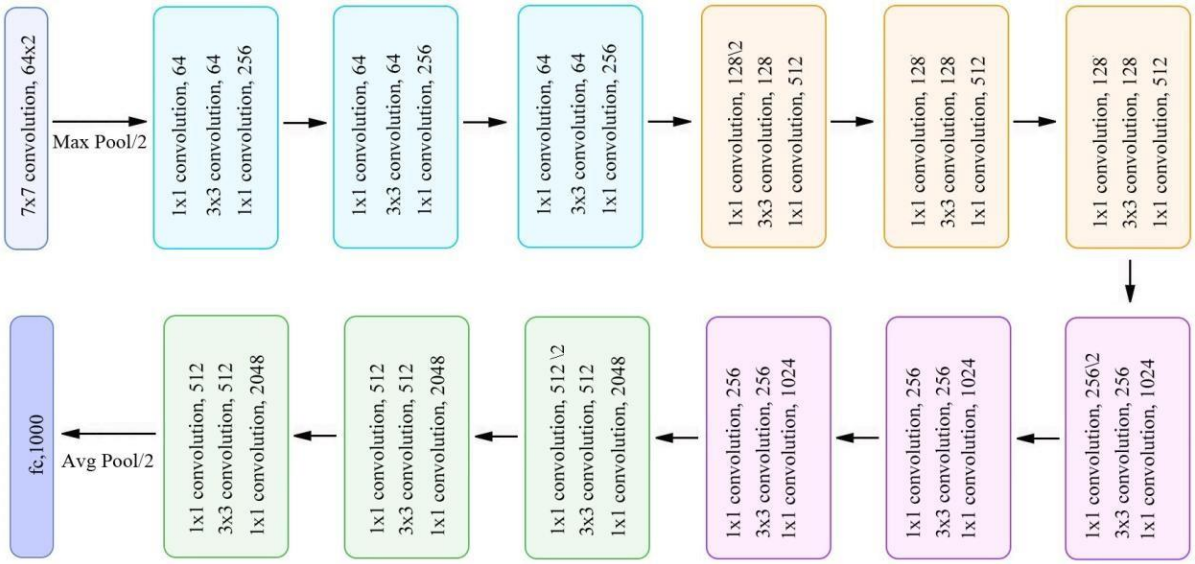

FIGURE 5:AnIllustration of ResNet-50 Network

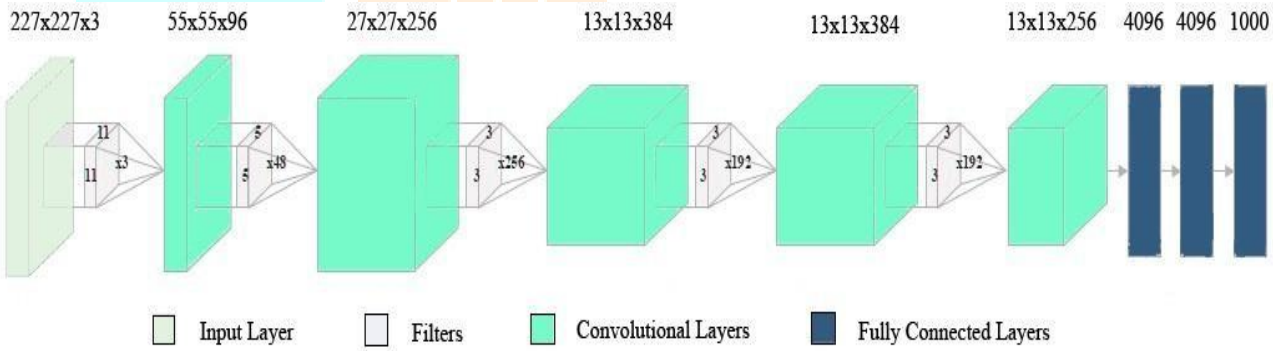

FIGURE 6.An Illustration of AlexNet Network

The first five layers are convolutional layers, and the last three layers are fully connected layers. In addition, as illustrated in Figure 6, the AlexNet network incorporates two normalisation, three pooling, seven ReLU layers, and a softmax layer for the learning and classification operations.

Then, in order to overcome the various problematic situations encountered in the classification process, such as time constraints, extreme dataset dimension, and so on, transfer learning algorithms were examined. The feature extraction method is carried out utilising pre-trained networks in the transfer learning approach. The classification process is then completed with a general classifier such as a support vector machine or softmax as the final step. This technique is adapted to the planned architecture in order to deal with the aforementioned demanding conditions. ImageNet dataset was used to determine which Resnet and AlexNet architectures were trained. The features obtained from the Resnet and AlexNet designs are then concatenated in the second stage to form a feature vector. This 4096-dimensional vector was produced. On each grayscale image frame, the features created by the pre-trained architectures of ResNet-50 and AlexNet are respectively the extraction of a 2048 dimensional feature from the final fully connected layer, as shown in Figures 5 and 6. After obtaining the combined feature vector with 4096 elements, the model's pre-training stage is completed. Finally, to achieve normalisation, the combined feature vector is fed into the softmax layer and fully connected layers. 4096 nodes seized Thesoftmax layer contains 57 outputs that correspond to 57 malware types, while the fully linked layers have 4096 nodes.

The goal of this layer is to improve the proposed network's learning capability (Figure4). Finally, utilising the exhaustive datasets as inputs to the trained model, experimental analysis of the proposed model was done.

IV EXPERIMENTAL RESULTS IS AND DISCUSSIONS

The implementation details, experimental findings, and evaluation of the suggested deep neural network model are all explained in this part. Our tests were conducted in a Linux environment with an Intel Core i9 processor running at 4.8 GHz and 32 GB of RAM memory. The Python programming language was used to implement the proposed design. Each data set's training, validation, and test data were chosen at random, and the assessment processes were carried out one by one. The selection rates of the available data for the training, validation, and testing stages are set at 70%, 10%, and 20%, respectively.

V. CONCLUSION

Despite extensive research on malware detection and categorization, detecting malware variants successfully remains a severe problem in the cyber security area. The identification of malware is difficult due to code obfuscation and packaging techniques. This paper offered a new deep learning architecture for detecting malware variants effectively. A hybrid architecture is proposed in the proposed architectureapproach. This method relies on the transfer learning method and contains numerous exhaustively pre-trained networks. Initially, malware data was gathered by combining multiple large databases. The features are then retrieved with the help of pre-trained networks. Finally, a supervised learning method is used to execute the training phase of deep neural network design.

REFERENCES

1. Aslan and R. Samet, "A Comprehensive Review on Malware Detection Approaches," in *IEEE Access*, vol. 8, pp. 6249-6271, 2020, doi: 10.1109/ACCESS.2019.2963724.
2. J. Jeon, J. H. Park and Y. -S. Jeong, "Dynamic Analysis for IoT Malware Detection With Convolution Neural Network Model," in *IEEE Access*, vol. 8, pp. 96899-96911, 2020, doi: 10.1109/ACCESS.2020.2995887.
3. E. M. Alkhateeb and M. Stamp, "A Dynamic Heuristic Method for Detecting Packed Malware Using Naive Bayes," *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2019, pp. 1-6, doi: 10.1109/ICECTA48151.2019.8959765.
4. A. Yilmaz, M. S. Guzel, E. Bostanci and I. Askerzade, "A Novel Action Recognition Framework Based on Deep-Learning and Genetic Algorithms," in *IEEE Access*, vol. 8, pp. 100631-100644, 2020, doi: 10.1109/ACCESS.2020.2997962.
5. S. A. Roseline, S. Geetha, S. Kadry and Y. Nam, "Intelligent Vision-Based Malware Detection and Classification Using Deep Random Forest Paradigm," in *IEEE Access*, vol. 8, pp. 206303-206324, 2020, doi:
6. S. Bozkir, A. O. Cankaya and M. Aydos, "Utilization and Comparision of Convolutional Neural Networks in Malware Recognition," *2019 27th Signal Processing and Communications Applications Conference (SIU)*, 2019, pp. 1-4, doi: 10.1109/SIU.2019.8806511.
7. S. Sriram, R. Vinayakumar, V. Sowmya, M. Alazab and K. P. Soman, "Multi-scale Learning based Malware Variant Detection using Spatial Pyramid Pooling Network," *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2020, pp. 740-745, doi: 10.1109/INFOCOMWKSHPS50562.2020.9162661.
8. X. Ma *et al.*, "How to Make Attention Mechanisms More Practical in Malware Classification," in *IEEE Access*, vol. 7, pp. 155270-155280, 2019, doi: 10.1109/ACCESS.2019.2948358.
9. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran and S. Venkatraman, "Robust Intelligent Malware Detection Using Deep Learning," in *IEEE Access*, vol. 7, pp. 46717-46738, 2019, doi:

[1] S. Luo and D. C. -T. Lo, "Binary malware image classification using machine learning with local binary pattern," *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 4664-4667, doi: 10.1109/BigData.2017.8258512.