International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

Intellectual Feature Ranking Model with Correlated Feature Set based Malware Detection in Cloud Environment using Machine Learning

Sanaboyina Madhusudhana Rao[1], Arpit Jain[2], PVSS Gangadhar[3], Vinay Sowpati[4]
[1]Research Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
Email: madhusudhanarao.s@nic.in
[2]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
Email: dr.jainarpit@gmail.com
[3]Scientist-E, NIC, Meity,GoI,Vijayawada,AP,India.
Email: pvss.gangadhar@nic.in
4Scientist-D, NIC, Meity,GoI,Vijayawada, AP,India
Email: vinay.sowpati@nic.in

Abstract—Malware detection for cloud systems has been studied extensively, and many different approaches have been developed and implemented in an effort to stay ahead of this ever-evolving threat. Malware refers to any programme or defect that is designed to duplicate itself or cause damage to the system's hardware or software. These attacks are designed specifically to cause harm to operational systems, but they are invisible to the human eye. One of the most exciting developments in data storage and service delivery today is cloud computing. There are significant benefits to be gained over more conventional protection methods by making use of this fast evolving technology to protect computer-based systems from cyber-related threats. Assets to be secured may reside in any networked computing environment, including but not limited to Cyber Physical Systems (CPS), critical systems, fixed and portable computers, mobile devices, and the Internet of Things (IoT). Malicious software or malware refers to any programme that intentionally compromises a computer system in order to compromise its security, privacy, or availability. A cloud-based intelligent behavior analysis model for malware detection system using feature set is proposed to identify the ever-increasing malware attacks. The suggested system begins by collecting malware samples from several virtual machines, from which unique characteristics can be extracted easily. Then, the malicious and safe samples are separated using the features provided to the learning-based and rule-based detection agents. To generate a relevant feature set for accurate malware detection, this research proposes an Intellectual Feature Ranking Model with Correlated Feature Set (IFR-CFS) model using enhanced logistic regression model for accurate detection of malware in the cloud environment. The proposed model when compared to the traditional feature selection model, performs better in generation of feature set for accurate detection of malware.

Keywords- Malware Detection, Cloud Environment, Cyber Physical Systems, Feature Extraction, Feature Selection, Correlated Feature Set.

I. INTRODUCTION

In order to penetrate and potentially destroy a computer system without the owner's knowledge, malicious software was developed. Malware is a catch-all term for any harmful software installed on a computer [1]. Malware can be broken down into two categories: those that infect files and those that can function independently. Malware can also be categorized by the harm they cause, such as worms, backdoors, trojans, rootkits, spyware, adware, and so on. It is becoming increasingly difficult to detect malware using traditional, signature-based methods, as all existing malware applications typically employ multiple polymorphic layers to evade detection, or use auxiliary mechanisms to automatically update to a newer version after relatively short intervals [2].

Instead of maintaining and managing own server, IT infrastructure, or data centres, can take advantage of the scalability and flexibility of a cloud computing platform [3]. Since Cloud computing resources are increasingly being offered as a remote service to consumers, hackers are increasingly trying to steal internet data without having to pay for the privilege [4]. Malware is a common source of disruption in the form of data leaks, downtime for a company website, and ransomware and WannaCry attacks.

Organizations and Internet users continue to face a significant challenge in the form of the rising frequency and severity of cyber attacks. By 2025, experts expect that cybercrime will cost the global economy over $8 trillion annually. Furthermore, it is taking longer to handle cyber attacks and costing organizations far more than before, with attacks involving malware being the most expensive at around $3.2

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

million per attack. Malware is any programme that is designed to cause harm when run on a computer. It's easy to set up automatically and from a distance, making it a potential attack vector [5]. Over the past few decades, malware has become increasingly diverse, stealthy, and complicated, rendering it undetectable by traditional antimalware methods. In 2019, Kaspersky found a total of 24,610,126 distinct pieces of malware, an increase of 14% from 2018. Hackers have resorted to developing malware that can take down entire networks because of the huge financial rewards it could bring. Due to malware's constant development and variety, an effective malware detection strategy is essential for protecting businesses [6].

Over the years, several different methods for detecting malware have been developed, with a growing emphasis on using Machine Learning (ML) methods [7]. A signature-based method, a behavioural approach, a heuristic approach, and a model-based approach have all emerged as a result of research into malware detection [8]. Both static analysis-based and dynamic analysis-based technologies exist for detecting malware [9]. Malicious files can be analyzed in both static and dynamic analysis approaches, but only the latter requires the files to be executed. Malicious files should only be run in an isolated setting. Static analysis can quickly determine whether a file is dangerous or not, so long as it is not encrypted or compressed [10]. However, since dynamic analysis occurs in real time, packing approaches have little to no impact on how files are analyzed as they are being run. However, modern malware can often tell whether it's in a virtual environment and will often simply cease running if it does. Because malware may only run under specific conditions, it is often impossible to obtain data on how it acts. The process of analyzing malware in cloud environment is shown in Figure 1.

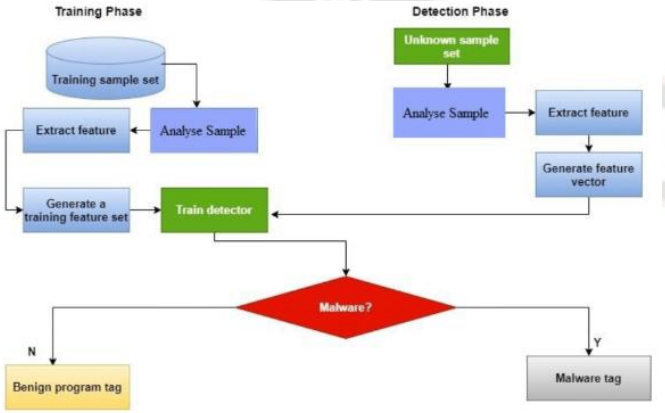

Fig 1: Malware Analysis in Cloud

The method of identifying malicious software by analyzing it for a set of features is known as malware detection. There is an urgent need to automatically and effectively detect malware so that it can be swiftly removed from systems. Because it is an NP-complete issue, detecting malware is a formidable challenge [11]. As a result of this difficulty, researchers have set out to develop better methods for identifying malicious software. Signatures derived from malware instances via reverse engineering form the basis of traditional malware detection techniques. Because malware authors regularly update their malicious signatures, these methods are now largely ineffective [12]. With the alarming rise of malware, the difficulty of detection has also increased significantly. Malware authors use a wide variety of methods to avoid being discovered. Despite the fact that a number of novel malware detection strategies employing machine learning have been developed, the accuracy of malware detection could still be enhanced [13]. The parameters used, the malware analysis procedure implemented, and the features employed can all have an impact on how well a machine learning technique performs [14]. As a result, one of the most pressing challenges in cyber security is the development of a reliable malware detection system by selecting the most relevant features. The malware detection process in cloud environment is shown in Figure 2.

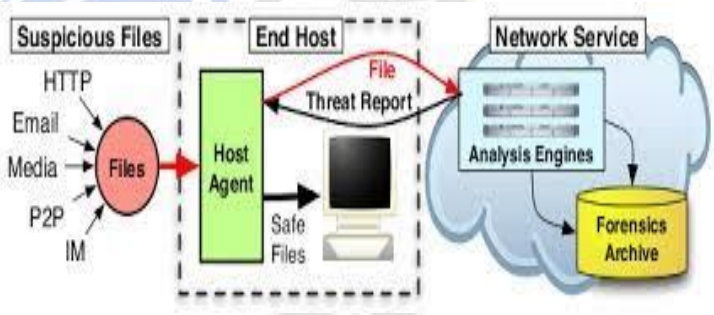

Fig 2: Cloud Environment Malware Detection

By sending malware detection tasks to security servers with a larger malware database and powerful computational capabilities, cloud-based malware detection improves the accuracy of detection for mobile devices [15]. Each mobile device decides its own offloading rate of application traces to the security server in the dynamic detection of malware amusement, leading to conflict between radio transmission bandwidths and data sharing with the secure server. Data sharing during the Cloud's security stage is used to build an ML based malware detection model [16]. At that time, depending on the characteristics of models propagation, the exact number of malware infected nodes that have physical infectivity to vulnerable nodes is estimated [17].

In machine learning, logistic regression is a well-liked classification algorithm employed frequently in the context of feature selection. It is a quick and easy method for extracting

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

useful information from a dataset and creating a classification model for future data [18]. Using logistic regression, one can estimate the likelihood that a given observation corresponds to a specified category. For malware prediction, for instance, the model will return a probability between 0 and 1 that indicates the possibility that a cloud environment has malware. Choosing the features that will ultimately be used to make predictions about the target variable is the first step in developing a logistic regression model [19]. Feature selection is a crucial part of the modeling process since it affects the model's accuracy and precision. One wrapper approach for categorical data is logistic regression [20]. The addition and regularizations to logistic regression was also implemented. The combination of the wrapper and filter techniques results in a hybrid or embedded approach. Further, it was discovered that selecting features for high-dimensional datasets in parallel could lessen the computing load of logistic regression [21].

Although it does a reasonable job of classifying data, logistic regression suffers from a lack of computing efficiency [22]. This is because it typically uses an optimization technique belonging to the hill-climbing Newton family, such as Newton-Raphson and its variants and extensions, to arrive at an estimate. Therefore, it appears that the conventional logistic regression has difficulty managing a large range of features [23]. When building a machine learning model, feature selection is used to narrow down the data points to a manageable collection. Improving the model's accuracy and decreasing the computational cost of training it make this a crucial step in the model-building process [24]. To overcome the limitations of logistic regression model, enhanced logistic regression model is designed in this research. The best parameters are chosen and sent on to the risk prediction model using an enhanced sparse logistic regression approach [25]. This regression strategy refines the model by adjusting the sparsity component with a logistic loss function. To generate a relevant feature set for accurate malware detection, this research proposes an Intellectual Feature Ranking Model with Correlated Feature Set model using enhanced logistic regression model for accurate detection of malware in the cloud environment.

II. LITERATURE SURVEY

More efficient low-latency integration of AI and cloud computing within the framework of sophisticated and intelligent IIoT ecosystems is made possible by 5G, making the entire industrial process more efficient. In addition, it poses serious security and privacy problems because it raises the underlying control system's functional complexity and introduces additional powerful attack vectors. The importance of security and privacy in this environment is illustrated by the fact that malware assaults have begun focusing on vulnerable but highly connected IoT devices. Ahmed et al. [1] designed a 5G-enabled system that uses deep learning to categorize malware attacks on the IIoT. This approach makes use of convolutional neural networks trained to distinguish between distinct types of malware attacks and a visual representation of the virus. The proposed architecture combines many layers to extract complimentary discriminative features, with a 97% success rate.

The number of apps designed specifically for smart devices or Android-based IoT has skyrocketed in recent years. Meanwhile, the number of malicious or poorly made apps has been increasing at an alarming rate. Many methods have been developed in recent years to tell malicious programmes from legitimate ones, helping to make app stores safer for users. Classification is a common use of machine learning. The success of detection using machine learning algorithms is dependent on an accurate description of app behaviour, or characteristics. Apps for Android are always changing. As time has progressed, app sizes have grown and complexity of app behaviour has increased. Therefore, it remains difficult to effectively and reliably extract representative information from apps. Wang et al. [2] characterized the behaviours of apps with different sorts of features in order to detect malapps, and to present a clear and thorough overview of the state-of-the-art work in this area.

Many businesses today depend on cloud computing to power their operations. With these services, firms can save costs associated with hardware management, scalability, and maintenance. Amazon Web Services, Microsoft Azure, and Google Cloud Platform are just a few of the leading CSPs that provide Infrastructure as a Service to businesses like these. With so many people using the cloud, it's no wonder that keeping data safe in the cloud is a major issue for CSPs. Malware is widely acknowledged as one of the most severe and pervasive dangers to IaaS in the cloud. In this paper, Kimmel et al. [3] investigated the efficacy of deep learning algorithms based on Recurrent Neural Networks (RNNs) for identifying malware in cloud VMs. In particular, the Long Short-Term Memory RNN (LSTM) and the Bidirectional RNN (BIDI) architectures of RNNs are examined. Malware behaviour is learned by these models over time by observing system parameters like CPU, memory, and disc utilization at runtime during malicious attacks. Using a dataset consisting of 40,680 malicious and benign samples, the author tested this method. In order to capture the true behaviour of stealthy and complex malware, the process stage features were collected by executing actual malware in an unrestricted internet cloud environment.

The rapid development of the industrial Internet has led to the emergence of cloud service as a cutting-edge industrial norm with game-changing possibilities for the business world. Many

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

businesses these days see the value in adopting cloud-based service models. Security risks, such as covert virus attacks on virtual domains, are a serious worry, though. To combat malware in cloud infrastructure, Mishra et al. [4] offered an introspection-based security method for protecting virtual domains within a cloud-based service platform. VMShield prevents malware from evading the security tool by performing virtual memory inspection from the hypervisor to collect the run-time behaviour of processes. The suggested method outperforms both static and dynamic state-of-the-art techniques, which are currently in use, because of its reliance on introspection. The VMShield uses the Bag of n-gram method to extract features from system calls and the meta-heuristic methodology binary particle swarm optimization to choose relevant features. By using a Random Forest (RF) classifier to divide monitored programmes into good and bad processes, it is possible to detect variations of malware, providing an improvement over the standard signature-matching method.

One of the most exciting developments in data storage and service delivery today is cloud computing. There are significant benefits to be gained over more conventional protection methods by making use of this fast evolving technology to shield computer-based systems from cyber-related threats. Assets to be safeguarded may reside in any networked computing environment, including but not limited to CPS, critical systems, fixed and portable computers, mobile devices, and the IoT. Malicious software refers to any programme that intentionally compromises a computer system in order to compromise its security, privacy, or availability. Aslan et al. [5] proposed a cloud-based intelligent behavior-based detection system to identify the ever-increasing malware attack surface. The suggested system begins by collecting malware samples from several virtual machines, from which unique characteristics can be extracted easily. Then, the malicious and safe samples are separated using the features provided to the learning-based and rule-based detection agents. The effectiveness of the suggested approach was assessed by analyzing 10,000 samples of code. The suggested method has a high rate of detection and accuracy for both known and undiscovered malware.

In light of recent advancements in computer systems, people are increasingly spending more time in simulated worlds. This process has been sped up by the Covid-19 diseases. The focus of cybercriminals has switched from the actual world to the online world. This is due to the fact that it is much less difficult to engage in criminal activity online compared to the real world. Cybercriminals routinely employ malicious software in order to initiate cyber attacks. Advanced obfuscation and packing techniques are being included into new malware versions. Malware identification and categorization are greatly complicated by these methods of concealment. To effectively tackle emerging malware strains, novel ways that are very different from conventional methods must be applied. Traditional ML methods, a subset of AI, have become insufficient to detect all new and complicated malware types. A promising potential solution to the problem of detecting all forms of malware is the deep learning technique, which is considerably distinct from conventional ML algorithms. In this research, Aslan et al. [6] presented a novel deep-learning-based architecture for malware classification that makes use of a hybrid model. The primary result of this research is a hybrid architecture for optimally combining two diverse pre-trained network models. Acquisition of data, design of deep neural network architecture, learning of the proposed architecture, and evaluation of the trained network are the four primary phases of this architecture.

Edge devices are used in the IIoT to mediate communications between sensors and actuators and back-end systems like application servers and cloud storage. Many organizations have turned to machine learning models to protect their edge devices from infection. These models, however, are susceptible to adversarial assaults, in which malicious actors deliberately alter samples of malware in order to trick a classifier into incorrectly labelling them as benign. As a countermeasure, adversarial retraining is proposed in the literature on deep learning networks, wherein malicious samples are mixed in with good ones to retrain the classifier. However, current methods randomly select these hostile examples, which is bad for the classifier's accuracy. Two new methods for choosing adversarial data sets to retrain a classifier are proposed by Khoda et al. [7]. Both the distance from the malware cluster centre and a kernel-based learning (KBL) probability measure are used. Both of the sample selection approaches exceed the random selection method, and the KBL selection method enhances detection accuracy by 6%, as demonstrated by the trials. While most prior works on adversarial retraining have focused on deep neural networks, the author also investigated the effects on these samples have on other classifiers and find that the proposed selective competitive retraining strategies improve performance for these classifiers as well.

As IoT devices become more commonplace in everyday life and the workplace, they become more of a target for cybercriminals. In this work, Haddadpajouh et al.[8] proposed using the grey wolves optimization (GWO) technique to train a multi kernel support vector machine (SVM) to detect malware in IoT cloud-edge gateways. This metaheuristic method is used to pick attributes at the IoT cloud-edge gateway that most effectively differentiate between malicious and benign apps. Opcode and Bytecode of IoT malware

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

samples are used to train the model, and the model is then tested with K-fold cross-validation. The training data set includes 271 benign and 281 malicious Cortex A9 samples. The author verified the proposed model's robustness by testing its ability to identify samples of IoT malware that have never been seen before. By combining the RBF with polynomial kernels, that was able to attain an accuracy of 99.72 percent. The training time for the suggested model is only 20 seconds, while the training time for the prior deep neural network (DNN) model was over 80 seconds.

III. PROPOSED MODEL

Cyber related attacks have increased in frequency and intensity. Cyber attacks are typically caused by various forms of malware. Malware, or malicious software, is defined as any programme with the intent to cause harm or financial gain by taking advantage of a system's or the Internet's weaknesses. Malware comes in many forms, the most common of which include viruses, worms, Trojan horses, backdoors, rootkits, and ransomware. Different goals were in mind when creating each family of malicious malware. Complex attacks typically employ a wide variety of malware. There has been a steady rise of malware variants throughout the years. Business and academic estimates estimate that every day, over 1 million new dangerous software variations are created. The vast majority of these malicious programmes are just new iterations of old ones. Malware-related attacks in the virtual world are on the rise due to a number of factors, including the proliferation of Internet of Things devices, the rapid development of new applications, and the sheer volume of information generated daily by social media platforms. However, modern malware uses advanced hiding tactics like obfuscation and packaging to avoid being detected. This renders traditional malware detection methods essentially useless for identifying and categorizing sophisticated malware.

Identifying whether a piece of software is malicious or safe is what malware detection is all about. A wide variety of both time-honored and novel techniques for finding malware have been detailed. Signature-based, heuristic-based, behavior-based, and model-checking-based methods are some of the more traditional methods of detection. While the signature-based detection method has shown effective against previously identified malware and its variants, it has proven unable to identify previously unknown malware due to the latter's unique signature. Some zero-day malware may be detectable using behaviour, heuristic, and model checking-based techniques. However, they are unable to identify malware that employs modern compression methods. There are a number of advantages to using cloud computing to identify malware. The cloud computing environment reduces expenses while increasing convenience, storage space on demand, computing power, and the size of databases. Using various VMs and servers, multiple execution traces of the same virus have been gathered. The detection performance is enhanced while the false positive and false negative rates are reduced when multiple algorithms are used.

The process of manually analyzing malware and extracting features is labor-intensive and time-consuming. A system that can automatically analyze the malware and extract features is, therefore, desperately needed. Malware, in order to conceal its true behaviours, performs both interrelated and unconnected tasks. This highlights the importance of identifying interconnected processes and extracting genuine features during the dataset development phase.

Malware poses a significant risk to cloud hosting systems. Static malware detection is a common technique that involves examining an executable's signature and comparing it to a library of known malicious signatures. In order to make static analysis less useful, attackers have used obfuscation and packing. Furthermore, static malware analysis cannot detect the constantly growing zero-day malware because it is restricted to analyzing only known malware executables. Because of these two main drawbacks, researchers have focused heavily on dynamic behavioural malware detection techniques. Malware can be detected both dynamically and online using behaviorally based techniques.

Dynamic malware detection techniques examine the behaviour of malicious programmes by executing them in a safe environment. This allows the detection system to analyze unique zero-day malware without relying on previously known signatures, instead using the executable's real behaviour. However, malicious software has evolved to recognize when defensive measures are being taken, such as with a sandbox, and to stop its malevolent behaviour. Malware can easily circumvent these rudimentary detection methods by exploiting a system's vulnerabilities. In order to safeguard a computer from malware, online malware detection analyses the system's activity. Online approaches monitor the performance of the entire virtual machine and trigger an alarm if any signs of malicious behaviour are identified at any time, as opposed to analyzing executables and their behaviour. Because of this, malware detection techniques implemented in the cloud are superior to traditional static and dynamic malware detection methods and qualify as continuous real-time detection systems. The proposed model framework is shown in Figure 3.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

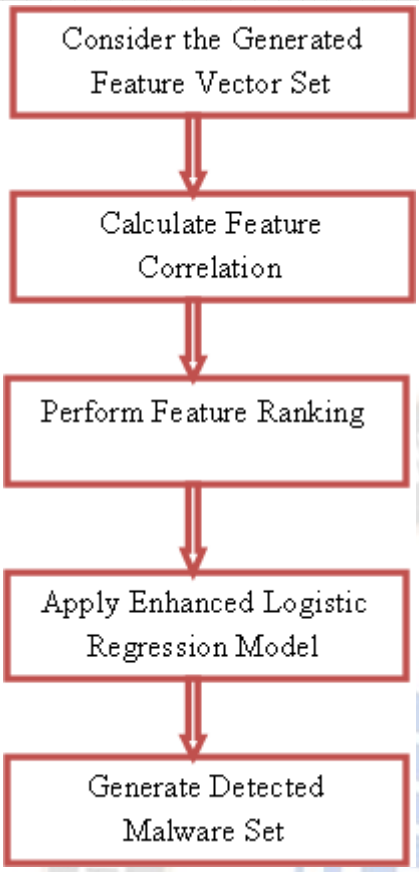

Fig 3: Proposed Model Framework

One of the most well-known Machine Learning algorithms, logistic regression belongs to the category of Supervised Learning. The categorical dependent variable can be predicted from a given collection of independent factors with this method. Results for a variable that is categorical can be predicted using logistic regression. This means that the result must take the form of a categorical or discrete number. Probabilistic values between 0 and 1 are provided instead of the exact values of Yes and No, 0 and 1, true and false, etc. The main difference between logistic regression and linear regression is in the application. Problems of regression are best tackled using linear regression, whereas those of classification with logistic regression.

Instead of a straight line, the "S" shaped logistic function with its two extreme predictions of 0 and 1 is fitted in logistic regression. If the cells are malignant, the logistic function curve will be skewed upwards; if a mouse is overweight, the curve will be skewed downwards. Since it can classify new data using both continuous and discrete datasets, Logistic Regression is an important machine learning approach. Logistic regression is useful for modeling and hypothesis testing with regards to associations between a categorical result and a set of predictors that can be either categorical or continuous. Creating categories for the predictor and computing the mean of the outcome variable for the categories in question are viable alternatives to using an ordinary least squares regression equation to characterize the two parallel lines. Sigmoidal or S-shaped shapes are notoriously challenging to explain with a linear equation for two main reasons. To begin, there is no linear trend between the two extremes. A second issue is that the errors do not follow a normal distribution or remain the same over different intervals of time or different datasets. The logit transformation of the dependant variable provides the answer in the context of logistic regression. The logit of Y is predicted from X using the logistic model. To generate a relevant feature set for accurate malware detection, this research proposes an Intellectual Feature Ranking Model with Correlated Feature Set model using enhanced logistic regression model for accurate detection of malware in the cloud environment. To generate a relevant feature set for accurate malware detection, this research proposes an Intellectual Feature Ranking Model with Correlated Feature Set (IFR-CFS) model using enhanced logistic regression model for accurate detection of malware in the cloud environment.

Algorithm IFR-CFS

{

Input: Feature Vector Set (FVset}

Output: Detected Malware Set {Malset}

Step-1: The features selected are extracted from the feature set and these features and the attribute ranges are analyzed. The selected features are used for analyzing the dissimilarities in attributes and then malware prediction is performed. The process of feature vector set analysis is performed as

$$FeatVec[N] \sum_{i=1}^{N} \frac{getmax(AttrSimm(i,i+1))}{\gamma} + \max\left(corr\left(AttrSimm(i,i+1)\right)\right) - \min\left(corr(AttrSimm(i,i+1)\right) + \frac{len(AttrSimm)}{\gamma}$$

$$FA[N] = \sum_{i=1}^{N} \max\left(FeatVec(i)\right) + \max\left(Corr\left(FeatVec(i,i+1)\right)\right) - \min\left(FeatVec(i)\right)$$

Here, getmax() is used to retrieve the maximum values from the feature set, corr() is used to identify the correlation factor among the features, γ is the total attributes count. len() is used to find the length of the feature set.

Step-2: When comparing the linearity of relationships between multiple features, correlation is the high standard.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

One feature can be predicted from another through the use of correlation. The idea behind utilizing correlation to choose features is that useful variables will have a high degree of correlation with the end result. It is typical for two features designed to assess distinct properties to be affected by the same cause and to exhibit correlated variation. The feature correlation calculation among the feature set is performed as

$$Fcorr[M] = \sum_{i=1}^{M} \frac{\max\left(FA(i), FA(i+1)\right)}{len(FeatVec)} + \sum_{i=1}^{M} \frac{(x - x^{'})(y - y^{'})}{(M-1)\sigma a * \sigma b}$$

$$x^{'} = \sum_{i=1}^{M} \frac{\sum_{i=1}^{M} x}{M}$$

$$y^{'} = \sum_{i=1}^{M} \frac{\sum_{i=1}^{M} y}{M}$$

$$\sigma a = \sum_{i=1}^{M} \sqrt{\frac{(x - x^{'})^2}{M-1}}$$

$$\sigma b = \sum_{i=1}^{M} \sqrt{\frac{(y - y^{'})^2}{M-1}}$$

Here x and y are the two features set considered for correlation calculation, x' and y' is the mean of feature set and σa and σb are the standard deviation of feature set.

Step-3: The feature correlation model generates a correlation range of every two features and these values range from highly correlated to low correlated. The feature ranking is the process of selecting the best correlated features that are independent with each other. The feature ranking process is performed as

$$FeatRank[M] = \sum_{i=1}^{M} \frac{\max\big(Fcorr(i, i+1)\big) + \max\big(simm(i, i+1)\big)}{len(Fcorr)}$$

$$\begin{cases} FeatRank(i) \leftarrow k++ & if Fcorr(i, i+1) > Th \\ FeatRank(i) \leftarrow 0 & Otherwise \end{cases}$$

Here Th is the threshold value of correlation range. k is the flag value initially considered as 1. simm() model is used to identify the similarity levels of the features set.

Step-4: Using an improved sparse logistic regression method, the optimal parameters are selected and forwarded to the malware prediction model. Using a logistic loss function, this regression method fine-tunes the model's sparsity. The enhanced logistic regression model is applied by considering the activation function as

$$ActFun[M] = \sum_{i=1}^{M} \frac{1}{1 + e^{-x}}$$

$$LogF[M] = \sum_{i=1}^{M} \frac{e^{(p+qx)}}{1 + e^{(p+qx)}}$$

$ElogF[M] = \sum_{i=1}^{M} \mu\big(ActFunc(i, i+1)\big) +$ max $(LogF(i))$+δ(max(FeatRank(i,i+1))

Here e is the natural logarithms base and x is the transformed value, LogF is the newly predicted value, p is the bias value and q is the coefficient. μ is the probabilistic function, δ is the tan function of the activation function.

Step-5: The enhanced logistic regression model performs the classification of the nodes attributes and then the nodes causing malware and affected with malware are detected and can be removed from the network for enhancing the security. The malware detection set is generated as

$$Malset[M] = \sum_{i=1}^{M} simm\big(Elogf(i, i+1)\big) + \max\big(FeatRank(i, i+1)\big) + \max\left(ActFun(i, i+1)\right)$$

$$\begin{cases} Malset \leftarrow 1 \; if \; simm(ElogF) < Th \; and \; FeatRank < V \\ Malset \leftarrow 0 & Otherwise \end{cases}$$

Here Th is the threshold logarithmic function and V is the threshold rank of the features.

}

IV. RESULTS

It is becoming increasingly difficult to identify malware in the cloud as cloud computing technology evolves. Malware programmes constantly update their source code as they spread over the cloud. However, many current relevant studies have concentrated on malware detection accuracy without giving adequate consideration to privacy protection of cloud users. Malware polymorphism and hostility present a challenge to traditional antivirus technologies based on signature scanning. Malware comes in many forms, the most common of which include viruses, worms, Trojan horses, backdoors, malware, ransomware, bots, and rootkits. Because identical functionality may be achieved by reusing large portions of code, attackers can create many variants of the same virus with just minor structural changes. While many malware detection systems have put their attention on machine learning techniques in an attempt to improve their detection rates, these methods have so far proven ineffective.

Most cloud users may not grant antivirus service providers access to private files due to privacy concerns, making malware removal function extraction unfeasible. While most signatures are derived from a static analysis of the code within an executable file, malware authors frequently avoid static

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

analysis by making obfuscation-based changes to their virus's code. The device under scrutiny has unreliable antivirus software. Only if the inspected instance is unaware of the inspection process and cannot affect the inspection process is the inspection mechanism regarded trustworthy. While some top-tier antivirus programmes use virtual environment simulation and file behaviour checks for dynamic analysis, sophisticated malware is still able to spot and avoid detection. To generate a relevant feature set for accurate malware detection, this research proposes an Intellectual Feature Ranking Model with Correlated Feature Set (IFR-CFS) model using enhanced logistic regression model for accurate detection of malware in the cloud environment. The proposed model is compared with the traditional Multilayer Deep Learning Approach for Malware Classification in 5G-Enabled IIoT (MDLA-MC) and Constructing Features for Detecting Android Malicious Applications (CFDAMA). The proposed model when compared to the traditional models performs better in feature selection for accurate malware detection.

The proposed model performs loading of dataset and analyzing the records for detection of attributes variations in the dataset. This analysis is used to identify multiple patterns so that malware can be easily detected. The Dataset Record Analysis Time Levels of the proposed and existing models are shown in Table 1 and Figure 4.

Table 1: Dataset Record Analysis Time Levels

Records Considered	Models Considered		
	IFR-CFS Model	MDLA-MC Model	CFDAMA Model
10000	10.8	15.8	14.3
20000	11	16.2	14.8
30000	11.2	16.7	15
40000	11.5	17	15.3
50000	11.7	17.5	15.6
60000	12	18	16

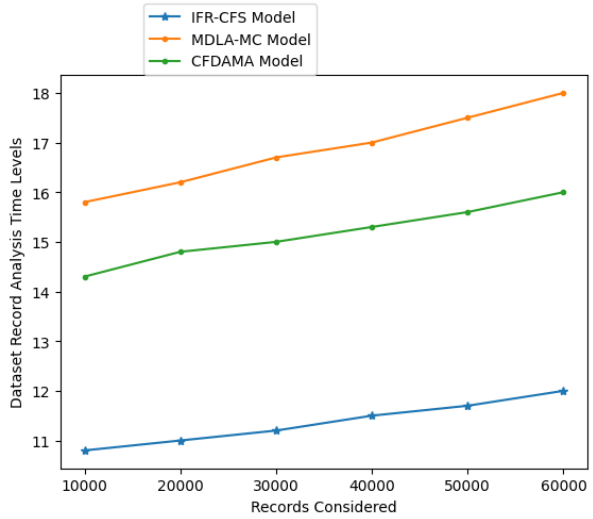

Fig 4: Dataset Record Analysis Time Levels

By extracting new features from an existing collection of features and deleting the set's original features, Feature Extraction can be used to reduce the dataset's feature count. These new, streamlined features should be able to effectively summarize the original features' worth of data. By combining the original features in this way, a condensed version of the full set can be made. The Feature Extraction Accuracy Levels of the proposed and traditional models are depicted in Table 2 and Figure 5.

Table 2: Feature Extraction Accuracy Levels

Records Considered	Models Considered		
	IFR-CFS Model	MDLA-MC Model	CFDAMA Model
10000	96.3	93.4	90.5
20000	96.7	93.7	90.7
30000	97.1	94.1	91
40000	97.4	94.4	91.3
50000	97.8	94.7	91.7
60000	98	95	92

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
Article Received: *12 July 2023* ***Revised:*** *04 September 2023* ***Accepted:*** *20 September 2023*

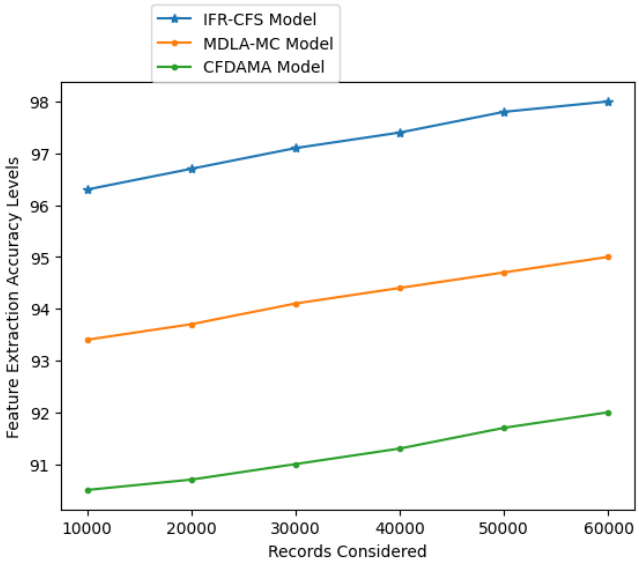

Fig 5: Feature Extraction Accuracy Levels

Selecting the k highest ranked elements according to feature extracted set S is an easy approach for feature selection using ranking. This is not ideal, but it is usually better than the alternatives. As it simply requires the computation and sorting of n scores, it is very efficient computationally. Rank the features for each model where their importance has been calculated using correlation. To find out where each feature falls on an aggregated list of all models' rankings, one can use the median function. The median rank should be used to sort the assembled list. The Feature Ranking Time Levels of the proposed and traditional models are shown in Table 3 and Figure 6.

Table 3: Feature Ranking Time Levels

Records Considered	Models Considered		
	IFR-CFS Model	MDLA-MC Model	CFDAMA Model
10000	14.7	17	20.7
20000	15	17.2	21
30000	15.3	17.3	21.2
40000	15.6	17.5	21.4
50000	15.8	17.6	21.7
60000	16	18	22

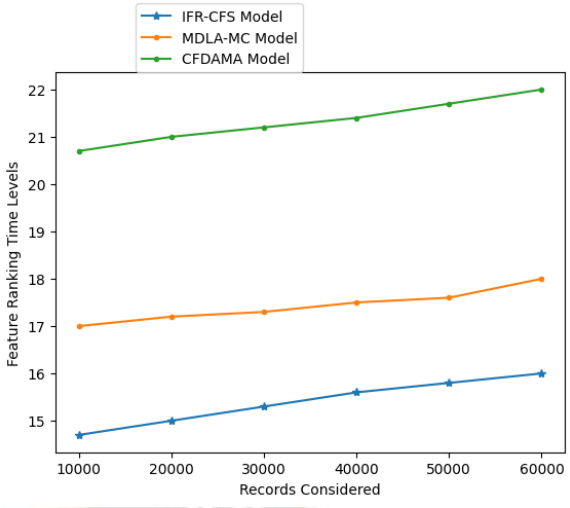

Fig 6: Feature Ranking Time Levels

Covariance or correlation is divided by the sum of the standard deviations of the two feature variables, yielding the correlation coefficient. The spread of a set of numbers around an average is quantified by its standard deviation. Covariance is a statistical measure of the relationship between two variables. For variables classified as either between or within intervals, the correlation coefficient is calculated as the square root of the total amount of squares for the interval type of variable divided by the overall sum of squares. The Correlation Calculation Accuracy Levels of the proposed and existing models are shown in Table 4 and Figure 7.

Table 4: Correlation Calculation Accuracy Levels

Records Considered	Models Considered		
	IFR-CFS Model	MDLA-MC Model	CFDAMA Model
10000	97.1	93.5	92.2
20000	97.4	94	92.5
30000	97.7	94.5	92.8
40000	98	95	93.2
50000	98.2	95.5	93.6
60000	98.4	96	94

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
Article Received: *12 July 2023* ***Revised:*** *04 September 2023* ***Accepted:*** *20 September 2023*

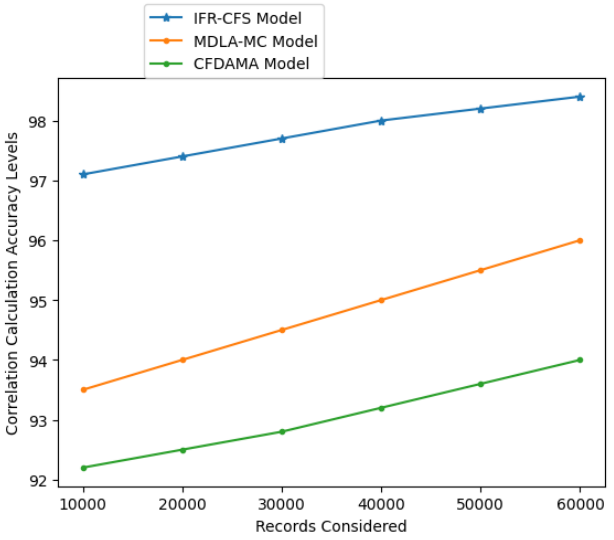

Fig 7: Correlation Calculation Accuracy Levels

The correlation factor among the ranked feature is selected and correlation ratio among the two feature sets is calculated repeatedly. Based on the feature coorelation ranking ranging from -1.0 to +1.0, based on the value range, correlation feature ranking is performed from positive range as high to negative range as low. The Correlated Feature Ranking Accuracy Levels of the proposed and traditional models are represented in Table 5 and Figure 8.

Table 5: Correlated Feature Ranking Accuracy Levels

Records Considered	Models Considered		
	IFR-CFS Model	MDLA-MC Model	CFDAMA Model
10000	97	94.3	92.9
20000	97.2	94.7	93.1
30000	97.4	95	93.3
40000	97.5	95.3	93.6
50000	97.6	95.5	93.8
60000	97.8	95.6	94

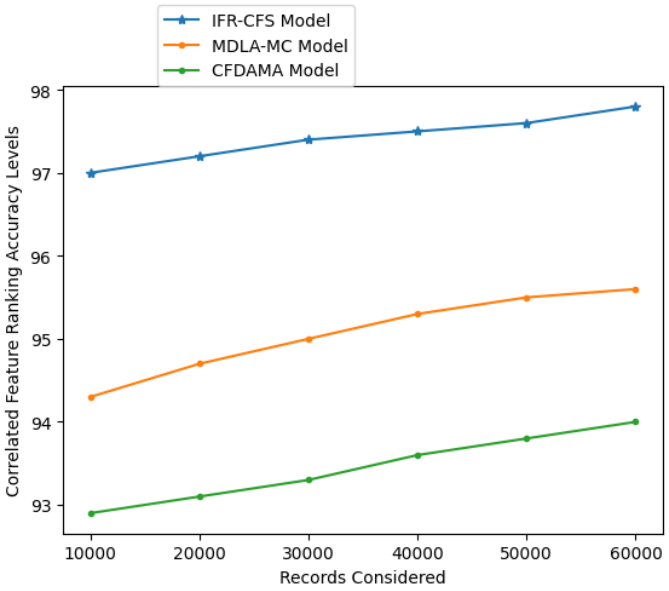

Fig 8: Correlated Feature Ranking Accuracy Levels

New features can be generated from existing features, or from a combination of existing features, for use in statistical analysis. In most cases, this involves gathering additional data that improves the model's predictive abilities. When there is an interaction between the features, model accuracy can be enhanced by feature generation. By using only the necessary data and filtering out irrelevant information, a process known as feature selection can drastically reduce the size of the model's input variable. It is when the machine learning model selects appropriate characteristics for the challenge at hand for accurate detection. The Feature Set Generation Accuracy Levels of the proposed and traditional models are shown in Table 6 and Figure 9.

Table 6: Feature Set Generation Accuracy Levels

Records Considered	Models Considered		
	IFR-CFS Model	MDLA-MC Model	CFDAMA Model
10000	97.5	93.5	94.2
20000	97.8	93.6	94.5
30000	98.1	94	94.7
40000	98.3	94.2	95
50000	98.4	94.6	95.3
60000	98.6	94.8	95.5

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
Article Received: *12 July 2023* ***Revised:*** *04 September 2023* ***Accepted:*** *20 September 2023*

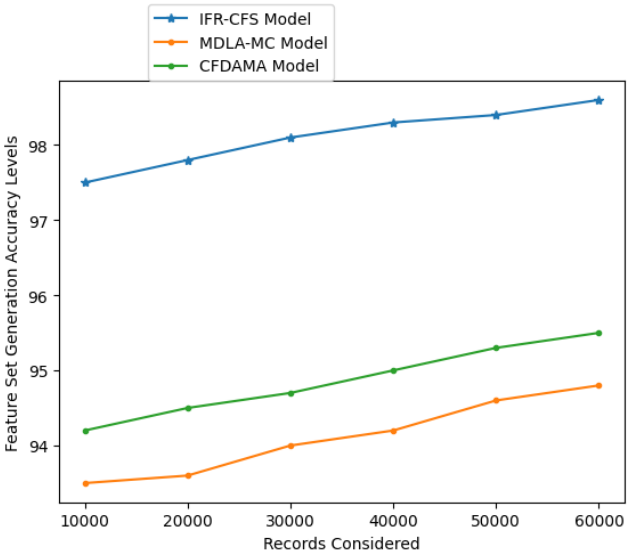

Fig 9: Feature Set Generation Accuracy Levels

V. CONCLUSION

Many cyber-security concerns, including national security threats, have their origins in malware. It is widely believed that cyber attacks provide the greatest global security threat currently. Cybercriminals and those who create security software must constantly stay one step ahead of each other. To generate a relevant feature set for accurate malware detection, this research proposes an Intellectual Feature Ranking Model with Correlated Feature Set model using enhanced logistic regression model for accurate detection of malware in the cloud environment. The suggested method utilizes a regression methodology in addition to a classification strategy and uses a feature engineering technique to drastically decrease the feature space, as opposed to existing methods, which treat malware detection as a classification problem. Extensive examination of performance demonstrates that the suggested mechanism keeps its excellent categorization capabilities. The proposed technique was more accurate than competing models by 98.2%. This research introduced a technique for detecting malware in a cloud-based computer setup. The proposed model selects the best feature set for analyzing the dynamic malware patterns that increases the security levels of cloud environment of malware detection accuracy is improved. The proposed model performs feature ranking on selected features that further reduces the feature set to reduce the memory utilization and training time. The suggested method is capable of accurately detecting both common and rare forms of malware in a wide variety of data sets. The suggested method achieves significantly higher detection rate and accuracies than existing systems, while achieving lower false positives and false negatives when these features are used for training the model. In future optimization strategies can be applied on the feature selection model for further optimizing the feature set and multi level malware analysis models can be designed using deep learning for strict detection and removal of malware patterns that can improve the quality of service levels.

REFERENCES

[1] Ahmed, M. Anisetti, A. Ahmad and G. Jeon, "A Multilayer Deep Learning Approach for Malware Classification in 5G-Enabled IIoT," in IEEE Transactions on Industrial Informatics, vol. 19, no. 2, pp. 1495-1503, Feb. 2023, doi: 10.1109/TII.2022.3205366.

[2] W. Wang et al., "Constructing Features for Detecting Android Malicious Applications: Issues, Taxonomy and Directions," in IEEE Access, vol. 7, pp. 67602-67631, 2019, doi: 10.1109/ACCESS.2019.2918139.

[3] J. C. Kimmel, A. D. Mcdole, M. Abdelsalam, M. Gupta and R. Sandhu, "Recurrent Neural Networks Based Online Behavioural Malware Detection Techniques for Cloud Infrastructure," in IEEE Access, vol. 9, pp. 68066-68080, 2021, doi: 10.1109/ACCESS.2021.3077498.

[4] P. Mishra et al., "VMShield: Memory Introspection-Based Malware Detection to Secure Cloud-Based Services Against Stealthy Attacks," in IEEE Transactions on Industrial Informatics, vol. 17, no. 10, pp. 6754-6764, Oct. 2021, doi: 10.1109/TII.2020.3048791.

[5] Ö. Aslan, M. Ozkan-Okay and D. Gupta, "Intelligent Behavior-Based Malware Detection System on Cloud Computing Environment," in IEEE Access, vol. 9, pp. 83252-83271, 2021, doi: 10.1109/ACCESS.2021.3087316.

[6] Ö. Aslan and A. A. Yilmaz, "A New Malware Classification Framework Based on Deep Learning Algorithms," in IEEE Access, vol. 9, pp. 87936-87951, 2021, doi: 10.1109/ACCESS.2021.3089586.

[7] M. Khoda, T. Imam, J. Kamruzzaman, I. Gondal and A. Rahman, "Robust Malware Defense in Industrial IoT Applications Using Machine Learning With Selective Adversarial Samples," in IEEE Transactions on Industry Applications, vol. 56, no. 4, pp. 4415-4424, July-Aug. 2020, doi: 10.1109/TIA.2019.2958530.

[8] H. Haddadpajouh, A. Mohtadi, A. Dehghantanaha, H. Karimipour, X. Lin and K. -K. R. Choo, "A Multikernel and Metaheuristic Feature Selection Approach for IoT Malware Threat Hunting in the Edge Layer," in IEEE Internet of Things Journal, vol. 8, no. 6, pp. 4540-4547, 15 March15, 2021, doi: 10.1109/JIOT.2020.3026660.

[9] I.Ahmed, M. Ahmad, A. Ahmad and G. Jeon, "Iot-based crowd monitoring system: Using SSD with transfer learning", Comput. Elect. Eng., vol. 93, 2021.

[10] Ahmed, M. Ahmad, J. J. Rodrigues and G. Jeon, "Edge computing-based person detection system for top view surveillance: Using centernet with transfer learning", Appl. Soft Comput., vol. 107, 2021.

[11] M. Ahmad, I. Ahmed and G. Jeon, "An IoT-enabled real-time overhead view person detection system based on cascade-RCNN and transfer learning", J. Real-Time Image Process., vol. 18, pp. 1129-1139, 2021.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8334
***Article Received:** 12 July 2023 **Revised:** 04 September 2023 **Accepted:** 20 September 2023*

[12] H. Jaidka, N. Sharma and R. Singh, "Evolution of IoT to IIoT: Applications & challenges", Proc. ICICC, 2020.

[13] A.Mahmood et al., "Industrial IoT in 5g-and-beyond networks: Vision architecture and design trends", IEEE Trans. Ind. Inform., vol. 18, no. 6, pp. 4122-4137, Jun. 2022.

[14] P. Varga et al., "5G support for industrial IoT applications–challenges solutions and research gaps", Sensors, vol. 20, no. 3, pp. 828, 2020.

[15] Mohammed Khairullah Mohsin, Mustafa A. Fiath. (2023). Development of Load Balancing Methodology in Cloud Computing Platforms. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 660–672. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2743

[16] P. Trakadas et al., "An artificial intelligence-based collaboration approach in industrial iot manufacturing: Key concepts architectural extensions and potential applications", Sensors, vol. 20, no. 19, 2020.

[17] P. Deflorin, M. Scherrer and K. Schillo, "The influence of IIoT on manufacturing network coordination", J. Manuf. Technol. Manage., vol. 32, no. 6, pp. 1144-1166, 2021.

[18] M. Compare, P. Baraldi and E. Zio, "Challenges to IoT-enabled predictive maintenance for industry 4.0", IEEE Internet Things J., vol. 7, no. 5, pp. 4585-4597, May 2020.

[19] B. Almadani and S. M. Mostafa, "IIoT based multimodal communication model for agriculture and agro-industries", IEEE Access, vol. 9, pp. 10070-10088, 2021.

[20] C. A. Ardagna, R. Asal, E. Damiani, N. El Ioini and C. Pahl, "Trustworthy IoT: An evidence collection approach based on smart contracts", Proc. IEEE SCC, pp. 46-50, 2019.

[21] J. Sengupta, S. Ruj and S. D. Bit, "A comprehensive survey on attacks security issues and blockchain solutions for IoT and IIoT", J. Netw. Comput. Appl., vol. 149, 2020.

[22] M. Anisetti, C. A. Ardagna, N. Bena and E. Damiani, "An assurance framework and process for hybrid systems", Proc. Int. Conf. E- Bus. Telecommun., pp. 79-101, 2020.

[23] M. Anisetti, C. A. Ardagna, N. Bena and A. Foppiani, "An assurance-based risk management framework for distributed systems", Proc. IEEE ICWS, pp. 482-492, 2021.

[24] M. Anisetti, F. Berto and M. Banzi, "Orchestration of data-intensive pipeline in 5G-enabled edge continuum", Proc. IEEE World Congr. Serv. (SERVICES), pp. 2-10, 2022.

[25] V. Sharma, I. You, K. Yim, R. Chen and J.-H. Cho, "Briot: Behavior rule specification-based misbehavior detection for IoT-embedded cyber-physical systems", IEEE Access, vol. 7, pp. 118556-118580, 2019.

[26] R. Sihwail, K. Omar, K. Zainol Ariffin, and S. Al Afghani, "Malware detection approach based on artifacts in memory image and dynamic analysis," Applied Sciences, vol. 9, no. 18, p. 3680, 2019.

[27] A.Pinhero, M. L. Anupama, P. Vinod et al., "Malware detection employed by visualization and deep neural network," Computers & Security, vol. 105, Article ID 102247, 2021.

[28] A.Yuan, J. Wang, D. Liu, W. Guo, P. Wu, and X. Bao, "Byte-level malware classification based on Markov images and deep learning," Computers & Security, vol. 92, Article ID 101740, 2020.

[29] R. Patil, H. Dudeja and C. Modi, "Designing in-VM-assisted lightweight agent-based malware detection framework for securing virtual machines in cloud computing", Int. J. Inf. Secur., vol. 19, pp. 147-162, 2020.