Engineering and Technology Journal e-ISSN: 2456-3358

Volume 08 Issue 03 March-2023, Page No.-2038-2046

DOI: 10.47191/etj/v8i3.06, I.F. – 7.136

A Novel Method for Managing System Vulnerability using Machine Learning Algorithms

Ramakrishna Hegde[1], Soumyasri S M[2]

[1]Dept. of Computer Science and Engineering, Vidyavardhaka College of Engg. Mysore-570002, India

[2]Dept. of MCA, JSS College of Arts, Commerce and Science, Mysore-570002, India

ABSTRACT: The extensive sharing of personal information over networks has given rise to an emerging malware industry. This has widened the scope of the organizations being vulnerable to malware - driven cybercrime. Such organized and distributed cyber-attacks can compromise the confidentiality, integrity and availability of any organization's valuable data and resources. The endpoints (Desktops, Laptops, Mobiles, Servers, etc.) are more vulnerable and hence mainly targeted by the cyber criminals. The aim of this study is to determine the probability of such endpoints being affected by cybersecurity threats, based upon certain characteristics of the particular endpoint. Using the machine learning techniques applied in this study, like missing data analysis and imputation (Multiple Imputation), ensemble learning algorithms (Bagging and Boosting), it can be predicted that which devices/systems in an organization are likely to be infected by malwares, ransomwares or other such threats. Based on such findings, proactive measures can be taken, and cyber security strategies can be devised which can help organizations prevent losses to the tune of millions of dollars and become cyber resilient.

I. INTRODUCTION

In 2016, several developments in learning algorithms techniques such as self-driving cars, linguistic exchange of information, health sector, and reasonable smart device were made in cyber security. To protect networks, devices, training programmes, and relevant data from attack vectors, damage, or unauthorised access, networks, devices, processes, and practises are referred to as being "cyber secure". They must be used to locate useful data from various audit datasets that are used in intrusion detection. We will apply Machine Learning technology to these concepts in cyberattacks in order to improve the security mechanisms inside the security. To begin, we must feed the data into the classification model. The dataset sample trains the model, resulting in a trained model. Following the feeding of the set of data sample, the supervised learning formula is used and implemented. In this malware detection, the computer vision formula is critical in increasing security protocols. The two categories of ML algorithms are supervised and unsupervised and unsupervised learning. By the relevant data (i.e., input) they choose, they can be distinguished from one another. Methodologies that are given such as labelled training data with the project of acknowledgement sets labels apart are said to be learning under supervision. Unsupervised learning describes techniques where algorithms are given set of unlabelled data and left to infer the classes on their own. Most of the time, labelled data is extremely rare, or even labelling the data itself is a laborious task, and we may not have been able to tell if labels are actually present.

The cybersecurity industry has grown in recent years, owing in part to the increasing availability of data (personal and organisational) on digital mediums. Today, cyber crime causes companies all over the world to lose millions of dollars. The Ponemon Institute conducted a study involving 507 organisations in sixteen countries and regions, across 17 industries, and determined also that global average average cost of a data breach for 2019 is $3.92 Million, an increase of 1.5 percent from the estimate for 2018. In order to address these issues, businesses all over the world are making significant investments in the capabilities of predictive using artificial intelligence and machine learning. In the government financial year 2020, according to a report by the Accenture Research Center (2019), 48 percent of organisations expect to increase their budget for using predictive analytics to improve cybersecurity by 29 percent. Senior executives report that 56% of cyberwarfare experts are overworked and that about 25% of them are unable to thoroughly investigate all issues. According to 64% of organisations, predictive analytics can cut the overall detection time by up to 12% while also lowering the cost of danger detection and response. Given the foregoing, it becomes critical for organisations to use predictive analytics to investigate access points that are highly probable to become infected by malware in the long run. Utilizing knowledge of the endpoint hardware and software specifications for an organisation, the study investigates this goal (Desktops, Laptops, Mobiles, Servers, etc.). The study considers various difficulties encountered when pursuing the goal and how those

difficulties were resolved. In the real world, many organisations frequently lack crucial knowledge about the various endpoints they employ. This data may include details about hardware specifications, software licencing, activation and expiration dates, the presence of a firewall, etc. These are problems with missing data, which, if not fixed, will have an effect on the final classification.

II. RELATED WORK

Ransomware classification is the process for figuring out whether a specific piece of software is malicious [1]. Conventional anti-malware systems rely on signature-based algorithms, but they are unable to identify newly created malware or malware that has been updated to use evasion methods including encryption, wadding, polymorphic, clouding, and metamorphosis. There have been several recent attempts to employ cutting-edge machine learning approaches in the field of malware categorization in order to notice these fresh varieties of malware. We review the most recent initiatives in this field in this subsection. We also draw attention to a number of comparable studies that used picture classification methods to find malware[2] and energetic analysis, as in earlier publications, is also available. Recent research has revealed that article-built malware classification approaches stay silent vulnerable to evasion strategies used by contemporary malware.

We gathered malware samples from Windows executable files disguised as innocuous files for the examination[3]. We label the samples using the Software Removal Tool. 525 harmful and 525 benign selected samples make up our dataset. The dataset is then divided into the training set (80%) and the test images (20%). (20 percent). In order to obtain an average set of data, we repeated the testing and training phases of our studies ten times each. We direct potential readers to a longer edition of this paper where we give updated techniques and a dataset evaluation using a sizable dataset[4]. This paper does parameter selection first, and their suggested model performs best at cut point 8. The model's performance will decline if the cut point is raised to 16 since additional random pixels cannot be trained effectively. It also slows classifiers down and overbites them. As a result, we decide to use the cut point 8 for the remainder of our studies. Table II presents the effectiveness of several image processing techniques and various picture attributes. Generally speaking, CNN outperforms GIST. The training and test sets yield the finest results for our HIT[5]. This indicates that the HIT can generalise effectively to samples that have not been encountered. Productivity on the training dataset is typically superior to that on the validation set.

Furthermore, there are no metrics for live cybersecurity audits, so the topic of cybersecurity audits is incompletely known because it changes so quickly[6]. According to Khan, in order to conceal a sizeable area to consider when creating an information security audit, auditors should encompass every pertinent region any organisation, including client processes, finance, human resources, IT application systems, legal, buying, regulatory affairs, physical security, and all relevant third - party that have friendships with the business[7].

Private businesses and governmental institutions must now deal with frequent and sophisticated cyber threats and cyberattacks. In order to protect themselves from cybercriminals, organisations should create and grow a level of security and awareness. To address cyber threats, cyber risks, and cyberattacks that emerge in a competitive cyber landscape, information systems audits such as IT and data security audits such as InfoSec, which have been previously cost-effective, attempt to merge into cyberwarfare audits [8]. The complexity of the cyber threat landscape and the increase in the quality and variety of cyberattacks, however, are posing a challenge to the current cybersecurity inspection models and providing justification for a network security audit model. The simplest techniques and methodologies used by global specialists in the area of cybersecurity confirmation and audit are reviewed in this text. In order to create a strong and coherent synthesis, the real scope, strengths, and weaknesses of these approaches and their theoretical foundation are highlighted through analysis. In order to conduct cybersecurity audits in organisations and Nation States, this text proposes an innovative and exhaustive cybersecurity audit model. For all areas where structure is useful, the Cyber Security Audit Model (CSAM) assesses and confirms audit, preventive, rhetorical, and detective controls. CSAM has undergone testing, enforcement, and validation alongside cybersecurity. Each model is being validated through a research case study, and as a result, the results are made public [9].

It is granted a completely unique method attempt to feature options to detect malicious nodes at their own command and control (C&C) segment. A significant drawback is that although researchers have suggested solutions based on their research, there is no way to evaluate these solutions because some of them might have a lower detection performance than alternatives. In order to achieve the intended objective, we identify the feature set that supports connections between botnets at their C&C section and maximises the rate at which those botnets are detected. Genetic formula (GA) has been selected as the option with the highest detection rate because it is familiar to users. We frequently employ the deep learning formula C4.5, which distinguished between connections that belonged to a botnet and those that did not[10]. A few experiments were conducted to introduce the GA's most basic parameters and the C4.5 formula. We typically conduct experiments jointly in order to obtain the most simple and direct array of choices for each analysed botnet in particular, as well as for each type of botnet in general. The findings, which include a significant reduction in features and an improved detection rate than for the related work conferred, are presented at the conclusion of the paper.

The issues with neural network-based intrusion detection, such as redundant data, a lot of data, and lengthy training, are easily solved at the local optimum. The use of DBN and probability - based neural networks (PNN) is proposed as an intrusion detection method. First, using the nonlinear intelligence of DBN, the raw data is converted to low-dimensional data while retaining the key attributes of the data. Second, the number of hidden-layer nodes per layer is optimised using the based on swarm intelligence formula to accomplish the simplest learning performance. Next, PNN is used to classify the information with low dimensions [11]. Subsequently, the dataset is used to evaluate the effectiveness of the strategies stated earlier. The research's findings indicate that approach outperforms PCA-PNN, standard PNN, and non-optimized DBN-PNN. ML has moved from the lab to the frontline of deployable techniques over the past couple of years. Machine learning is used frequently by Amazon, Google, and Facebook to improve customer experiences, guide purchases, connect people generally with fresh apps, enable intimate life. The strong capacity of machine learning is also present in cybersecurity. Machine learning can be used by cybersecurity to improve malware detection, organise events, recognise breaches, and notify organisations of security issues[12]. In an hour, malware alone could represent three million new samples. Malware analysis and detection techniques from the past are unable to keep up with modern attacks and variants. Cyberattacks are being delivered at alarming rates thanks to sophisticated malware and new attacks that are organized to evade detection of end - points and connectivity. To address the growing malware downside, new methods like computer vision should be used. This claim describes how machine learning can be used by cyber defence analysts to find and highlight sophisticated malware. The findings of our preliminary analysis are presented, along with a discussion of potential follow-up research to improve machine learning.

Secure and dependable networks are essential given the rapid expansion of computer networks and user content consumption. Because it has been established that there are more and more different types of network attacks, it is essential to develop a supply of reliable automated tools for attack detection. One of the threat systems that tries to find intrusions coming back from the internet is the antivirus software. The literature has identified a number of methods for detecting attacks. To visualise intrusion detection in the recent times, mining techniques were popular[13]. The features of inbound interferences were determined by using well data from network's data. When an exactly equal entity is discovered within the features of the off well data, it is classified as an interference. As a result of the recent analysis, various anti - malware models were developed to support this criterion, and the accuracy has improved. A quick review of the earlier approaches is performed. Metadata preprocessing strategies and detection approaches make up the entire strategy. Additionally, there are two types of information pre - processing stage approaches: feature extraction and transformation models, which encourage able to operate research methods over the alternatives. The detection methods are categorised similarly as learning algorithms and organic process methods[14].

When a group of auditors conducts an IT audit, a data protection audit, or an audit standards, there are recurring phases such as designing, defining objectives and scope, defining terms of the agreement, conducting the audit, gathering supporting evidence, assessing risks, reporting the audit report, and scheduling carry tasks. However, due to the high quality of many cybersecurity domains, this will require a significant amount of effort. Moreover, most internet capabilities are not covered by the scope of inside reviews. This background comprises threat management, advancement life - cycle, security programme, 3rd managerial staff, knowledge managerial staff, user access, threat/vulnerability management, the need for word, which will be achieved through administration review meetings, cyber threat assessments, information conservation and sustainable, risk analytics, crisis management, and response. Furthermore, Deloitte's framework is consistent with trade frameworks such as the National institute Of standards (NIST), the data Machinery Setup Papers (ITIL), the Review panel of Sponsoring Organizations Commission (COSO), and the International Organization for Standardization (ISO).

Malware categorization founded on malware photos and deep learning has emerged as one of the efficient answer since it avoids a lot of story trade work, which is one of the constraints of standard machine learning methods. Deep learning-based malware classification has grown more appealing during the last two years. For the categorization of malware photos, for instance, a convolutional neural network (cnn with weighted softmax) loss was developed. As shown in figure 2.1.6 planned a ransomware arrangement technique called MCSC, which transferred the separated malware code into grey images predicated on simhash and then recognised their families by fully convolutional. Their experiments showed that the new calculation can fit both these distinctive convolution neural network with a better classification performance. Additionally, a CNN-based framework for spyware categorization was suggested by Kalash et al. (2018). The pre-training VGG16 models were employed in this technique, which had great accuracy[15]. Similar to this, Rezende et al. (2017) have applied the Multilayer perceptron neural architecture for malware analysis using transfer learning.

I. METHODOLOGY

A. System Architecture

As in figure 1 system's structure and behaviour are defined by its system architecture, which is a design phase. A system's structural properties are logically supported by an architecture description, which is a concise overview of the system. It outlines the system's constituent parts or key components and

offers a blueprint for the development of systems and products that will cooperate to implement the whole scheme.

The System architecture is shown below.

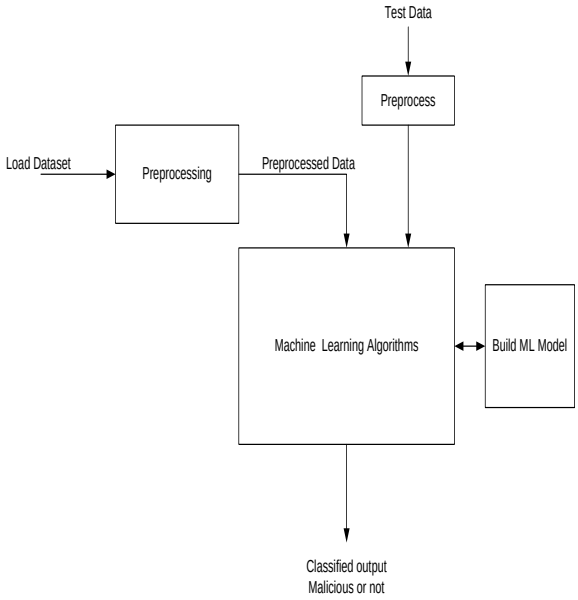


Figure 1 : System Architecture

B. Classes Designed for the system

In the Unified Modelling Language (UML), a diagram is a type of static structure that illustrates the classes, attributes, and relationships between the classes in a system to describe the system's structure. Here is the class diagram figure 2.

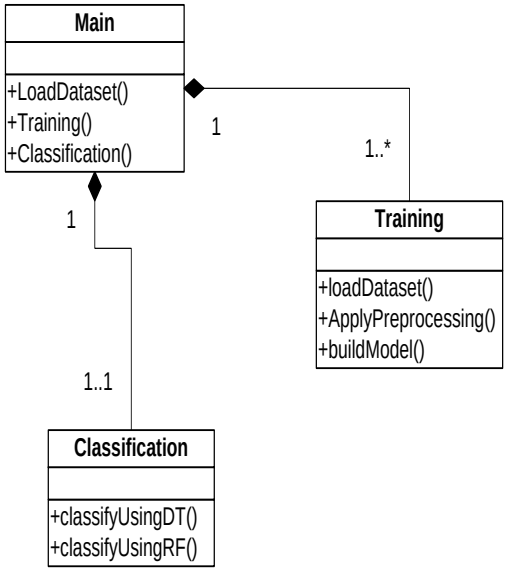


Figure 2 : Classes designed for the system

C. Diagram of the system's use cases

A use-case analysis is used to create a specific kind of behavioural diagram as in fig 3 known as a use case diagram. Its objective is to provide a graphical overview of a system's i. functionality in aspects of actors, their objectives (defined as use cases), as well as any roles and responsibilities among those use cases.

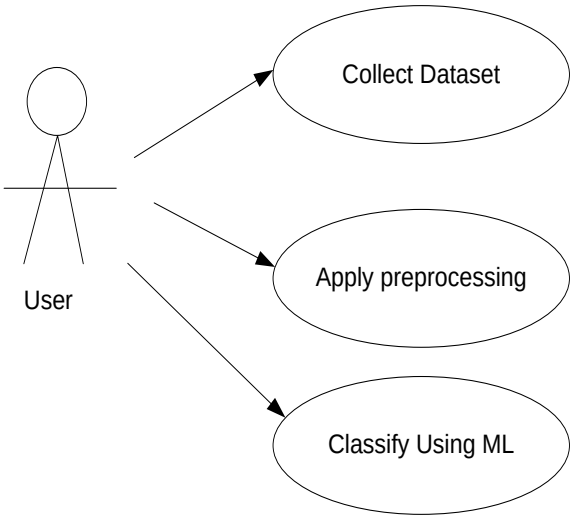


Figure 3 : Use cases

D. System operation sequence diagram

In Unified Modeling Language (UML), a sequence diagram is a type of communication graph that represents how processes interact with each other and in what order. It is a Chart construct. The sequence diagrams as in figure 4 are displayed below.

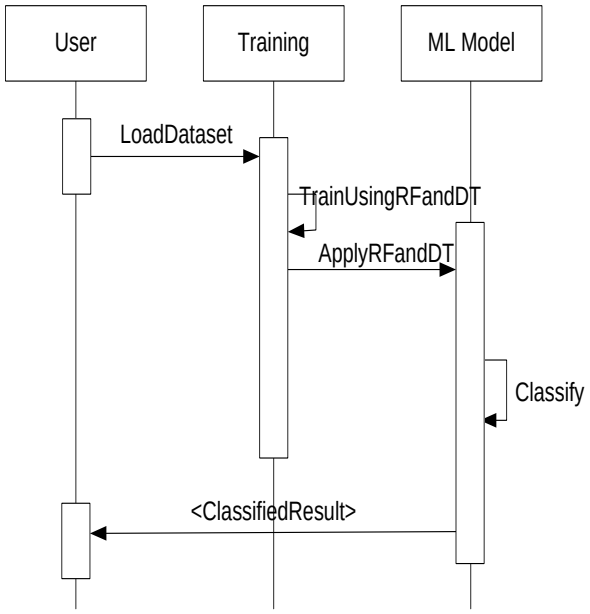


Figure 4 : System Operations

E. The system's data flow diagram

A information diagram shows how data "flows" through such an information system graphically. DFDs can be utilised to visualise data analysis (structured design). On a DFD, an internal process transfers data items from an external device or domestic data store to an outer data sink or internal data store.

Level 0 Data flow diagram

A frame of reference or level 0 data flow depicts the interaction of the system with external agents that serve as data sources and sinks. The logical model represents the system's interactions with outside world solely in terms of flow of data across the system boundary. As in figure 5 context diagram depicts the a

whole process of the system with no indication of its internal organisation.

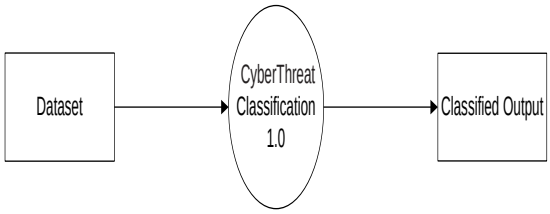


Figure 5 : Data flow

ii. Data flow diagram at Level 1

The Level 1 DFD demonstrates how well the system is broken down into smaller units (processes), which each deals including one or more data streams to or by an external entity and which, when combined, provide the system's entire functionality. In Figure 6 Additionally, it shows this same flow of data among the various system components and recognises domestic data stores that are essential for the organisation to function.

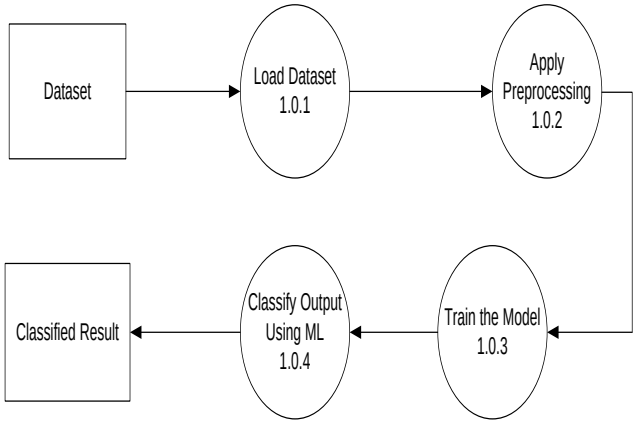


Figure 6 : Level 1 data flow

Once the data set is loaded into the machine then , we will apply preprocessing concept to the model and the model will be trained by machine learning techniques by this model will be created and for that model we will classify output using ML. The result will be generated based on performance measure and classified the result.

The following tasks are necessary for the implementation phase.

- Meticulous planning.
- System and constraint investigation.
- Development of transitional methods.
- A review of the switchover procedure.
- Making the right decisions when choosing the platform
- Deciding on the right language to use when developing applications.

iii. Platform used for Implementation

A platform is an important component of software development. Simply put, a platform is "a place to launch software." Windows XP is used in this project for implementation purposes, and the following are the reasons why: Support for Integrated Connectivity, more secure and reliable than the prior version, include a restore option and remote desktop connection Enhancements to the device driver verifier, Significantly fewer scenarios requiring a reboot enhanced coding protection, Support for parallel DLLs, Windows File Protection, and proactive multitasking architecture, support for scalable processor and memory, IPSec, Kerberos, Smart Card Support, Internet Explorer Add-on Manager, Firewall, Windows Defender security Center, and Encrypting File System (EFS) with Multi-User Support modern aesthetics.

F. Proposed System

A solution was created using the framework to identify endpoints that are strongly likely to become infected by malware.

f. Data cleaning and Imputation:

The data for the study was sourced from the Microsoft Malware Classification Challenge (BIG 2015). For the purpose of this study, 50,000 random rows of data were taken, as would be the case in any mid-size organization in terms of number of endpoints. The variables present in the data and their description is described in Table 1. Alike categorical and continuous data are present in the dataset. A significant portion of the data - set in this research was missing, which posed a number of difficulties. The data upon further analysis was categorizes as Missing at Random (MAR). The misplaced figures credited by means of the Multiple Imputation technique by using the MICE package available in R. Further, exploratory data analysis helps in understanding class imbalance and correlation.

g. Dataset

The dataset contains both numerical and categorical data as in table 2. A few of the basic challenges encountered in this study was the subject of a huge proportion of incomplete information in the dataset. Following further analysis, the records was classified as Missing at Random (MAR). Further, exploratory data analysis helps in understanding class imbalance and correlation. Here system features is taken as dataset and each attributes is divided as train set and test set. In numerical data, 80% of the data is used as a training set, while 20% is used as a testing dataset, and it is classified as malware or not.

Table 1: List of attributes used

Variable	Description
MachineId	Individual machine ID
ProductName	Type of Endpoint Protection enabled e.g. 'windefender
HasTpm	True if the machine has tpm (Trusted Platform Module) enabled
Platform	Version of Windows installed
Processor	Process architecture of the installed operating system
SkuEdition	SKU Edition of the Windows Version
IsProtected	If the Machine is Protected by an active and up-to-date Antivirus Product
Firewall	If the Windows Firewall is enabled
AdminApprovalMode	Whether the "administrator as Admin Approval Mode" user type is disabled or enabled
DeviceType	Type of the Device eg - Notebook, Laptop, Desktop
PrimaryDiskTotalCapacity	Amount of disk space on primary disk of the machine in MB
PrimaryDiskTypeName	Friendly name of Primary Disk Type - HDD or SSD
SystemVolumeTotalCapacity	The size of the partition that the System volume is installed on in MB
HasOpticalDiskDrive	True indicates that the machine has an optical disk drive (CD/DVD)
TotalPhysicalRAM	Retrieves the physical RAM in MB
AutoUpdate	Friendly name of the WindowsUpdate auto-update settings on the machine
GenuineStateOS	Indicates the authenticity of the OS version
IsSecureBootEnabled	Indicates if Secure Boot mode is enabled
IsPenCapable	Is the device capable of pen input?
IsAlwaysOnAlwaysConnectedCapable	Retrieves information about whether the battery enables the device to be AlwaysOnAlwaysConnected
IsGamer	Indicates whether the device is a gamer device or not based on its hardware combination
IsInfected	Indicates if the machine has been diagnosed as Malware affected

Table 2 : Each Attributes having system features

MachineId	ProductName	HasTpm	Platform	Processor	SkuEdition	IsProtected	Firewall	AdminApp	DeviceType
000002896	win8defender	1	windows10	x64	Pro	0	1	1	Desktop
00000753E	win8defender	1	windows10	x64	Pro	0	1	1	Notebook
00000790E	win8defender	1	windows10	x64	Home	0	1	1	Desktop
00000b11E	win8defender	1	windows10	x64	Pro	0	1	1	Desktop
000014a5F	win8defender	1	windows10	x64	Home	0	1	1	Notebook
000016191	win8defender	1	windows10	x64	Pro	0	1	1	Desktop
0000161e9	win8defender	1	windows10	x64	Home	0	1	1	Notebook
000019515	win8defender	1	windows10	x64	Home	0	1	1	Notebook
00001a027	win8defender	1	windows10	x64	Pro	0	1	1	Notebook

PrimaryDi	PrimaryDi	SystemVo	HasOptica	TotalPhys	AutoUpda	GenuineS	IsSecureB	IsPenCapa	IsAlwaysC	IsGamer	IsInfected
476940	HDD	299451	0	4096	UNKNOW	Invalid	1	0	1	0	0
476940	HDD	102385	0	4096	UNKNOW	OFFLINE	1	0	1	0	0
114473	SSD	113907	0	4096	NOTIFY	Invalid	1	0	1	0	0
238475	UNKNOW	227116	0	4096	NOTIFY	Invalid	1	0	1	0	1
476940	HDD	101900	0	6144	NOTIFY	Invalid	1	0	1	0	1
114473	SSD	113671	0	8192	NOTIFY	Invalid	1	0	1	0	1
476940	HDD	458782	0	4096	NOTIFY	Invalid	0	0	1	0	1
305245	HDD	290867	1	4096	NOTIFY	Invalid	1	0	1	0	0
305245	HDD	303852	0	4096	NOTIFY	Invalid	1	0	1	0	0

Cross validation and model building: Instead of splitting the data into training and testing data subsets, a K- Fold cross validation approach is adopted. The K-Fold cross validation technique randomly creates K subsets from the data of approximately equal size. From these subsets, the first subset is treated as the validation or the test set and the remaining K-1 subsets are used to train the predictive model. For the purpose of this study, the value of K is taken as 10. Ensemble learning models were created on this 10-Fold cross validated data. They are described below:

1. Decision Tree Algorithm
2. Random Forest Algorithm
3. K-Nearest Neighbor Algorithm
4. Naïve Bayes Algorithm
5. Logistic Regression

h. Model Evaluation

Training a model and fitting a model to available training data are the two stages of supervised learning. Using the trained model on new samples to make predictions.

The assignment:

• We are given a collection of objects

• Feature set X is used to represent each object.

• Each object is labelled as Y or mapped to the correct answer.

This training data is used during the training phase to find the best model that will generate the appropriate description Y for previously unseen objects given the feature set X. In the case of malware classification, X could be some characteristics of file content or behaviour patterns, such as file statistics and a list of API functions used. Labels Y could be vulnerable to hacking or benign, or they could be more specific, such as a virus, Trojan-Downloader, or adware.

During the training phase, we must choose a model family, such as neural networks or decision trees. Each method in a family is usually represented by its parameters.

Training entails looking for the model from the chosen family with a specific set of parameters that provides accurate answers for the trained model over the set of reference objects according to a specific metric. In other words, we 'learn' the best parameters for defining a valid mapping from X to Y.

As in figure 6.1 second step, applying the model to new objects, can be done when a model has been trained and its quality has been established. The model's type and its parameters stay the same during this phase. Only predictions are produced by the model.

This is the protective phase in the case of malware discovery. Customers frequently receive a trained model from vendors, and the product then takes decisions on its own based on model predictions. Making mistakes can have disastrous effects on a user, such uninstalling an OS driver. The vendor must make a wise choice of model family.

II. RESULTS

The outcomes of the proposed methods are discussed further below.

Table 3: Classification Report for Supervised Learning Algorithm

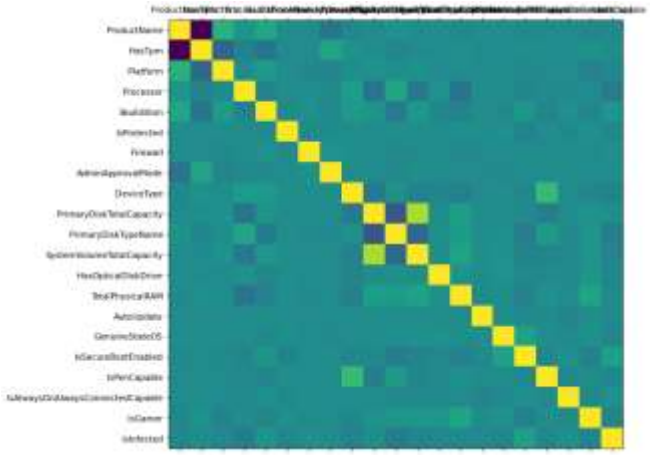
Algorithm	Classification	Precision	Recall	F1 score	Support	Accuracy
KNN	0	0.52	0.28	0.37	2464	91%
	1	0.63	0.83	0.72	3649	
NB	0	0.57	0.30	0.39	2464	93.68 %
	1	0.64	0.85	0.73	3649	

RF	0	0.55	0.37	0.44	2464	93.71 %
	1	0.65	0.80	0.72	3649	
LR	0	0.61	0.18	0.28	2464	93.44 %
	1	0.62	0.92	0.74	3649	
DT	0	0.55	0.38	0.45	2464	93.71 %
	1	0.65	0.79	0.71	3649	

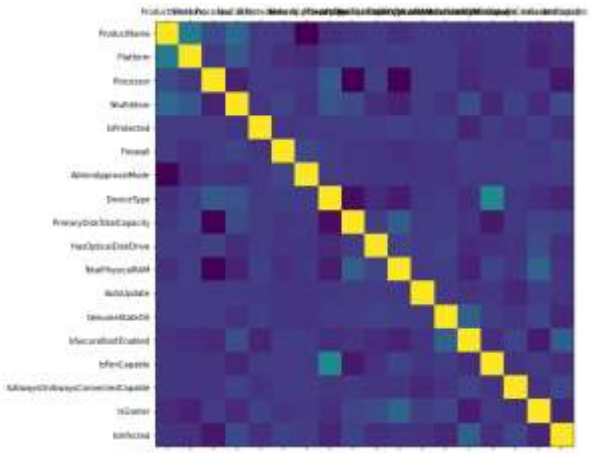
The above table 3 shows the performance measure of the machine learning model where having malware class and no malware class, the classification is based on numeric value where 0 is identified as no malware class and 1 is identified as malware class achieved out of 30,000 datasets by applying pre-processing technique that is data cleaning, data reduction and data transformation.

The goal of this research is to determine the likelihood of such endpoints being impacted by cybersecurity risks based on certain endpoint characteristics. It is possible to predict which devices/systems in an organisation are likely to be infected by spyware, ransomwares, or other such threats using the machine learning techniques used in this study, such as having missed market research and imputation (Multiple Probabilistic reasoning), supervised learning methods (Bagging and Boosting).

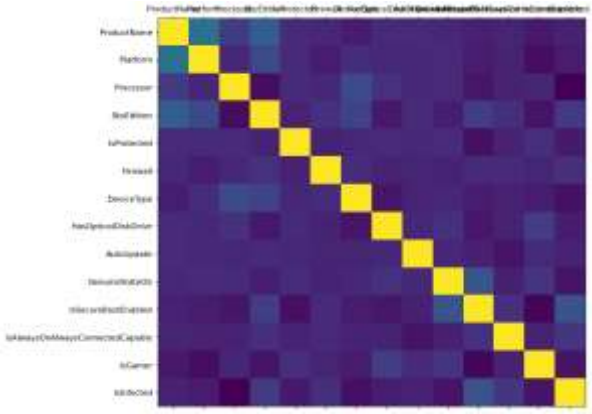
Below graph shows the one to one correlation applied in this study as follows:



Graph 1: Correlation 1

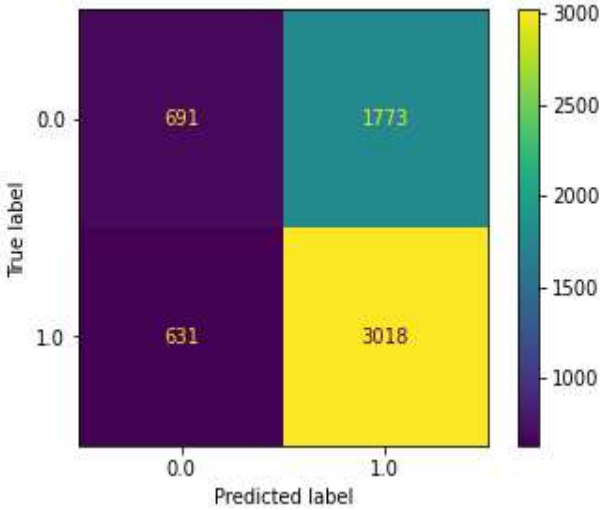


Graph 2: Correlation 2

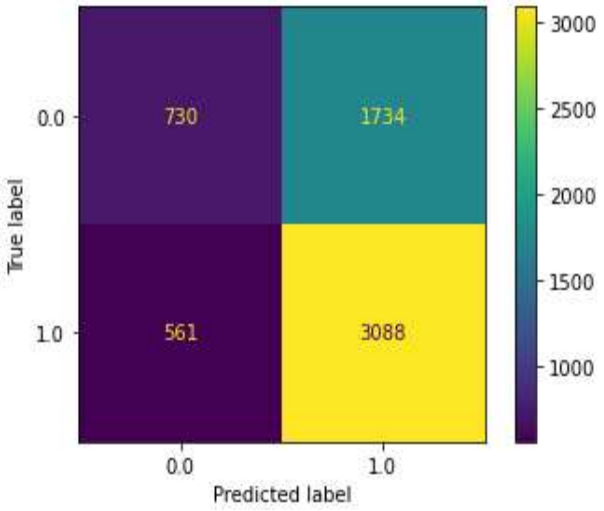


Graph 3: Correlation 3

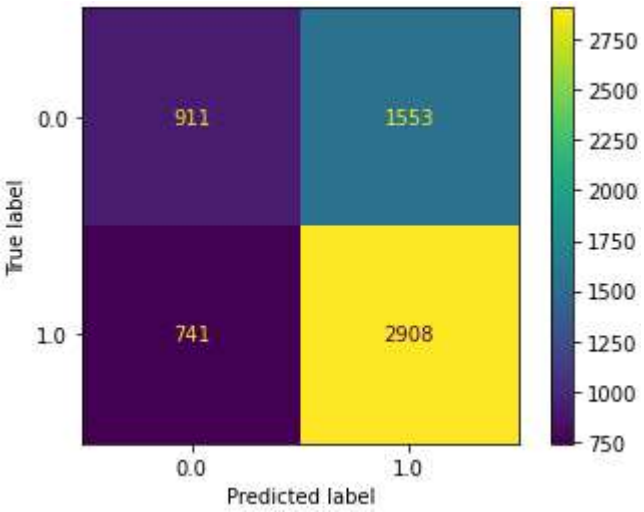
As above graph 1, 2 and 3 shows the correlation applied in system attributes by deleting null values, reducing the size of dimension and transformation of data from one form to another such as Product Name, platform processor, skuEditon, IsProtected, Firewall, DeviceType, HasOptical DiskDrive, Auto Update, Genuine State OS, IsSecure Boot Enabled, IsAlwalys On Always Connected Capable, IsGramer, IsInfected. All these attributes are correlated one – one as in graph yellow colour and purple represents the correlation.



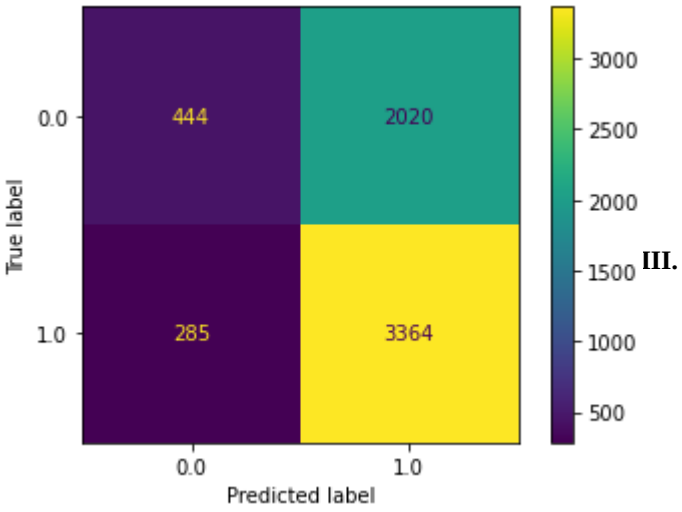
Graph 4: KNN confusion matrix



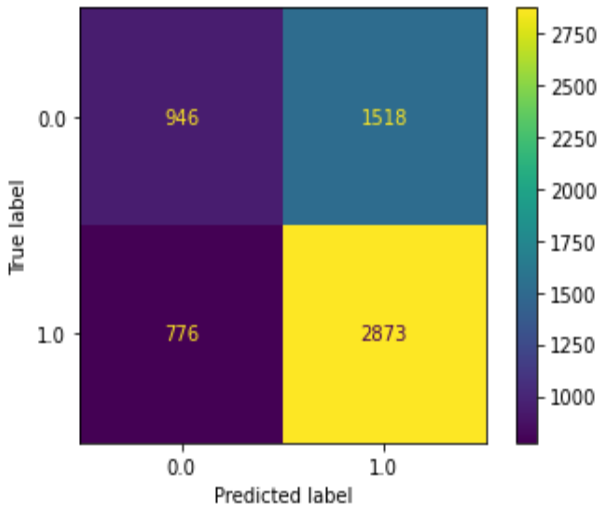
Graph 5: Navie Bayes confusion matrix



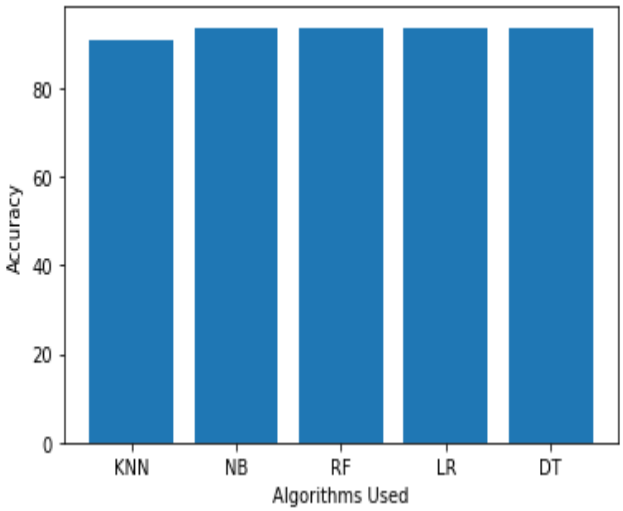
Graph 6: Random Forest confusion matrix



Graph 7: Logistic Regression confusion matrix



Graph 8: Decision tree confusion matrix



Graph 9: Comparison of algorithms

By referring to the above graph, it shows us the analysis and comparison of the supervised learning algorithms applied in this study such as KNN algorithm results with 91% of accuracy, Navie bayes results with 93.68% of accuracy, Random forest scores with 93.71% and Logistic regression scores with 93.44%, Decision tree scores with 93.71%. By analysing this algorithms we conclude one of the algorithm is the best classifier.

III. CONCLUSION

The experiment indicates the hypothesis that it is possible to predict an endpoint's likelihood of becoming poisoned by spyware and many other cyber security threats if one has the proper information of the standards of organisational end devices (both software and hardware). Numerous actual data challenges, such as understanding missing data, intimation of forgetting metadata interpolation methodology, difficulties with bridge of data and summative assessment of the models, had to be overcome in order to accomplish the study's goal. The description of variables of the study to create the model is by no means complete; an amount of additional measurements and

indicators may be adjusted depending on their accessibility and suitability to different organisations in order to create concepts that are more credible and valid.

REFERENCES

1. Journal homepage: www.elsevier.com/locate/compeleceng
2. Duc-Ly Vu1 Trong-Kha Nguyen2 Tam V. Nguyen3 Tu N. Nguyen4 Fabio Massacci1 Phu H. Phung "HIT4Mal: Hybrid image transformation for malware classification",2019.
3. Journal homepage: www.elsevier.com/locate/cose
4. "A Convolutional Transformation Network for Malware Classification",2019 6th NAFOSTED Conference on Information and Computer Science.
5. Baoguo Yuana , Junfeng Wang "Byte-level malware classification based on markov images and deep learning",2020.
6. Abdurrahman Pektaş1, Tankut Acarman "Malware classification based on API calls and behaviour analysis", The Institution of Engineering and Technology,2017.
7. Xiaopeng TIAN, Di TANG, "A Distributed Vulnerability Scanning on Machine Learning", 2019 6th International Conference on Information Science and Control Engineering.
8. J. Cano, "Cyberattacks-The Instability of Security and Control Knowledge", *ISACA Journal*, vol. 5, pp. 1-5, 2016.
9. C. Hollingsworth, "Auditing from FISMA and HIPAA: Lessons Learned Performing an In-House Cybersecurity Audit", *ISACA Journal*, vol. 5, pp. 1-6, 2016.
10. Li X, Wang J, Zhang X, "Botnet Detection Technology Based on DNS", J. Future Internet, 2017.
11. Y J Hu, Z H Ling, "DBN-based Spectral Feature Representation for Statistical Parametric Speech Synthesis", *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 21-325, 2016.
12. Dinil Mon Divakaran et al., "Evidence gathering for network security and forensics", *Digital Investigation*, pp. 56-65, 2017.
13. S Fong, R Wong, A V Vasilakos, "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data", *IEEE Transactions on Services Computing*, vol. 9, no. 1, pp. 33-45.
14. M. Khan, "Managing Data Protection and Cyber security Audit's Role", ISACA Journal, vol. 1, pp. 1-3.
15. Bharadwaj R. K. Mantha, Borja Garcia de Soto, "Cyber security challenges and vulnerability assessment in the construction industry", Accepted 7 June 2020.