IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning-Based Risk Assessment and Visualization for Cybersecurity Resilience in Organizations

[1]S. Suriya, [2]Dr. Jayanthi M G
[1]Mtech Student, [2]Professor
[1]Department of CSE, Cambridge Institute of Technology, Bengaluru
[2]Department of CSE, Cambridge Institute of Technology, Bengaluru

Abstract: Organizations are more vulnerable to cyber-attacks, especially malware, due to the growing reliance on digital networks for the transmission of critical information. Computers, smartphones, tablets, and servers used by organizations are the target of this research, which seeks to determine how susceptible they are to cyberattacks. Organizations may take proactive steps to improve their cybersecurity resilience if they are able to estimate the possibility of malware and similar threats compromising these endpoints. The research does this by analyzing endpoint attributes and predicting their attack vulnerability using sophisticated machine learning methods. These approaches include multiple imputation for addressing missing data and ensemble learning algorithms like boosting and bagging. To guarantee a thorough assessment of endpoint vulnerabilities, the main approaches center on processing and modeling data taken from a publicly available cybersecurity dataset. The results show that certain endpoint characteristics are strongly associated with increased cyberattack risks, which helps to identify which devices in a company are most at danger. Organizations may successfully minimize possible risks by implementing customized cybersecurity strategies based on these findings. This research provides a data-driven strategy for improving corporate cybersecurity via the identification of high-risk endpoints. This will help avert significant financial and reputational losses that may be caused by cyber intrusions.

***Index-Terms* - Phishing websites, Machine Learning, Content based features, URLs, Attacks.**

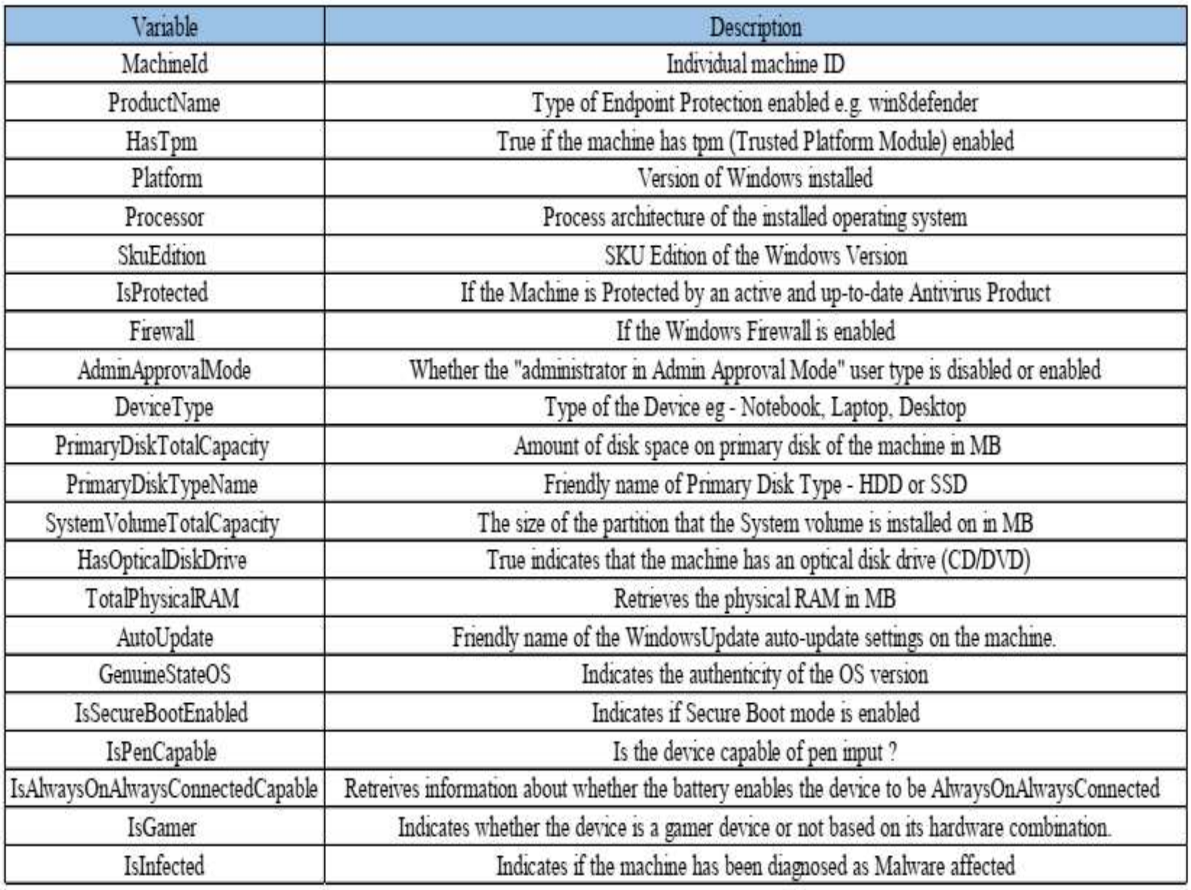
I INTRODUCTION

As businesses rely more and more on linked networks to handle mission-critical operations and private data, the cybersecurity environment has changed drastically in the digital age. Cybersecurity, especially the protection of endpoints like computers, smartphones, tablets, and servers, has become an urgent issue due to our increasing dependence on digital infrastructure. Cybercriminals target endpoints because they hold or provide access to the most valuable corporate assets. Therefore, endpoint security is vital.

Malicious software, or malware, has proliferated in the last ten years. Its goal is to harm or disrupt computer systems or even get unauthorized access to them. Cybercriminals are always coming up with new ways to circumvent standard security measures, which has led to malware assaults becoming more complex and diverse. Because even a single compromised endpoint may cause extensive data breaches, operational interruptions, and significant financial losses, the ever-changing threat environment is a major concern for businesses of all kinds.

There is a large amount of literature on endpoint security, covering topics such as how to build better detection and prevention systems. Endpoint security has always relied on anti-malware and firewall programs; however, these outdated methods are no match for the increasingly complex threats that are appearing in the modern day. Machine learning (ML) approaches to strengthen cybersecurity defenses have recently attracted a lot of attention as a result. Through the analysis of massive amounts of data, this research have shown that ML algorithms, especially ensemble approaches like as bagging and boosting, can detect and reduce risks. Despite these improvements, there are still a number of unanswered questions and holes in the existing research. Predicting which endpoints are most vulnerable before to an attack has received less attention than post-infection malware detection. Furthermore, complete datasets that account for the wide range of endpoint settings and possible threats are typically lacking. Another issue that has not been systematically handled is the management of missing data, which is a typical problem in cybersecurity datasets. This might result in predictions that are incomplete or biased.

To fill these gaps, our study will concentrate on predicting endpoint vulnerabilities before cybercriminals can exploit them. The goal of this work is to improve the accuracy of endpoint risk assessments by using advanced machine learning methods, such as ensemble algorithms like boosting and bagging and multiple imputation to deal with missing data. The results are relevant and applicable to real-world settings since the study is based on a publicly accessible cybersecurity dataset. Organizations may use the research's findings to better target their security efforts and safeguard their most susceptible assets. The Following table 1 details about the dataset used, additional to that data updated with location and crime rate in this research.



Variable	Description
MachineId	Individual machine ID
ProductName	Type of Endpoint Protection enabled e.g. win8defender
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 in 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	Retreives 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 1: List of Attributes Used

What follows is an outline of the rest of the paper: Data sources, preprocessing procedures, and machine learning approaches are described in depth in the next part of the suggested methodology. After that, the study's techniques, particularly those involved in endpoint vulnerability prediction, will be thoroughly explained. The analysis's outputs, such as a comparison of the models' performance and important discoveries, are presented in the results section. The study concludes with a brief summary of its key points, a discussion of its limitations, and some recommendations for further research.

II RELATED WORK

Cyberthreats and attacks are becoming more sophisticated and persistent, affecting organizations in both the public and commercial sectors. Businesses may better defend themselves against hackers if they encourage cybersecurity knowledge and cultivate a security-conscious culture. In an effort to keep up with the always evolving cyber dangers, risks, and attacks, the once cost-effective data technology (IT) and data security (InfoSec) audits are now attempting to combine into cybersecurity audits [1]. Currently, cybersecurity audit models are being used; however, due to the emergence of new threats and the increasing diversity and complexity of assaults, these models are becoming outdated and a new alternative must be developed. This book provides a concise overview of the fundamental approaches used by the leading cybersecurity assurance and audit professionals. Examining these methodologies and their theoretical underpinnings is crucial for arriving at a strong and cohesive synthesis. This will expose their actual breadth, strengths, and flaws. Accordingly, this study presents a comprehensive and novel technique to conduct cybersecurity audits in organizations and states. The CyberSecurity Audit Model (CSAM) checks and evaluates audit, preventative, rhetorical, and investigative controls for each and every one of a structure's practical aspects [2]. On the cybersecurity front, CSAM has undergone testing, enforcement, and validation. We are now validating each model using a research case study, and we will share the findings with the public.

Machine learning approaches for feature selection in botnet detection

A new method of providing alternatives for observing is introduced to the Botnets' Command and Control (C&C) segment. Researchers have proposed solutions based on their experiences, but there is currently no mechanism to assess their efficacy. This is a big concern as some of these solutions may not have the best detection rate compared to others. At the moment, our goal is to find the feature set that supports connections at botnets' C&C section in order to enhance the detection rate. As a set of known and selected possibilities, a genetic formula (GA) provided the greatest detection rate. We often refer to the machine learning algorithm C4.5 when determining whether a connection is associated with a botnet. Information culled from the ISOT and ISCX databases is used in this piece [3]. Several experiments were carried out to induce the minimal set of GA parameters and the formula C4.5. We often collaborate on trials to find the simplest set of options for all botnet types (specific) and all generic botnet types (generic). The results, presented at the beginning of the article, show that a higher detection rate was attained with a smaller number of characteristics as compared to the related research.

Deep Belief Network-Based Intrusion Detection

Several problems plague neural network-based intrusion detection, including repetition, large data sets, and long training periods. Being mired in a rut is easy. We provide a technique for intrusion detection that utilizes deep belief networks (DBN) in conjunction with probabilistic neural networks (PNN). The first step is to utilize DBN's nonlinear wit to reduce the data to a lower dimensional format while keeping its essential characteristics. Second, the particle swarm optimization technique is used to optimize the number of hidden-layer nodes per layer for the simplest feasible learning results. Data with low dimensions is then sorted using PNN [4]. Finally, we test all of these methods on the KDD CUP 1999 dataset. Results from experiments show that the technique is more effective than PCA-PNN, non-optimized DBN-PNN, and baseline PNN. Previously considered a specialized technology, machine learning is now an integral part of many business processes. Every day, machine learning aids tech giants like Facebook, Amazon, and Google in their efforts to enhance user experiences, facilitate purchases, and unite people via shared interests, applications, and personal connections. Cybersecurity is another area where machine learning has the potential to make significant strides. Machine learning has the potential to improve malware detection, event categorization, breach detection, and breach notification for hackers. Organizations, infrastructure, and potential exploits and weaknesses that can be advantageous to both sides can be uncovered via machine learning. Significant changes will be wrought in the cybersecurity area as a result of machine learning [5]. Three million new malware samples might be represented in only one hour. Antivirus and malware detection systems that are more than a decade old are unable to deal with the sheer volume and sophistication of today's threats. Cybercriminals are always inventing new malware and attack techniques to circumvent network and endpoint detection. Modern approaches, including machine learning, are needed to counter the ever-growing danger of malware. Using the machine learning methods outlined in this proposal, cyber security professionals can detect and bring attention to complex malware. The results of our preliminary analysis are presented in detail, and we also discuss potential areas for future study that might enhance machine learning.

The Impact of Machine Learning on Intrusion Detection Systems: A Comparative Survey

Because of the skyrocketing growth of laptop networks and user content consumption, trustworthy networks are urgently needed. Developing a supply of trustworthy automated methods to detect attack scenarios is crucial, since the frequency of network assaults has been shown to be rising over time. Intruder detection systems, which are a kind of attack system, may detect incursions that have returned from the internet. Several approaches have been laid forth in the literature on network intrusion detection. Recently, mining algorithms became popular for displaying intrusion detection [6]. It was possible to identify incoming intrusions by comparing the supplied network information with the well-mined data. Discovering two or more items in the well-mined data that have the same attributes is called an incursion. Several intrusion detection models were developed and used in this investigation to improve accuracy, meeting these requirements. A quick assessment is carried out on the sooner approaches. In all, the approach consists of two distinct procedures: one for data preparation, and another for detection. Information preparation operations are categorized into feature extraction, and there are transformation models that provide operating ways over the possibilities. Similarly, the detection methods are classified using machine learning and organic process methodologies.

Modelling Enhancement for Cybersecurity

All audits pertaining to IT, data security, or compliance follow the same standard procedure: planning, goal and scope definition, engagement terms explanation, audit execution, evidence provision, risk assessment, audit reporting, and follow-up activity planning. The framework of a cybersecurity audit is identical to that of any other audit. Nevertheless, a great deal of work will be required because of the high quality of the many cybersecurity domains. However, most cyber capabilities are not covered by the scope of the internal audits. An integral part of this specific framework are management reviews, cyber risk assessments, information management and protection, risk analytics, crisis management and resilience, security operations, security awareness and training, development life cycle, security program, third-party management, information/asset management, access management, threat or vulnerability management, and the implementation of cybersecurity controls as part of an overall strategy and framework. Not only that, but Deloitte's framework is interoperable with other trade frameworks like NIST, ITIL, COSO, and ISO.

Due to the fast-paced nature of change and the absence of standards for real-time audits, cybersecurity audits are also a poorly understood issue. Khan claims that auditors need to involve the appropriate departments—client operations, financials, human resources, law, purchasing, regulatory affairs, physical security, and any pertinent third parties with whom the company does business—in order to thoroughly plan a cybersecurity audit.

Database Intrusion Detection System Using Octraplet and Machine Learning.

When it comes to intrusion detection systems, host systems and networks are often the targets. The field of information intrusion detection is, without a doubt, severely lacking in foundational works. A technique proposed by Chung et al. in a recent paper proposes an approach to information intrusion detection that is based on misuse detection. Here, typical data patterns have been extensively mined and continue to exist as classic profiles. The main negative is that there are no role profiles. According to their job descriptions, the users do a wide range of activities. Searching through user profiles is not enough. Users may carry out actions that are supported by roles, and malicious intent can be identified. Lee et al. created an intrusion detection system that operates in real-time by using time signatures. There is a temporal component to real-time information systems, and the values of this component do change over time [7]. With each upgrade, a new gadget dealing is born. Updates to the temporal information are performed over a predetermined time period. Any attempt by a transaction to modify the updated temporal information by more than that amount will cause an alert to go out. The fact that this method places all of its focus on updates and not on job profiles is a possible downside. Using log data, Hu Panda generates user profiles. Tables and data that are often used for research and those that are kept for comparison. The problem with this approach is that knowledge becomes very difficult to hold on to when both the number of users and the dimensions of information are rapidly increasing. Because modern cyber dangers are so dynamic and diverse, relying just on human intervention is insufficient. Machine learning also provides capacity and improved speed when dealing with large quantities of attacks with various variations. Combining AI with human intellect—a combination of speed, strength, talents, and judgment—is the $64,000 key to investing in cyber protection. Cyber security breach investigations may benefit greatly from the use of machine learning and AI. Machine learning makes human work much more efficient and precise. We can assist the US in detecting cyber security attacks by using various machine-learning approaches.

III PROPOSED METHODOLOGY

The approach used to forecast the susceptibility of organizational endpoints to cyberattacks is detailed in this section. A number of critical modules make up the technique, and they all deal with different parts of data analysis and machine learning. Information Gathering, Data Pretreatment, Feature Selection, Model Building, and Model Assessment are the components that make up this suite of modules.

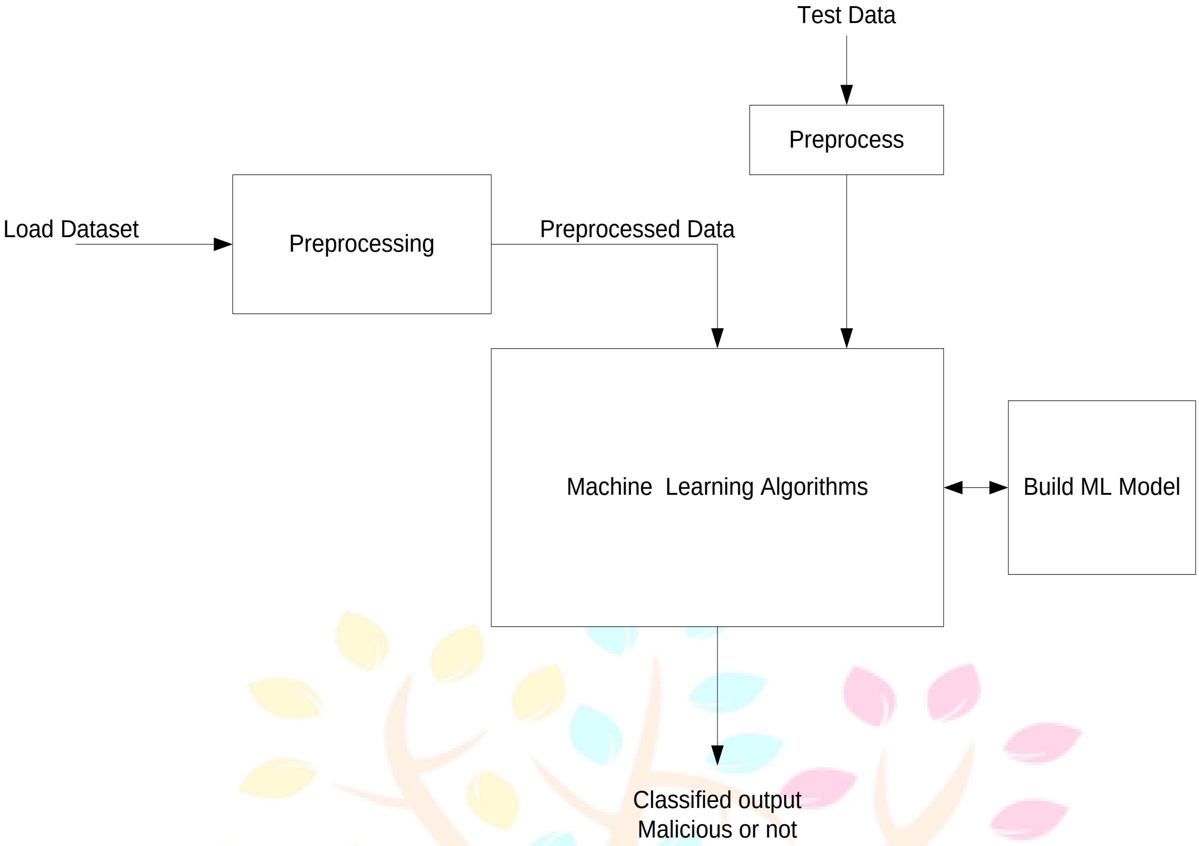


Figure 1: Proposed Architecture

1. Gathering Information

Identify endpoints' characteristics and the cybersecurity risks they face by collecting pertinent data.

Various endpoint features, such as device type, operating system, installed software, network settings, and historical records of cyber events, are included in a publicly accessible cybersecurity dataset gathered from Kaggle.

Quantitative and qualitative variables coexist in the dataset, with some missing values that must be filled up prior to analysis.

2. Cleaning Up Your Data

In order to clean, convert, and encode the raw data in preparation for machine learning analysis.

Dealing with Data Absences: One typical problem with cybersecurity databases is missing data. For missing value cases, this research employs a variety of imputation methods. The research makes sure there isn't any bias and that the dataset is strong by repeatedly imputing missing data.

In order to make them more amenable to machine learning algorithms, categorical variables like operating system and device type are transformed into numerical representations using data transformation processes like one-hot encoding.

To enhance the convergence of gradient-based learning algorithms and to make sure that all features contribute equally to the model, numerical features are normalized.

3. Feature Selection

The goal is to determine which characteristics are most important for endpoint vulnerability prediction.

We start by looking at the characteristics and how they relate to the endpoint vulnerability, which is our target variable. To simplify the model, we eliminate variables that aren't important or redundant and give more priority to those with strong correlation.

To take the selection process to the next level, we use advanced approaches including feature significance ratings using tree-based models. The characteristics that significantly affect the model's prediction performance may be better identified with the use of these ratings.

4. Constructing Models

In order to create machine learning models that can forecast the probability of a cyberattack compromising a certain endpoint.

The research improves the accuracy and stability of the models by using ensemble learning methods, particularly boosting and bagging algorithms.

Boosting: This approach, which encompasses algorithms like as AdaBoost and Gradient Boosting, constructs models in a sequential fashion, with each iteration aiming to rectify the shortcomings of its predecessor. Incorporating the strengths of several underperforming learners into a single robust prediction model is facilitated by this.

One method, known as "bagging," uses methods like Random Forest to train several models simultaneously on separate data sets and then average their predictions. Bagging is useful for avoiding overfitting and lowering variance.

To train the models, we first use the preprocessed and chosen features, and then we tune their hyperparameters so that they perform as well as possible. The models' ability to generalize to new data is tested via cross-validation.

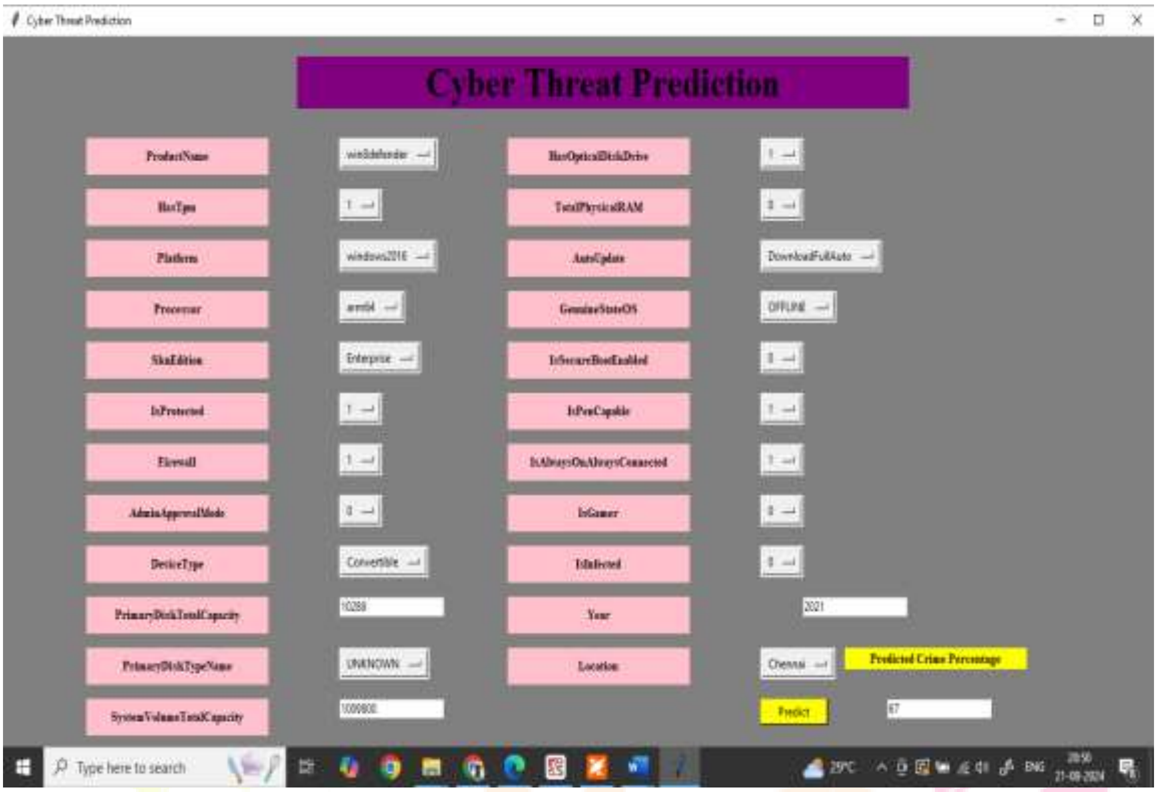
5. Evaluating the Model

In order to find out how well the created models predicted endpoint vulnerabilities and how well they performed overall.

To find the optimal method, it is necessary to compare the results of several models. With an eye toward their potential use in practical cybersecurity situations, the findings are examined to comprehend the benefits and drawbacks of each method.

IV. RESULTS AND DISCUSSION

This section details the study's user interface and the outcomes it produced. In addition to displaying crime-related information, the user-friendly interface includes input capabilities for forecasting endpoint susceptibility to cyber-attacks.



Cyber Threat Prediction

ProductName	win8defender	HasOpticalDiskDrive	1
HasTpm	1	TotalPhysicalRAM	0
Platform	windows2016	AutoUpdate	DownloadFullAuto
Processor	arm64	GenuineStateOS	OFFLINE
SkuEdition	Enterprise	IsSecureBootEnabled	0
IsProtected	1	IsPenCapable	1
Firewall	1	IsAlwaysOnAlwaysConnected	1
AdminApprovalMode	0	IsGamer	0
DeviceType	Convertible	IsInfected	0
PrimaryDiskTotalCapacity	10288	Year	2021
PrimaryDiskTypeName	UNKNOWN	Location	Chennai
SystemVolumeTotalCapacity	1009800	Predicted Crime Percentage	67

Figure 2: User Interface shows the output crime percentage on the user inputs.

Figure 2 shows the user interface (UI) that will be used to anticipate the vulnerability of an organization's endpoints to cyberattacks based on the information that users provide. In order to give extra insights, the UI also allows visualization of crime information, including the crime rate broken down by year.

Type of Endpoint, OS, installed applications, network settings, and security incidents history are some of the endpoint parameters that users may provide. To determine the endpoint's vulnerability, these data points are incorporated into a machine learning model. The user interface shows the expected perrcent chance that the endpoint is infected with malware or comparable threats after processing the inputs. Users can easily grasp the amount of danger thanks to this percentage, which is computed using the model's output.

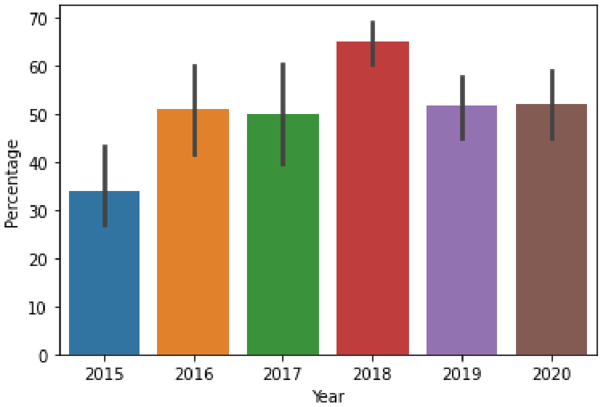


Figure 3: Graph of Year-wise Crime Rate.

Users may see patterns in crime data over time, with an emphasis on cybercrime, in Figure 3, which displays the visualization of year-wise crime rate.

Information Obtained: The study's cybersecurity dataset is where the data for the crime rates is retrieved year-wise. Included in this data set are year-by-year records of cyber occurrences.

The user interface has a bar graph that shows the annual crime rate graphically. The number of cyber events registered for a given year is represented by the height of the corresponding bar.

Analysis: Users may quickly see patterns in cybercrime over time, such spikes or dips in the event count, thanks to the bar graph. With this data, we can trace the history of threats and prepare for the ones to come.

V. CONCLUSION

Predicting the susceptibility of organizational endpoints to assaults, especially malware, has been tackled by this research. Cybersecurity measures are more necessary than ever before due to the increasing dependence on digital networks and devices, which in turn increases the danger environment. This study offers significant insights that may help firms boost their cybersecurity defenses by concentrating on the vulnerability of different endpoints, such as PCs, smartphones, tablets, and servers.

Extensive use of ensemble learning methods such as boosting and bagging, as well as multiple imputation to deal with missing data, were used in the study to construct accurate prediction models. To guarantee that these models are applicable to real-world situations, they were trained and verified using a publicly accessible cybersecurity dataset. In order to determine which devices inside an organization are most susceptible to cyberattacks, the findings showed that certain endpoint attributes are significantly associated with a higher risk of such assaults.

The data-driven approach to improving organizational cybersecurity is one of the main contributions of this research. Organisations may improve the efficacy of their security measures, budget allocation, and cybersecurity strategy execution by precisely identifying which endpoints pose the most threat. Preventing major financial and reputational damages that might occur from cyber-attacks is possible with this proactive strategy, which also assists in minimizing prospective dangers.

Finally, this study offers a thorough methodology for evaluating and enhancing firms' cybersecurity resilience. The research adds to the continuing fight against cyber threats by providing fresh insights into endpoint security and the use of machine learning methods. To further strengthen the accuracy and usefulness of the prediction models, future study might build on this research by examining other traits and datasets and adding real-time threat information.

ACKNOWLEDGMENT

Thanks to the Management of Cambridge Institute of Technology and my Research Guide “NAME” for providing the e-resources and guidance.

REFERENCES

[1] J. Cano, "Cyberattacks-The Instability of Security and Control Knowledge", *ISACA Journal*, vol. 5, pp. 1-5, 2016.

[2] C. Hollingsworth, "Auditing from FISMA and HIPAA: Lessons Learned Performing an In-House Cybersecurity Audit", *ISACA Journal*, vol. 5, pp. 1-6, 2016.

[3] Li X, Wang J, Zhang X, “Botnet Detection Technology Based on DNS”, J. Future Internet, 2017.

[4] 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.

[5] Dinil Mon Divakaran et al., "Evidence gathering for network security and forensics", *Digital Investigation*, pp. 56-65, 2017.

[6] 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, 2016.

[7] M. Khan, "Managing Data Protection and CybersecurityAudit's Role", *ISACA Journal*, vol. 1, pp. 1-3, 2016