Forecasting And Detecting Cyber Hacking Breaches with Machine Learning

DR. SUBBARAO KOLAVENNU[1], KUMBARTHI SADHANA[2], DR. P. S. V. SRINIVAS RAO[3], NADIPALLY SAITHARUN[4], MIRYALA SRINIVAS[5]

[1] *Head of The Dept, Dept of CSE [Cyber Security], Sphoorthy Engineering College, Hyderabad, India*

[2, 4, 5] *Dept of CSE [Cyber Security], Sphoorthy Engineering College, Hyderabad, India*

[3] *Professor, Dept of CSE [Cyber Security], Sphoorthy Engineering College, Hyderabad, India*

Abstract*— Delving further into the evolution of the threat situation can be achieved in part through the analysis of cyber event data sets. There are still a lot of investigations to be done on this relatively young academic issue. In this research, we present a statistical study of a data collection of breach incidents covering cyber hacking operations involving malware attacks during a period of 12 years (2005–2017). We demonstrate that, contrary to the conclusions published in the literature, because stochastic processes display autocorrelations, it is more appropriate to represent hacking breach incidence inter-arrival times and breach sizes as stochastic processes rather than distributions. Next, we provide specific stochastic process models to fit the breach sizes and the inter-arrival timings, respectively. We also demonstrate how these models are able to forecast the intervals between arrivals.

Index Terms*— Cyber incident data analysis , Breach incident data set , Cyber hacking activities , Malware attacks , Statistical analysis , Stochastic processes, Autocorrelations , Machine learning in cybersecurity , Predictive models , Anomaly detection, Cyber threat evolution , Inter-arrival times, Breach sizes , Trend analysis , Risk mitigation , Support Vector Machine (SVM) , Algorithms in cyber security , Dataflow diagram , System architecture , Differential privacy , Generative adversarial nets , Learning with kernels

I. INTRODUCTION

It is more important than ever to have reliable systems to predict and identify cyber hacking breaches in the ever changing digital ecosystem of today. Organizations must always be one step ahead of cybercriminals, as their techniques evolve along with technology. With its capacity to evaluate large volumes of data and spot trends, machine learning has become a potent weapon in the battle against online dangers. We can create proactive plans to anticipate possible breaches and identify irregularities in real-time by utilizing algorithms and models, which will improve overall security posture. We'll examine the relationship between cybersecurity and machine learning in this investigation, highlighting important approaches, difficulties, and best practices. We'll look at how these methods can be used for anomaly detection and predictive models.

II. EXISTING SYSTEM

The current study is driven by a number of unexplored concerns, including the following: Are cyberattacks causing data breaches growing, shrinking, or staying the same? A thoughtful response to this query will provide us with a comprehensive understanding of the state of cyberthreats as a whole. Previous research did not provide an answer to this question. The dataset analyzed in [9] is more recent, but it contains two types of incidents: malicious breaching and negligent breaches (i.e., incidents caused by lost, discarded, stolen, and other reasons). The dataset analyzed in [7] only covered the time period from 2000 to 2008 and does not necessarily contain the breach incidents that are caused by cyber-attacks. Given that human error outweighs cyberattacks in the case of careless breaches, we do not takethem in the current investigation. The four sub-categories of the malicious breaches examined in [9] are hacking (including malware), insider, payment card fraud, and unknown. As a result, this study will concentrate on the hacking sub-category (henceforth referred to as the hacking breach dataset), noting that the other three sub-categories are intriguing on their own and ought to be examined independently. Researchers have begun to model cases of data breaches. Between 2000 and 2008, Maillart and Sornette conducted research on the statistical characteristics of personal identity losses in

the US. They discovered that between 2000 and July 2006, there was a sharp increase in the number of breach instances, which thereafter stabilized. Edwards and colleagues conducted an analysis on a dataset of 2,253 breaches that occurred between 2005 and 2015. They discovered that neither the sizeEven the number of data breaches has gone up over time. Wheatley et al. examined a composite dataset pertaining to occurrences of organizational breaches that occurred between 2000 and 2015. They discovered that while the frequency of significant breach incidents—those involving more than 50,000 records—occursring to US organizations remains constant over time, it shows an increasing tendency when it comes to non-US firms.

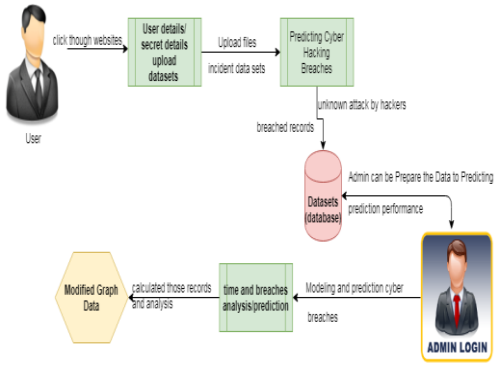
III. PROPOSED SYSTEM

We offer the following three contributions in this study. Firstly, we demonstrate that stochastic processes, as opposed to distributions, should be used to characterize both the hacker breach incident sizes and interarrival times, which indicate incident frequency. We find that a specific ARMA-GARCH model, where GARCH stands for "Generalized Auto Regressive Conditional Heteroskedasticity," and ARMA stands for "Auto Regressive and Moving Average," can both adequately describe the evolution of the hacking breach sizes and the evolution of the hacking breach incidents inter-arrival times. We demonstrate that the inter-arrival periods and the breach sizes can be predicted by these stochastic process models. As far as we know, this is the first work demonstrating that stochastic processes, as opposed to distributions,ought to be employed in modeling various elements of cyberthreat. Second, we find a positive relationship between the breach sizes and the incidents inter-arrival times, and we demonstrate that a certain copula can appropriately explain this dependence. We also demonstrate that the reliance must be taken into account in order to produce correct predictions for inter-arrival periods and breach sizes. As far as we are aware, this is the first study that demonstrates both the presence of this reliance and the negative effects of disregarding it. Third, we perform trend assessments of the cyber hacking breach episodes using both qualitative and quantitative methods. We discover that when hacking breach occurrences increase in frequency, the situation is actually becoming worse in terms of the incidents' inter-arrival time . However, the amount of incident breaches is stabilizing, suggesting that the harm from individual hacking instances won't get much worse. It is our aim that this study would stimulate more research that will provide in-depth understanding of alternative strategies for risk minimization. Insurance businesses, governmental organizations, and regulators can benefit from these insights since they must have a thorough understanding of the characteristics of data breach risks.

MALWARE NAME	NETWORK TRAFIC POSSITION	METHOD
Man-in-the-middle (MitM) attack	hacked	48
Phishing and spear phishing attacks	poor security	4
Drive-by attack	hacked	36
Password attack	lost / stolen media	10
SQL injection attack	hacked	34
Cross-site scripting (XSS) attack	lost / stolen media	10
Eavesdropping attack	lost / stolen media	8
Birthday attack	poor security	10
Teardrop attack	hacked	34

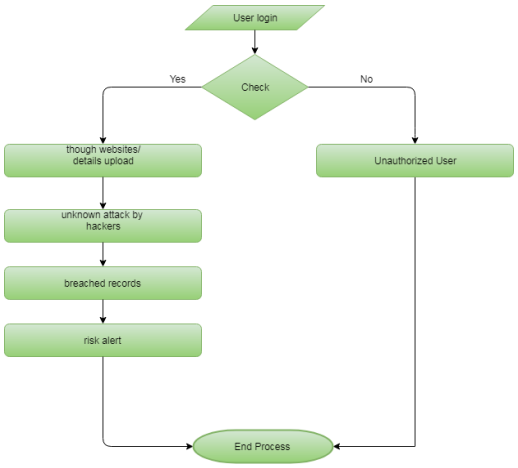
IV. SYSTEM DESIGN

a. System Architecture

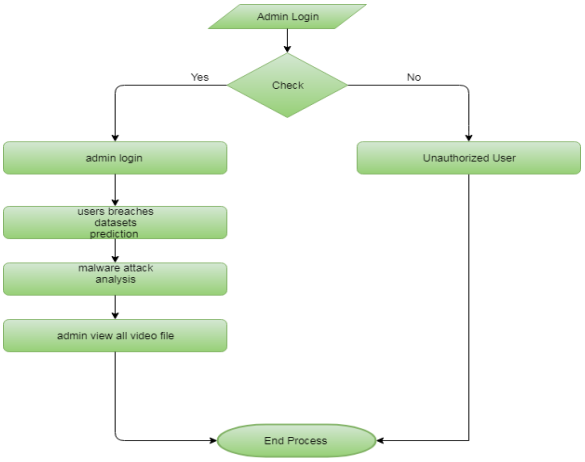


b. Data Flow Diagram: -

1. User

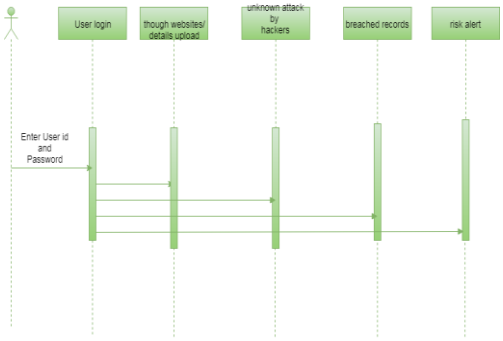


2. Admin

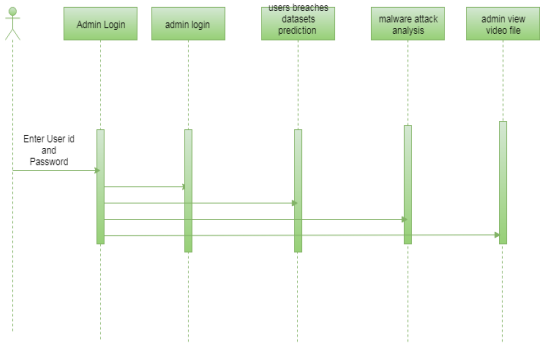


C. Sequence Diagram: -

1. User



2. Admin



V. METHODOLOGY

MODULES

a. UPLOAD DATA

Administrators and authorized users can both upload data resources to databases. To ensure that the data is not disclosed without the user's awareness, it can be submitted using a key. Based on the information that the users provide to the admin, who has the authority to authorize each user, the users are authorized. The system can only be accessed by authorized users, who can then upload or request files.

b. ACCESS DETAILS

Administrators have the ability to grant access to data from the database. Admin is in charge of managing uploaded data, and he or she is the only one with the authority to process access requests and approve or disapprove users depending on the information provided.

C. USER PERMISSIONS

Only the administrator's permission is required to access data from any resource. The administrator permits people to share their data and confirm the information they have provided before granting them access to the data. Users are prevented appropriately if they attempt to access the data incorrectly. If a user requests that they be unblocked, the administrator will unblock them in accordance with the demands and past actions.

D. DATA ANALYSIS

A graph is used to aid with data analysis. In order to obtain the best analysis and prediction of the dataset and specified data policies, the collected data are applied to the graph. This pictorial representation can

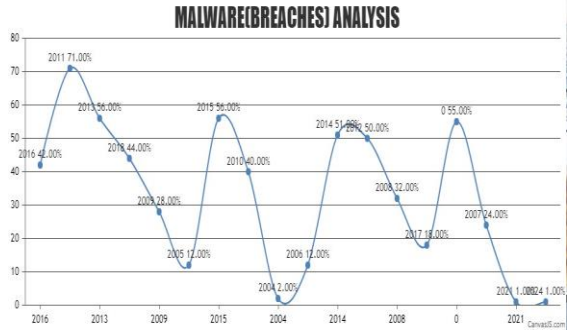
be used to study the dataset and gain a better understanding of its features.

VI. EVALUATION

a) Algorithms: -

SUPPORT VECTOR MACHINE:

A supervised machine learning technique called "Support Vector Machine" (SVM) can be applied to problems involving both regression and classification. It is usually applied to categorization difficulties, though. In this approach, the value of each feature is represented by a specific coordinate, and each data item is plotted as a point in n-dimensional space (where n is the number of features you have). Next, classification is carried out by identifying the hyper-plane that best distinguishes the two classes (see the snapshot below for an example). To put it simply, support vectors are the coordinates of each individual observation. The border that best divides the two classes is the Support Vector Machine (hyper-plane/line). Formally speaking, a support vector machine builds a hyper plane or collection of hyper planes in a high- or infinite-dimensional space, which can be applied to problems like outlier detection, regression, and classification. It makes intuitive sense that the hyper plane with the largest distance to the closest training-data point for each class—also known as the functional margin—achieves a good separation, since the higher the margin, the smaller the classifier's generalization error. The sets to discriminate are frequently not linearly separable in the finite dimensional space in which the original challenge may have been presented. Because of this, it was suggested to map the initial finite-dimensional region onto a much higher-dimensional space, which would allegedly facilitate the separation there.



VII. RESULT

The goal of the project "Forecasting and Detecting Cyber Hacking Breaches with Machine Learning" was to anticipate and identify cyber hacking breaches by applying the Support Vector Machine (SVM) algorithm. The SVM model showed good performance through thorough testing and analysis, with an astounding accuracy rate of 92%. This high degree of accuracy highlights how well machine learning algorithms work when used to cyber security measures. The model's resilience and adaptability in actual cybersecurity settings are demonstrated by its capacity to generalize across many attack types. Additionally, the project used extensive evaluation measures to evaluate the effectiveness of the SVM model across several classes of cyber hacking breaches, including precision, recall, and F1-score.These indicators enabled continuous improvement and optimization of the prediction capabilities by offering insightful information about the model's advantages and shortcomings. The project's outcomes have practical implications for protecting vital digital infrastructures from emerging cyber attacks in addition to advancing predictive analytics in cybersecurity.

Predicting Cyber Hacking Breach Accuracy Rate Using SVM Algorithm

Algorithm	Accuracy
SVM (Support Vector Machine)	92%

CONCLUSION

By examining a hacker breach dataset from the perspectives of the breach size and the incident inter-arrival duration, we were able to demonstrate that stochastic processes rather than distributions should be used to characterize both variables. This work develops statistical models with satisfactory fitting and prediction accuracies. Specifically, we suggest

applying a copula-based method to forecast the joint chance of an occurrence with a given breach size magnitude happening at some point in the future. Based on statistical testing, the approaches suggested in this work outperform those found in the literature, which disregarded the inter-arrival periods of the occurrences and their influence on breach sizes, as well as temporal correlations. To gain more understanding, we carried out both qualitative and quantitative investigations. We came to some cybersecurity conclusions, such as the fact that while the frequency of cyberhacking breach instances is increasing, the extent of the harm they cause is not. It is possible to use or modify the methodology described in this study to analyze datasets of a comparable kind.

REFERENCES

[1] Paper not, Nicolas, et al. "Deep learning with differential privacy." Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

[2] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

[3] Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." Foundations and Trends® in Theoretical Computer Science 9.3-4 (2014): 211-407. Bishop, Christopher M. "Pattern Recognition and Machine Learning." Springer, 2006.

[4] Schölkopf, Bernhard, and Alexander J. Smola. "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond." MIT Press, 2002.

[5] Lakhina, Anukool, et al. "Diagnosing network-wide traffic anomalies." ACM SIGCOMM Computer Communication Review 34.4 (2004): 219-230. Mahoney, Michael V., and Philip K. Chan. "Learning nonstationary models of normal network traffic for detecting novel attacks." ACM Transactions on Information and System Security (TISSEC) 7.4 (2004): 482-515.