Journal of Engineering Excellence
ISSN: Applied | Volume 1 Issue I Oct-Dec 2025 | Inaugural Edition
www.jeejournal.com JEE-03

# A MACHINE LEARNING APPROACH TO CYBER BREACH PREDICTION USING SUPPORT VECTOR MACHINES

**J.SWATHI,** *Assistant Professor & HOD,*
*Department of CSE,*
**TRINITY COLLEGE OF ENGINEERING & TECHNOLOGY(AUTONOMOUS), PEDDAPALLI.**

**ABSTRACT:** The evolution and development of cyber dangers can only be understood with the use of data collected from cyber occurrences. Since this field of study is still in its infancy, there is much more to discover. Cyberespionage activities, primarily malware infections, that occurred between 2005 and 2017 are the focus of this statistical study. We found that distributions based on autocorrelation do not adequately describe the frequency and severity of cyberattacks. Rather, random procedures work better. Then, we display the bespoke stochastic process models we developed to account for different breach magnitude and inter-breach time span. On top of that, we prove that these models are highly predictive of both arrival timeframes and damage levels. We use quantitative and qualitative trend studies to find out how cyber events have evolved over the years. Concerning security, one can say many things. The frequency of leaks has increased, but the total number of leaks has remained unchanged.

***Keywords:*** *Analysis cyber incidents, stochastic process, prediction of hacking.*

## 1. INTRODUCTION

Information theft is one of the most serious issues that can arise during an online contact. There were 7,730 instances of unlawful access to data that allowed 9,919,228,821 records to fall into the wrong hands between 2005 and 2017, according to the Privacy Rights Clearinghouse. There were 1,093 cases of data theft in 2016, according to Cyber Scout and the Identity Theft Resource Center. The number of complaints has increased by 40% from 780 in 2015. Data breaches affected 4.2 million active and retired federal workers in 2015, according to the US Office of Personnel Management (OPM). To learn more about background checks, we also combed through federal contractor and employee data. The files contained 21.5 million Social Security numbers. Organizations risk losing a substantial amount of capital due to data breaches. IBM discovered in 2016 that the average cost of losing or misusing a record containing private or sensitive information was $158. According to Net Diligence, 1,339 records were compromised in 2016, with an average cost of $39.82 per record. The worst case scenario included costs of $665,000. The cheapest item cost sixty thousand dollars. Data breaches are troublesome regardless of the availability of technological solutions to prevent hackers from gaining access to computer systems. As a result, we need to demonstrate the evolution of data leaks. The frequency of data breaches and potential mitigation strategies, such as insurance, will be the focus of this research. Due to a lack of comprehensive modeling tools and an incomplete

understanding of data theft, we are unable to create reliable cyber risk indicators for the distribution of insurance premiums. Despite widespread agreement that insurance is a worthwhile investment, this issue persists. Models of data breach attempts have just lately been developed by researchers. Data on cases of identity theft in the United States were examined between the years 2000 and 2008. The number of persons taking longer holidays increased dramatically from July 2000 to July 2006. Finally, a state of peace was reached.

## 2. LITERATURE SURVEY

Hammouchi et al. wrote it. Al designed a STRisk forecasting system that uses social media seamlessly to make the prediction assignment more relevant. About 3,800 US businesses were investigated. Victims and non-victims were involved. Any company's profile includes externally judged social and technical traits. The researchers admit non-victim sample flaws and offer a solution for organizations with inaccurate IDs to reduce unreported cases. They use machine learning models to predict how soon hackers could get into a company's security. Social and technical difficulties can be overcome to reach an AUC exceeding 98%. This result beats the technical-only AUC by 11%. Our research shows that expired certificates and accessible ports are the most dependable technical signals. However, agreeableness and information dissemination are the best social accomplishment indicators. Mandal continued the party. Social mood categories were refined by everyone's social feelings, interactions, and events. The proposed system organizes and notifies users about major social events. Twitter text data was analyzed using an aspect-based sentiment model. It works better than modern approaches.

All Poyraz's workers. Multifaceted data invasion costs are analyzed. Here's how to calculate data breach costs. The system separates stolen US citizen data into SPII and PII. They assess the independent variables' effects on multiple levels using polynomial and factorial algebra in a complex stepwise regression approach. Three major advances have occurred in this field. Our strategy first links data breach costs to income, PII leaks, and group lawsuits. Separate "sensitive" from "non-sensitive" concealed information gives a more complete view of expenses than previous research. Finally, all independent variables interact factorially. Guru Akhil and colleagues studied 2005–2018 cyber espionage data, notably breaches. They found that stochastic cycles better describe hacking attempt frequency, length, and severity than distributions. Internal reliance is the cause, and cycles show it. After that, they recommend testing several stochastic cycle models to see which best depicts the break length-entry time relationship. These algorithms properly predicted large breaches on the 21st and 22nd. Their data is analyzed using quantitative and qualitative patterns to understand about cyber attack strategies quickly. Network security data supports the claim that intrusions are becoming more serious and less frequent.

The Fang Gang. Our team sought strategies to analyze and predict organizational data breach risk to begin our investigation. A corporation with few

incursions swiftly abandons conventional statistical models since there isn't enough data to train them. They suggest a novel statistical strategy that uses the interconnectedness of many time series as a first step toward a solution. The methodology was tested using real-world corporate hackers. The inquiry shows its ability to assess and predict corporation wrongdoing.

criminal group created by Kure. The assets it protects, the efficacy of present measures, and the risk types forecast determine the quality of cyber security risk management (CSRM). The proposed unified approach uses a comprehensive assessment model (CAM) to assess defenses, machine learning classifiers to detect threats, and fuzzy set theory to value assets. To assess risk, examine the linkages between VERIS community dataset (VCDB) subsets and key CSRM concepts such asset, threat agent, attack pattern, strategy, method, and procedure (TTP). The experiment found that fuzzy set theory improves risk management by determining item importance. The findings show that machine learning models can detect cyber espionage, crime ware, and denial of service attacks. Accurate risk prediction helps organizations determine the best strategies to prevent issues.

The paper was written by Subramanian et al. An algorithm based on machine learning was created to protect websites against hostile hacking. We focus on a machine learning model that can learn from data and adapt to new assault patterns. This model will monitor a system or website live. One solution is a Django-based web app. This software would collect data from Shop Clues, Amazon, Flipkart, and Snap deal. This app lets you securely see data online. The suggested framework includes regular site monitoring. Our platform's data will be secure and organized to prevent theft. The algorithm's daily forecasts use the latest assault data and public information. We will use pre-existing datasets and website hack and attack data to develop this model.

.

## 3. PROPOSED SYSTEM

Three perspectives are used to analyze this topic. To begin, we show that stochastic processes better characterize the average time between cyber breach incidents and the total number of breaches than static distributions. You can estimate frequency using the median interval between them. The hacker break-in time was determined using a specific point technique. We also found that a certain ARMA-GARCH model represents criminal breach magnitudes over time well. GARCH means "Generalized Auto Regressive and Conditional Heteroskedasticity," while ARMA means "Auto Regressive and Moving Average." Nobody has proved that stochastic processes are better than distributions for simulating significant cyber risks. Furthermore, we demonstrate that a copula captures the direct link between the number of occurrences and their timing. We also show how important dependencies are when projecting defense distances and arrival times. Not considering this will lead to erroneous projections. We know of no other study that analyzes such a link and the consequences of ignoring it. We employ qualitative and quantitative methods to uncover cyber attack patterns. Cyber

Journal of Engineering Excellence
ISSN: Applied | Volume 1 Issue I Oct-Dec 2025 | Inaugural Edition
www.jeejournal.com JEE-03

invasions have been relatively consistent in severity, while the intervals between breaches have decreased. Further invasions are unlikely to worsen the damage.

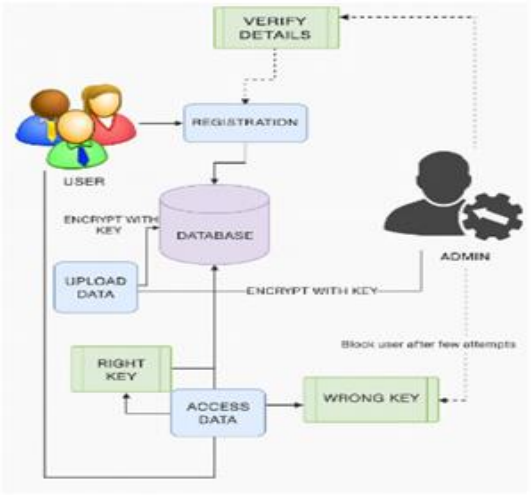



Fig. 1: block diagram of proposed system.

We hope our study will stimulate other researchers to dig deeper and find novel risk-reduction methods. Understanding data breach intricacies benefits politicians, government agencies, and insurance companies.

**Support Vector Machine**

SVMs tackle classification and regression problems in supervised machine learning. The word is widely used in classification debates. Each data point represents a feature, and the number of features is proportional to the space's dimensions. This matrix expresses features by points representing their values. Next, find the data-cutting hyper plane (see image). Support vectors are based on observation data. A support vector machine (SVM) generates hyper planes in a multidimensional space to do classification, regression, and outlier detection. The functional margin, the hyper plane at the maximum distance from the nearest training-data point in any class, can identify classes. This notion is supported by a wider margin lowering the classifier's generalization error. Even in a small space, distinguishing between areas of interest is not always as easy as drawing a straight line. The objective was to substantially extend the little space to emphasize contrast.

## 4. RESULTS MODULES UPLOAD DATA

User permission from the database administrator allows data resource addition. Data encryption protects it against unlawful access. Only the administrator can authorize utilizing the information given. Only authorized users can upload and download files.

**Access Details**

Users can access database information with administrator permission. Only the administrator may manage user accounts and assign access rights based on credentials. The administrator must ensure data integrity during transfer.

**User Permissions**

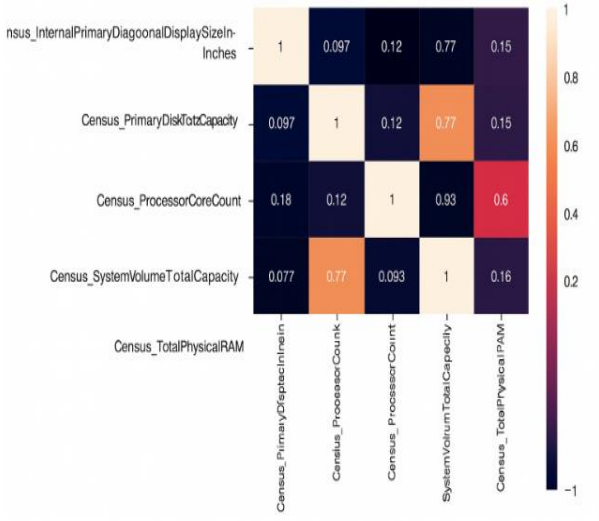
Users without administrator privileges cannot access resource data. If the user verifies that giving the information is appropriate, the administrator will authenticate their identification. If they fail many times, a user will be permanently barred from accessing the same data. Management will consider the user's requests and history before unblocking.

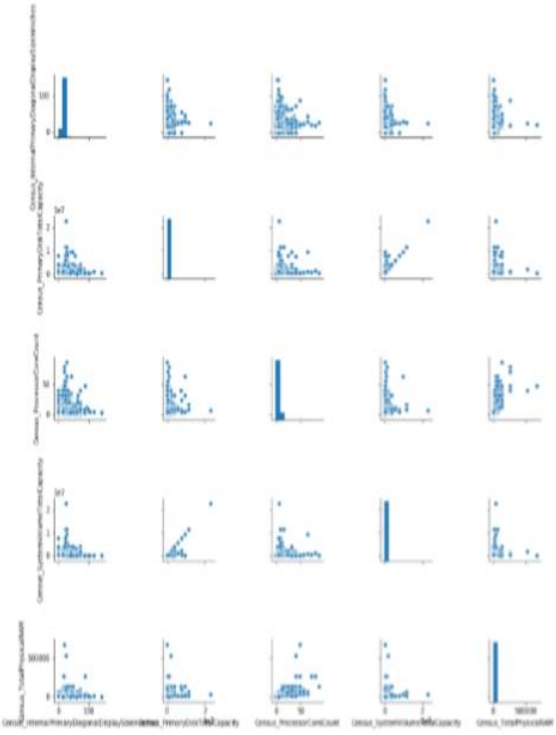
**Data Analysis**

Graphs are often used to analyze data. Data visualization enables for precise analysis and projection while complying to data standards. Visualizing data helps readers understand it.

Journal of Engineering Excellence
ISSN: Applied | Volume 1 Issue I Oct-Dec 2025 | Inaugural Edition
www.jeejournal.com
JEE-03

## EDA results

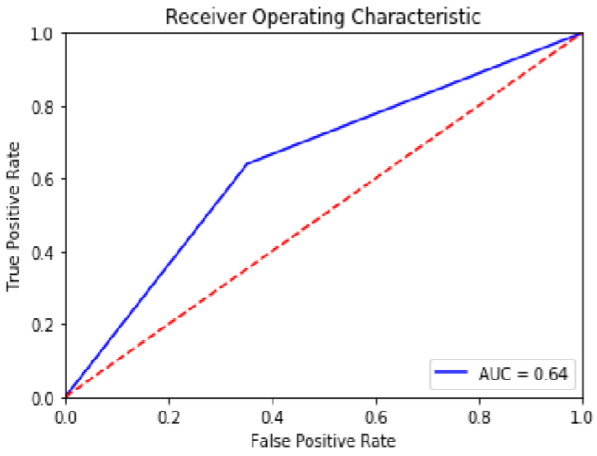


## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.64 | 0.65 | 0.65 | 49659 |
| 1 | 0.64 | 0.64 | 0.64 | 49341 |
| accuracy | | | 0.64 | 99000 |
| macro avg | 0.64 | 0.64 | 0.64 | 99000 |
| weighted avg | 0.64 | 0.64 | 0.64 | 99000 |

## Feature Importance




| | feature | importance | normalized_importance | cumulative_importance |
|---|---|---|---|---|
| 0 | index | 131055 | 0.068687 | 0.068687 |
| 1 | AvSigVersion | 129938 | 0.068102 | 0.136789 |
| 2 | CityIdentifier | 118341 | 0.062024 | 0.198812 |
| 3 | Census_InternalPrimaryDiagonalDisplaySizeInInches | 108861 | 0.057055 | 0.255867 |
| 4 | Census_SystemVolumeTotalCapacity | 106640 | 0.055891 | 0.311758 |
| ... | ... | ... | ... | ... |
| 65 | Census_IsAlwaysOnAlwaysConnectedCapable | 185 | 0.000097 | 0.999942 |
| 66 | OsVer | 111 | 0.000058 | 1.000000 |
| 67 | Census_IsPortableOperatingSystem | 0 | 0.000000 | 1.000000 |
| 68 | Census_DeviceFamily | 0 | 0.000000 | 1.000000 |
| 69 | SMode | 0 | 0.000000 | 1.000000 |

70 rows × 4 columns

## 6. CONCLUSION

Analyzing a cyber event database, we calculated the average time between assaults and disclosure size. Our research reveals that stochastic processes are better for expressing these two variables than ranges. This study's statistical models excelled at fitting data and making predictions. We recommend utilizing a copula-based technique to estimate the chance of an incoming event with a specific violation amount. Evaluations used qualitative and quantitative methodologies to improve understanding. Our cybersecurity survey shows that cybercrime is rising but not getting worse. Use these strategies to analyze connected data.

**Future work**

More research is needed now and in the future to resolve several concerns. Estimating large amounts and correcting information gaps like publicly publicized security vulnerabilities are intriguing and difficult. Knowing potential breaches is a bonus. To maximize forecast accuracy, breach predictability research is needed.

**REFERENCES**

1. P. R. Clearinghouse. Privacy Rights Clearinghouse's Chronology of Data Breaches.
2. Accessed: Nov. 2017. [Online]. Available: https://www.privacyrights.org/data-breaches
3. ITR Center. Data Breaches Increase 40 Percent in 2016, Finds New Report From Identity Theft Resource Center and Cyber Scout. Accessed: Nov. 2017. [Online]. Available: http://www.idtheftcenter.org/2016databreaches.html
4. C. R. Center. Cybersecurity Incidents. Accessed: Nov. 2017. [Online]. Available: https://www.opm.gov/cybersecurity/cybersecurity-incidents
5. IBM Security. Accessed: Nov. 2017. [Online]. Available: https://www.ibm.com/security/data-breach/index.html
6. Net Diligence. The 2016 Cyber Claims Study. Accessed: Nov. 2017. [Online]. Available: https://netdiligence.com/wp-content/uploads/2016/10/P02_NetDiligence-2016-Cyber-Claims- Study-ONLINE.pdf
7. H. Hammouchi, N. Nejjari, G. Mezzour, M. Ghogho and H. Benbrahim, "STRisk: A Socio-Technical Approach to Assess Hacking Breaches Risk," in IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2022.3149208.
8. Mandal, S., Saha, B., Nag, R. (2020). Exploiting Aspect-Classified Sentiments for Cyber- Crime Analysis and Hack Prediction. In: Kar, N., Saha, A., Deb, S. (eds) Trends in Computational Intelligence, Security and Internet of Things. ICCISIoT 2020. Communications in Computer and Information Science, vol 1358. Springer, Cham. https://doi.org/10.1007/978-3-030-66763-4_18
9. Poyraz, O.I., Canan, M., McShane, M. et al. Cyber assets at risk: monetary impact of U.S. personally identifiable information mega data breaches. Geneva Pap Risk Insur Issues Pract 45, 616–638 (2020). https://doi.org/10.1057/s41288-020-00185-4
10. Guru Akhil, T., Pranay Krishna, Y., Gangireddy, C., Kumar, A.K. (2022). Cyber Hacking Breaches for Demonstrating and Forecasting. In: Kumar, A., Mozar, S. (eds) ICCCE 2021. Lecture Notes in Electrical Engineering, vol 828. Springer, Singapore. https://doi.org/10.1007/978-981-16-7985-8_106
11. Z. Fang, M. Xu, S. Xu and T. Hu, "A Framework for Predicting Data Breach Risk: Leveraging Dependence to Cope With Sparsity," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2186-2201, 2021, doi: 10.1109/TIFS.2021.3051804.
12. Kure, H.I., Islam, S., Ghazanfar, M. et

al. Asset criticality and risk prediction for an effective cybersecurity risk management of cyber-physical system. Neural Comput & Applic 34, 493–514 (2022). https://doi.org/10.1007/s00521-021-06400-0

13. R. R. Subramanian, R. Avula, P. S. Surya and B. Pranay, "Modeling and Predicting Cyber Hacking Breaches," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 288-293, doi: 10.1109/ICICCS51141.2021.9432175.