WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Cyber Hacking and Breaches Prediction and Detection Using Machine Learning

Sufiyan Ansari[1], Manish Kumar Rathore[2], Prof. Priya Mourya[3]

[1,2]Student, [3]Guide

Department of Computer Science & Engineering Shri Shankaracharya Technical Campus (CSVTU), Junwani, Bhilai, Chhattisgarh, India

[1]E-mail: sufiyan1440@gmail.com, Manishrathore2505@gmail.com

ABSTRACT

Cyber hacking breaches prediction is one of the emerging technologies and it has been a quite challenging task to recognize breaches detection and prediction using computer algorithms. Making malware detection more responsive, scalable, and efficient than traditional systems that call for human involvement is the main goal of applying machine learning for breaches detection and prediction. Various types of cyber hacking attacks any of them will harm a person's information and financial reputation. Data from governmental and non-profit organizations, such as user and company information, may be compromised, posing a risk to their finances and reputation. The information can be collected from websites that can trigger cyberattack. Organizations like the healthcare industry are able to contain sensitive data that needs to be kept discreet and safe. Identity theft, fraud, and other losses may be caused by data breaches. The findings indicate that 70% of breaches affect numerous organizations, including the healthcare industry. The analysis displays the likelihood of a data breach. Due to increased usage of computer applications, the security for host and network is leading to the risk of data breaches. Machine learning methods can be used to find these assaults. By research, machine learning models are utilized to protect the website from security flaws. The dataset can be obtained from the Privacy Rights Clearinghouse. Data breaches can be decreased by educating staff on the use of modern security measures. This can aid in understanding the attacks knowledge and data security. The machine learning models like Random Forest, Decision Tree, k-means and Multi- layer Perceptron are used to predict the data breaches.

Keywords—Cyber hacking breaches, Machine learning, Algorithms, Prediction.

1. Introduction

Over the years, corporations have increasingly become the target of cyberattacks. Among the ransomware that causes such severe damage, the hackers may employ many attack types. Today, maintaining system security, including the confidentiality of both corporate and personal data, is quite difficult. Millions of cyberattacks per day cause tremendous financial losses. Our study has three main objectives. Using actual cybercrime data, the initial phase is to predict a cybercrime strategy, and the accuracy results are then compared. The second is to examine if the information at hand can be used to predict cybercrime perpetrators. It is employed to hide a system's data. Information theft is caused by sensitive and highly confidential data as well as poor management. The hackers' techniques might be found in two different methods. One is to move through with legal action, get in touch with the victim, and let them know about the violations [1]. The organizations should be aware of the sorts, trends, and patterns of assaults for the purpose of enabling them to monitor the system. We present a study on the consequences of these kinds of attacks in an effort for managing the prevention of occurring the beaches. We provide comprehensive study of the breaches that have occurred by the various organizations and financial effect. Because of improvements in information technology, declining prices for memory and storage devices, and the expansion of the digital economy, businesses and governmental organizations now acquire more data every day. Businesses and organizations have the threat of data attacks because of the collecting of personal data on their computers. The Privacy Rights Clearinghouse (PRC) found a large number of records that were exposed due to data breaches between 2005 and 2019. The organization may face legal action as a result of data breaches that cause financial and reputational harm. Computer networks are used in manufacturing, healthcare, research. This information is transferring every second through network. These attacks are used for profit and destroy the important information and use that for own need which rises the risk of data [2]. Hybrid based detection is used to detect the high false positive rates and low false positive rates. Anomaly based detection analyzes the behavior of traffic, where signature- based detection has the previous attacks records and able to detect the possibilities. over the Internet. One of the most prominent services offered by cloud computing is cloud storage. Cloud storage is a term that refers to an online space that can be used to store data. In more strict way, cloud storage is a service model in which data is maintained, managed and backed up remotely and made available to users over a network.

There are supervised and unsupervised methods of machine learning techniques. In a supervised method, the model will be trained by the labelled input [4]. By that label the model will distinguish between many classes existing from the records of taken dataset, where as in another model, system will be trained by the unlabeled input in order to determine the content of exactly same existed in the dataset of input Using Twitter data that has been scraped

and categorized, online social networks can serve as platforms and routes for exchanging information [5]. Probabilistic methodology evaluates the relationship between user group sentiment and potential cyberattacks We use a statistical modelling approach to address these problems and apply them for PRC data, which demonstrates It shows that neither the number of breaches nor their frequency have grown over time. We distinguish between two types of breaches: negligent and malicious, which happen when personal information is accidentally revealed. The dataset contains, both the negligent and risk sizes contain also persist the constant. PRC is a nonprofit organization with a focus on privacy problems [6]. Only the 4571 breaches that involved the delicate data contains corresponding data sum up in the dataset. Sizes of data breaches (records exposed) across a ten- year time frame. confined our analysis to the 2253 breaches in this subset. These statistics have two major drawbacks. The collection only includes breaches that have been publicly recognized, Hence the list of details mentioned for therefore the number of records listed for a breach each is simply the approximation of the total number of people suffered. PRC dataset, on the other hand, is the biggest and most comprehensive open dataset of its kind. There's a chance that several data breaches go unreported.

2. RELATED WORK

Between 2005 and 2011, Ayyagari examined the records of breaches and found that occurrence of hacking attacks was declining [8]. Smith focused data breaches that occur on health organizations and he also studied the relationship between breaches and storage. He discovered that 72% of attacks come from health companies and involve digital and electronic data. Shu looked into the intended company and implemented a number of measures to prevent the release of personal information. In contrast to Algarni, Malaiya explored the elements that contribute to data breaches and looked at how calculators are used to determine these factors. Horawalavithana investigated the vulnerability prediction activities.[9] Security- related algorithms used Protecting information on a public website against unauthorized scraping or illegitimate usage is essential for the today's technology environment. Data is crucial and has a big financial influence on many business executives and website owners. Security procedures such as honeypots are used to steer a masquerader in the incorrect direction [10]. Gaul and Rehman researched prediction techniques. Jenkins looked examined the correlation between data breach features using the speech act theory. Chen examined how phishing attempts and other organizational characteristics effects the stock value of multinational corporations [11].

McLeod and Dolezel discovered the levels of exposure, security, and how they can result in data breaches after coming across data breaches at health organizations. Kafali investigated the link between regulations and data leaks. Sen and Borle tested a number of data breach theories and discovered that forgoing information technology protection can increase the likelihood of a data breach Xu examined hacking techniques that were employed. Sen and Borle discovered that tight legislation can lower the chance of violations [12].

A model for data security was put forth by Kantarcioglu and Shaon to safeguard the systems. Bertino and Ferrari [13,14] researched the ideas around big data security and organizational security. Data on cyberbreach incidents are broken down into technical, social, and socio-technical categories. Liu researched network anomalies and malware using Risk teller to analyze 600 thousand of devices and examined their malware infections. Many organizations had the breaches 67.8% and 55.5%, of file sharing activities [15].

Martin suggested a deep learning architecture based on randomized ensembles for identifying breaches. caminero applied NSL-KDD using data from the AWID database [16]. To demonstrate how more fragile random forest is, Hang created a gradient tree technique. Huang introduced the extraction and classification of multiscale guided features. Iman extracted the chosen features from the dataset using the bortua feature selection algorithm [17,18].

Gwebu discovered that companies with a poorer reputation suffer a lower stock market value return than companies with a positive reputation [19]. According to Khandpur's research on social media, 71% of data breaches on Twitter occur there. Shu employed social media as a sensor to decipher social behavior from cyberattacks. Ritter used historical data to identify cyberattacks like denial of service and data breaches. Sarkar researched the dark web to understand the weaknesses [20].

Zhang researched the issues and threats of data security while Abouelmehdietal performed the current challenges with large data security. Ikegami and Kikuchi examined the model design of the breach dataset of Japan. Using the PRC dataset, Peng conducted research on cyberattacks. Eling and Loperfido conducted research on the type-based analysis of data beaches.

A panel regression that evaluates the possibility of data breaches was proposed by Buckman. A fixed-effect model was developed by Romansky [22] to determine the impact of data breaches.

Edwards looked at the frequency and significance of data breaches. The researchers showed that a log-normal distribution [23] may be used to match the volume breach and compared with binomial distribution which is negative that can be used to match the frequency. Sun developed a model that is applied to compile any level of breach data in order to know the rate making during underwritten of cybersecurity insurance. We have examined a range of subject areas, such as data clustering and privacy frameworks, before using a multidisciplinary approach to show why data privacy and security are crucial. The author provided a mathematical foundation description to characteristics and analyze privacy to aid in his work on preventing breaches that result in privacy.

3. EXISTING SYSTEM

Several elements, such as unexpected application existence, network port usage, and strange network activity. Different kinds of attacks that occur on cloud environment include details hijacking, mischievous information of client manipulation, denial of service, risky VM migration, and sniffing/spoofing

of virtual networks. All of these cutting-edge techniques might be employed by hackers to attempt to seize control of the cloud service. Data on user profiles, hosts, connections, protocols, and devices are tracked by an intrusion detection system. Firewalls and system monitors check the systems and websites used by businesses that handle sensitive data frequently to help combat these issues. Present, to detect the data breaches third parties are used. However, a lot of hackers and security experts make an effort to undermine a company's security measures out of personal hatred or for other reasons.

Each instance of a security breach contains a succinct narrative, the date the breach was first discovered, and the count of breaches, records and the specifics of offence. We only maintain breaches caused by hacking operations. Veris Community Database is the second. Launched in 2013, the Veris Community Database is a Verizon Threat Task that is meant to collect and disseminate information on security incidents from multiple reliable sources. A most recent edition has almost 8000 incidents and is divided into seven threat action categories: ransomware, hackers, interpersonal, exploitation, physical ones, errors, and environmental. The major subjects of this study are malware and hacking because they have a lot in common with outside hacking activities.

A security breach is indeed a security alert where sensitive information from a service or company is fraudulently accessed. Events like the Marriott data breach, which took almost 4 years to be detected, and a company like Verizon, which took almost 6 months to identify a data breach in 2016, are examples of how big corporate behemoths routinely disregard security fundamentals. Information breaches are when confidential or secure information about an organization is intentionally or accidentally collected. Decision tree algorithm works better with eccentricity but fails over time. In regression models as well, the threshold value significantly biases the results. If the barrier is too high, the entire process will stop working. Although very sophisticated, neural Networks need a lot of information at first. The machine learning framework used for security, which is currently being built in the background, will keep an eye on the website from both an internal and external perspective. With the help of several datasets, this model was created. To maintain the system under control, the model requires a variety of actions.

4.PROPOSED SYSTEM

A. Data and Methodology

The dataset used for prediction is taken from Kaggle as its data source. The dataset contains 300 instances with organization, website and social network details and year, records, organization type and sources are the attributes in the dataset. Several methods were employed to alter the dataset's structure and impute missing values during the preprocessing stage. During the preparation stage, the dataset was analyzed, and a number of approaches were utilized to modify the structure.

9	9		Air Canada	2018	20000	transport	hacked	[19]

Table 1 Attributes and its characteristics

Attributes like Entity, year, records, organization type and method are used in the table 1. The characteristics like Airtel and years 2018,2019,2020 and healthcare, social networking are used in the dataset.

B. Data Preprocessing

In order to remove the null and duplicate values data preprocessing is performed. The adjustments made to the data before we send for the particular algorithm is known as data pre- processing. The Data Preprocessing used for transforming noise data into clear datasets. In other words, the data is collected in the raw form from different sources, which makes it impossible for the evaluation. The data must be organized properly in order to achieve better outcomes from the machine learning app roach used to apply the model. To check if there are any missing values they can be detected in preprocessing method. The System efficiency and accuracy may affect if we do not perform data preprocessing. Here the output for the null values is not displayed as they are not consisted with numerical values. The machine learning models have specific requirements for the information's format; for instance, the K means method does not tolerate null values. Therefore, null values from the initial raw data collection must be controlled in order to run the k means method.

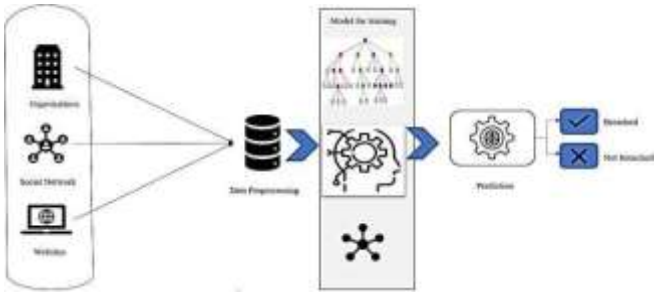

Figure 1 Architecture

Data is collected from various organizations, social network and websites. Then the data is aggregated and preprocessed. That data is sent to the training models to train the data and algorithms are applied on them. Prediction of data breaches whether occurred or not is identified. By using machine learning models it able to detects the breaches of data.

C. Feature Selection

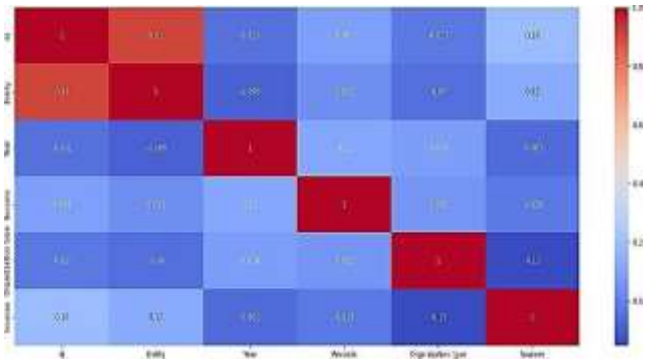

Figure 2 Feature Selection

D. Implementation

We are using Decision Tree, Random Forest, K-Means and Multi-layer perceptron algorithms for predicting the cyber breaches.

Decision Tree Model:

The decision tree algorithm is a supervised machine learning algorithm. It is used for both classification and regression problems. It is used to make predictions by taking the answers of questions previously noted. Two types of decision trees are existed, Categorial decision tree which predicts in discrete form of the data belongs and another type is regression tree is Finding the best feature from the features that are present in training data is the process of feature selection. Correlation coefficient method is used for the feature selection. It is the technique of selecting only suitable data and removing the unnecessary data. The main aim of the feature model in machine learning is to build useful models. It reduces the number of input variables when executing the model. By using correlation method, we can predict the variable from the other. It mainly used because the optimal variables are correlated. If the two variables are correlated, we take one of them which is more adequate with the destinated variable. subgroup in a node has the entire target variable, at which point the iteration is finished.

$(l, k) = (l_1, l_2, l_3, l_{m,k})$

The target variable, l, is what we're attempting to comprehend, categorise, or generalise. The input variables l_1, l_2, l_3 are

of decision trees by taking different samples and takes the majority vote for regression and classification.

Random Forest Model:

Random forest algorithm is a supervised machine learning algorithm. It is the collection of decision trees. The next decision tree will be error free and efficient than the previous one. In this way the decision trees are formed which provides the efficiency. It will provide the understandable predictions. It can be used for regression and classification problems. It builds various number The decision tree algorithm is a supervised machine learning algorithm. It is used for both classification and regression problems. It is used to make predictions by taking the answers of questions previously noted. Two types of decision trees are existed, Categorial decision tree which predicts in discrete form of the data belongs and another type is regression tree is regression tree in which the predictions in it can be considered as actual number. The different types of terminologies like Gini index, Information gain, Chi-Square are used by decision tree to work with variables of nodes and sub nodes. This algorithm automatically learns the breaches signatures and divide the task as breached or not. The collection of sources is divided into sub parts based on the value of attribute. The process of repeating this procedure on every outcome subset is referred as recursive partitioning.

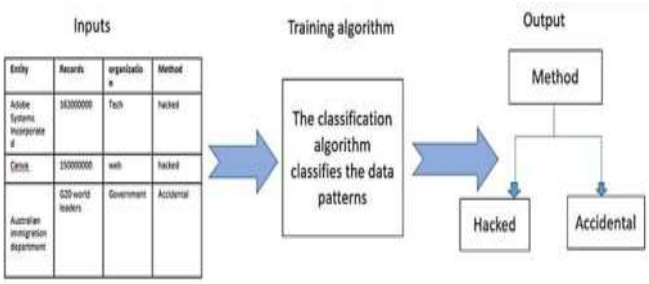

Figure 3 Working of Decision Tree Algorithm

The division no longer creates value to the predictions, or the The tree with the number n was chosen at random from a pattern.

Each category tree inside the ensemble is built using a distinct subset of the training data, B (l, m), if B (l, m) is the representation of the dataset. Then, each tree functions as usual decision trees. Data is divided into segments according to a randomly chosen value until it is completely partitioned or the maximum depth is achieved.

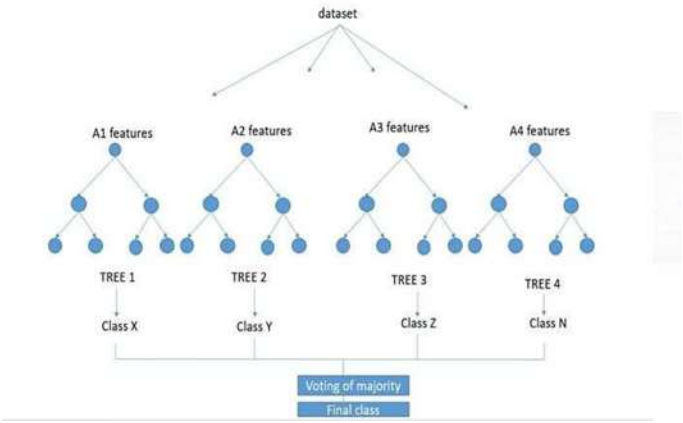

Figure 4 Working of random forest classifier

K means model:

K- Means Clustering is an unsupervised learning algorithm, which is used to group the unlabeled dataset into different clusters. It enables us to divide the dataset into various clusters so that we may identify the different groupings within the unlabeled dataset by itself without training. It is an algorithm based on centroid. Each centroid will be associated by cluster. The main goal of this algorithm to reduce the sum of the distance between data point and its clusters.

The unlabeled input dataset is divided into a variety of k- number of clusters, and this procedure will be repeated before the best cluster groups are discovered which will be the result. It works by first choosing the value of k in order to determine how many clusters will be produced. Then, it chooses the randomly chosen points that will serve as centroids. Next the data points are placed based on the distance of centroid they are placed whether nearest to the centroid or closest to the centroid. A new set of random points(centroid) is placed for each cluster. It repeats this process until it finds the optimal clusters.

The objective of this approach is to minimize an objective function, in this case, the squared error function.

$$= \sum -1 \sum -1 \| \quad ( ) - \|2 \quad (2)$$

Where $\| ^{( )} - \|$ is a selected measure of the distance between a data point and the cluster centre is an indication of how far the n data points are from each cluster's center. Centroid is the unknown locality of the clusters center. From the above formula x, y are the variables in squared error function

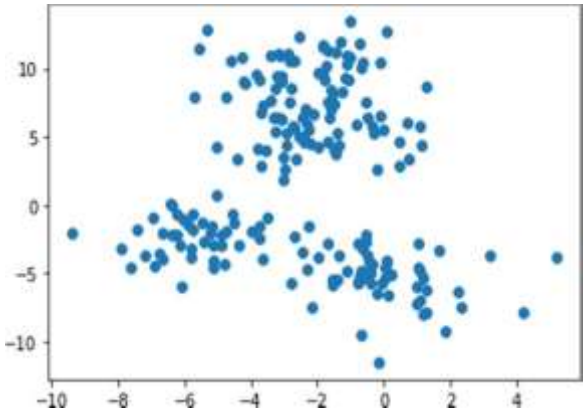

Figure 5 Before applying k-means Clustering

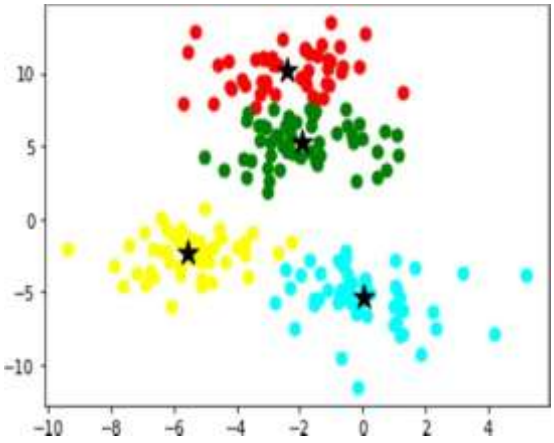

Figure 6 After applying k-means clustering

In Figure 5 the training examples are shown as dots. Here all different categories are mixed up before applying the k-means clustering. In the figure 6 the centroids are in star shape. Here after applying k-means clustering the different categories are divided into different clusters. The clusters are grouped in categories according to the similar properties.

Multi-layer perceptron model:

Multi-layer perceptron (MLP) contains multiple dense layers which converts any input dimension to desired dimension. It is a neural network which combines the neurons together in which the output of some neurons is the input of other neurons. This model has three layers, Input layer, hidden layer and output layer. Input layer takes the input and forward it for the further process and the hidden layer do the same process and forward it to the output layer. The output layer gives the predictive model result.

Input Layer:

The input layer takes the dataset as input known as visible layer also because it can be shown in the network. According to the unit neuron the neural network is drawn. The data is accepted by this layer and passed to the rest of network.

Hidden Layer:

Hidden layers come after input layer. These are called as hidden layers because they are not directly connected to the input. It adds the weights to the input and transfer them to output by using activation function and the layers exposed to output layer which gives the output.

Output Layer:

This layer gives the desired output. This layer has its own biases and weights which make it predict the desired output. From the hidden layer it directly gives the prediction.

Signals travel chronologically across various layers of the MLP network's main component, from the input layer to the output layer. Multiplications are added up in each hidden layer node and then transferred through a transfer function, like a nonlinear sigmoid function. Various transfer functions are used in neural networks, in this model, logistic sigmoid function is used which is given by

$$( ) = \frac{1}{1+(-)} \quad (3)$$

Sigmoid function is used because its probability is between 0 and 1. It is used to predict the outcome of model in the form of probability value. Here is the variable and f(x) are the sigmoid function.

70% of the data set is used for training during the network evaluation, and the weights and deviations can be changed in accordance with the network and the desired output value. 15% is utilized for verification in order to prevent overfitting before the network stops training and 15% is tested to determine the network's performance.

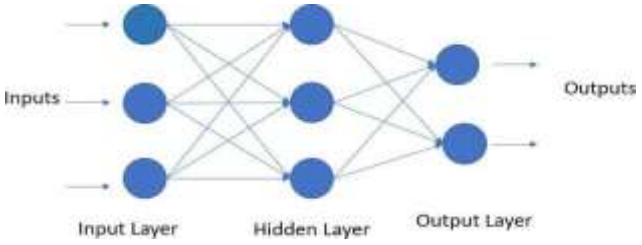

Figure 7 Multilayer perceptron

4.RESULTS

We test various predictive models using the prediction methods covered. The two classifiers that perform the best are K-Means and MLP.

	Algorithms	**Accuracy**
1	Decision Tree	90.86
2	Random Forest	91.80
3	K means	94.19
4	MLP	98.72

Table 2 Comparison of Results

Among Decision Tree, Random Forest, k-means, MLP classifiers, and we present the findings in table 2. K Means and MLP gives the better accuracy. we present the findings which is comparison of results shown in the table 2. Decision tree algorithm gives the 90.86 accuracy and Random Forest gives the 91.80 whereas K means gives 94.19 which is better and MLP gives the better accuracy compared to other machine learning models.

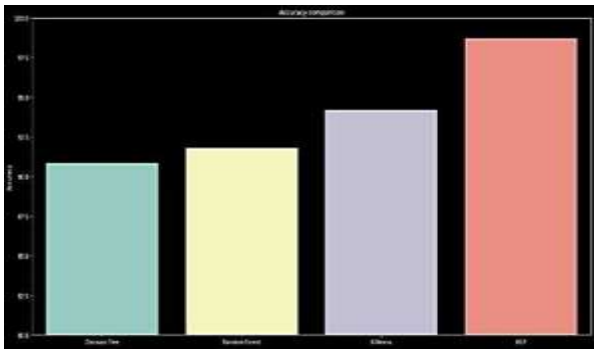

Figure 8 Comparison of results

5.CONCLUSION

A method for assessing the risk of hacker intrusions, addressing the issue of unreported intrusions, and estimating the exposure of enterprises. Since machine and deep learning techniques are increasingly being employed for a variety of purposes, including cyber security, it is imperative to determine whether and which category of algorithms can deliver adequate results. Spam detection, malware analysis, and intrusion detection are three important aspects of cyber security that are explored for these methodologies. According to our research, there are still a number of problems with current machine learning algorithms that reduce their value for cyber security. The dataset containing 300 instances is trained by using machine learning algorithms like Random Forest model, decision tree model, MLP, K-Means model. Through this method has achieved MLP and K-Means higher accuracy and yielded better output. The proposed system can be efficiently applied to detect the breaches and predict them.

***References**

1. M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu, “Modeling and predicting cyber hacking breaches,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 11, pp. 2856–2871, 2018.
2. IBM. (2019). Cost of a data breach report. IBM Security, 76. [Online]. Availablehttps://www.ibm.com/downloads/cas/ZBZLY7KL
3. Fernandez Maimo et al., “A self-adaptive deep learning-based system for anomaly detection in 5G networks,” IEEE Access, vol. 6, pp. 7700–7712, 2018.
4. Kantarcioglu M and Ferrari E (2019) Research Challenges at the Intersection of Big Data, Security and Privacy.
5. Verizon, “Data breach investigations report,” 2019. [Online]. Available:https://enterprise.verizon.com/resources/reports/dbir/
6. H. Hammouchi, O. Cherqi, G. Mezzour, M. Ghogho, and M. El Koutbi, “Digging deeper into data breaches: An exploratory data analysis of hacking breaches over time,” Procedia Computer Science, vol. 151, pp. 1004– 1009, 2019.

7. rack T. Majority of malware analysts aware of data breaches not disclosed by their employers. http://www.threattracksecurity.com/press-release/majority-of-malware-analysts-aware-of-data- breaches-not-dis closed-by-their-employers.aspx

8. K. Pujitha , Kattamanchi Prem Krishna , K. Amala , Annavarapu Yasaswini , Sivakumar Depuru , Parking Spaces", Journal of Pharmaceutical Negative Results, vol. 13, no. 4, pp. 1010–1013, Nov. 2022.

9. Sivakumar Depuru , Anjana Nandam , P.A. Ramesh , M. Saktivel , K. Amala , Sivanantham. (2022). Human Emotion Recognition System Using Deep Learning Technique. Journal of Pharmaceutical Negative Results, 13(4), 1031–1035. https://doi.org/10.47750/pnr.2022.13.04.141 (Original work published November 4, 2022)

10. S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish and P. K. Raju, "A Machine Learning based Malware Classification Framework," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1138-1143, doi: 10.1109/ICSSIT55814.2023.10060914

11. S. Depuru, K. Vaishnavi, B. Manogna, K. J. Sri, A. Preethi and C. Priyanka, "Hybrid CNNLBP using Facial Emotion Recognition based on Deep Learning Approach," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 972-980, doi: 10.1109/ICAIS56108.2023.10073918.

12. Ayyagari, R. (2012). An exploratory analysis of data breaches from 2005-2011: Trends and insights. Journal of Information Privacy and Security

13. Algarni, A. M., Malaiya, Y. K. (2016, May). A consolidated approach for estimation of data security breach costs. In 2016 2nd International Conference on Information Management (ICIM) (pp. 26-39). IEEE.

14. Kafali, Jones, J., Petruso, M., Williams, L., Singh, M. P. (2017, May). How good is a security policy against real breaches? A HIPAA case study. In 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE) (pp. 530-540). IEEE.

15. Sen, R., Borle, S. (2015). Estimating the contextual risk of data breach: An empirical approach. Journal of Management Information Systems, 32(2), 314-341.

16. Bertino, E., & Ferrari, E. (2018). Big data security and privacy,". In A comprehensive guide through the Italian database research over the last 25 years (pp. 425–439). Springer. Gray, J., Gerlitz, C., & Bounegru, L. (2018).

17. Smith, T. T. (2016). Examining Data Privacy Breaches in Healthcare

18. A. Bachar, N. E. Makhfi, O.E. Bannay, "Towards a behavioral network intrusion detection system based on the SVM model", in 2020 1st international conference on innovation research in applied science engineering and technology (IRASET), Meknes, Morocco, 2020, pp. 1-7 Kopparam Runvika, "Development of Secured Online

19. A. Bachar, N. E. Makhfi, O.E. Bannay, "Towards a behavioral network intrusion detection system based on the SVM model", in 2020 1st international conference on innovation research in applied science engineering and technology (IRASET), Meknes, Morocco, 2020, pp. 1-7

20. L. Bilge, Y. Han, and M. Dell'Amico, "Riskteller: Predicting the risk of cyber incidents", in Proc. of the 2017 ACM SIGSAC conf. on Computer and Communications Security, 2017, pp. 1299–1311.

21. S. Sarkar, M. Almukaynizi, J. Shakarian, and P. Shakarian, "Predicting enterprise cyber incidents using social network analysis on dark web hacker forums", The Cyber Defense Review, pp. 87–102, 2019

22. M. Lopez-Martin, B. Carro, J. I.Arribas, and A. Sanchez-Esguevillas, "Network intrusion detection with a novel hierarchy of distances between embeddings of hash IP addresses",Knowledge-based Syst.,vol 219,2021.