Metadata of the chapter that will be visualized in SpringerLink

Book Title	Computational Intelligence for Business Analytics
Series Title	
Chapter Title	A Proposal for Data Breach Detection in Organizations Based on User Behavior
Copyright Year	2021
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Switzerland AG

Author	Family Name	**Palacios**
	Particle	
	Given Name	**René**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Universidad Autónoma de Ciudad Juárez
	Address	32310, Ciudad Juárez, Chihuahua, México
	Email	
Corresponding Author	Family Name	**Morales-Rocha**
	Particle	
	Given Name	**Victor**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Universidad Autónoma de Ciudad Juárez
	Address	32310, Ciudad Juárez, Chihuahua, México
	Email	victor.morales@uacj.mx

Abstract	Data breach has become a big problem for organizations, as the consequences can range from loss of reputation to financial loss. A data breach occurs through outsiders and insiders; however, threats from insiders are the most common and, at the same time, the most difficult to prevent. Data loss detection systems are increasingly implemented in organizations to protect information with techniques like content-based and context-based checking. Machine learning techniques have proven to be useful for data breach detection. In this work, a statistical analysis of data breach incidents is presented. Also, a user behavior characterization is made, mainly based on incidents reported by various organizations. Part of this characterization is used to create a machine learning model with a long short-term memory network with an autoencoder, in order to identify anomalies in user behavior to detect data breaches from insiders.

Keywords (separated by '-')	Data breach detection - Machine learning - Information security - Information processing - Analytics

A Proposal for Data Breach Detection in Organizations Based on User Behavior

René Palacios and Victor Morales-Rocha

Abstract Data breach has become a big problem for organizations, as the consequences can range from loss of reputation to financial loss. A data breach occurs through outsiders and insiders; however, threats from insiders are the most common and, at the same time, the most difficult to prevent. Data loss detection systems are increasingly implemented in organizations to protect information with techniques like content-based and context-based checking. Machine learning techniques have proven to be useful for data breach detection. In this work, a statistical analysis of data breach incidents is presented. Also, a user behavior characterization is made, mainly based on incidents reported by various organizations. Part of this characterization is used to create a machine learning model with a long short-term memory network with an autoencoder, in order to identify anomalies in user behavior to detect data breaches from insiders.

Keywords Data breach detection · Machine learning · Information security · Information processing · Analytics

1 Introduction

The National Institute of Standards and Technologies (NIST) [1] defines information security as "The protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction to ensure confidentiality, integrity, and availability". This definition provides three information security objectives confidentiality, integrity, and availability, also known as the CIA triad.

According to the NIST standard "FIPS 199" [2], confidentiality deals with "preserving authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary information".

R. Palacios · V. Morales-Rocha (✉)
Universidad Autónoma de Ciudad Juárez, 32310 Ciudad Juárez, Chihuahua, México
e-mail: victor.morales@uacj.mx

W. Pedrycz et al. (eds.), *Computational Intelligence for Business Analytics*,
Studies in Computational Intelligence 953,
https://doi.org/10.1007/978-3-030-73819-8_17

483702_1_En_17_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2021 Pages: 18 Layout: T1-Standard

Loss of confidentiality occurs when there is a data breach, which is defined as "An incident that involves sensitive, protected, or confidential information being copied, transmitted, viewed, stolen, or used by an individual unauthorized to do so. Exposed information may include credit card numbers, personal health information, customer data, company trade secrets, or matters of national security" [3].

The number of incidents related to data breaches increases every year, directly or indirectly affecting organizations and users around the world. In a report from the Identity Theft Resource Center [4] there were 1244 data breach incidents reported in 2018, exposing a total of 446,515,334 records. The number of exposed records had an increase of 126% compared to the previous year.

The threat of data breach has become a major problem for organizations as the consequences can range from loss of reputation to financial loss. There are two types of costs when a data breach occurs, according to [5], namely, tangible and intangible costs. Intangible costs include, but are not limited to, identity theft, criminal charges against staff members, the increased risk of future attacks on the organization, as well as loss of reputation. A report in [6] shows that when a data breach occurs, 65% of those affected lose their trust in the organization as a result of the incident, and 85% will tell others about their negative experience.

On the other hand, tangible costs refer to the loss of items directly related to the budget. Depending on the nature of the breach, a variety of financial problems can arise. For example, the costs of investigating the causes or vulnerabilities that allowed the incident to occur, the costs of restoring the data if it was deleted, the legal costs of defending against a customer, the cost due to the temporary or permanent loss of availability of the data, loss due to use of the stolen data by a competitor, the costs for paying customers who have suffered some loss or who have been defamed due to disclosure, among others. According to [7], the average cost in 2019 for a data breach was $3.9 million, and since the average of records lost that year was 25,575, the cost per record was approximately $150.

As data breach threats are a source of potential loss, it is important that organizations focus on preventing the loss of sensitive and confidential data as part of a comprehensive business intelligence strategy. A data breach occurs through outsiders and insiders; however, threats from insiders are the most common and, at the same time, the most difficult to prevent.

Data loss prevention has been addressed in different ways. According to the Forrester Wave report in [8], most of the first data loss prevention solutions focused on finding sensitive data by monitoring it at the network level. In the second stage, as removable storage devices matured, data loss prevention solutions began to focus on detecting data breach directly on the devices (workstations, servers, laptops) and providing actions, for example, avoid copying sensitive information to USB devices or CD/DVD, even when the device is not connected to the network. Protection normally begins with the ability to detect potential breach through heuristics, rules, patterns, statistics, classification, and search for anomalies. Prevention occurs as a consequence of detection [9, 10].

Data loss prevention solutions must consider three key objectives, according to [9]:

- Data loss prevention must have the ability to analyze the content and context of confidential data.
- It must be possible to implement data loss prevention to provide protection of confidential data in one or different states, that is, in transit, in use, and at rest.
- They must have the ability to protect data through various corrective actions, such as notification, auditing, blocking, encryption, or quarantine.

Techniques for preventing data breach are based on either content-based checking (analyzing the content of the file or body of text) or context-based checking (analyzing the information beyond the data itself, such as the size of the file, destination, type of file, time of delivery, among others). Machine learning techniques have proven to be useful for data breach prevention and detection. In this work, we propose to analyze users' behavior using long short-term memory network with an autoencoder to prevent a data breach from insiders.

The remainder of this work is organized as follows. Section 2 describes the methodology used in this work, which includes the understanding of the problem, the characterization of the user behavior, and the process of machine learning used to detect anomalies on user behavior. Section 3 presents the conclusions of the work and suggests future directions for research.

2 Methodology

This section describes the methodological approach used in this work. First, the causes that cause data breach in organizations are analyzed. For this purpose, a dataset containing a large number of data breach records was used. Then, we describe the characteristics that we consider to be important to create a user behavior profile, which is later used to create a model that will be approached with a machine learning technique. Finally, using the dataset, the anomalies associated with user behavior are identified.

2.1 The Problem in Numbers

An analysis of data breach has been performed with the dataset in [11]. This dataset contains data breach incidents from 2004 to 2019; each incident has at least more than 30,000 lost records. Each incident is classified according to the breach cause, and a group of incidents was analyzed qualitatively to determine the root cause of the incident. Table 1 describes the fields in the dataset used for the purposes of this work.

Figure 1 shows the number of incidents and records exposed over the years. It shows that the situation has been worsening, as the number of incidents and the number of records affected increases each year.

Table 1 Fields from the dataset

Field	Description
Entity	Affected organization
Records lost	Records reported in the data breach incident
Year	Year in which the incident occurred
Story	Summary of how it happened
Sector	Affected business sector
Method	The method that caused the incident
Source name	The entity that posts the incident
1st source link	Link with the reference
2nd source link	Second link with the reference

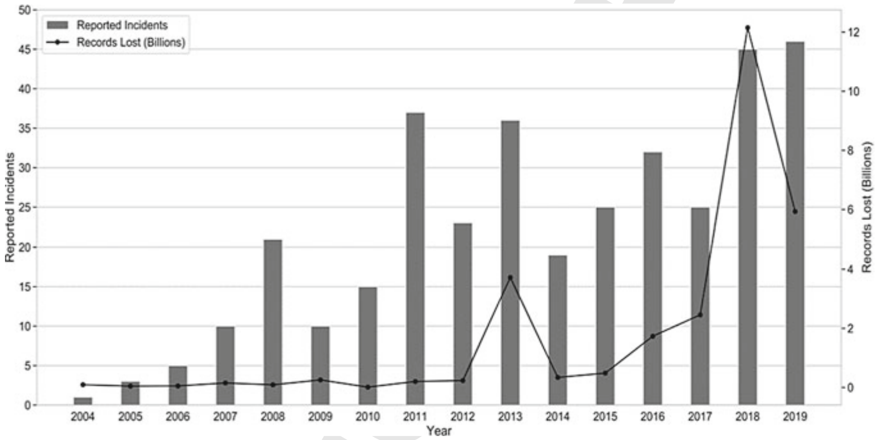

Fig. 1 Number of registered incidents and records compromised per year, from 2004 to 2019

Figure 2 lists the economic sectors most affected by a data breach in terms of incidents and compromised records. It should be clarified that the sector of large web companies, such as Facebook, Apple, Twitter, Dropbox, among others, has been ruled out in this analysis since they are usually specific targets of external intruders and represent a large part of a data breach. The focus of this work will be on organizations where a data breach is most likely due to actions of internal personnel, either accidentally or intentionally. Figure 3 shows the most affected sectors once the Web companies have been discarded.

In Fig. 4 we can see the methods used for a data breach. The hacked method accounts for 8 billions of the 16 billions of total compromised records. By obtaining the top offenders in the percentage of the total records, we can see that the top offender has been "hacked" with 53% and 8.6 billion records compromised, "poor security" with 29% and 4.7 billion records, "oops!" (accident) with 15% and 2.4

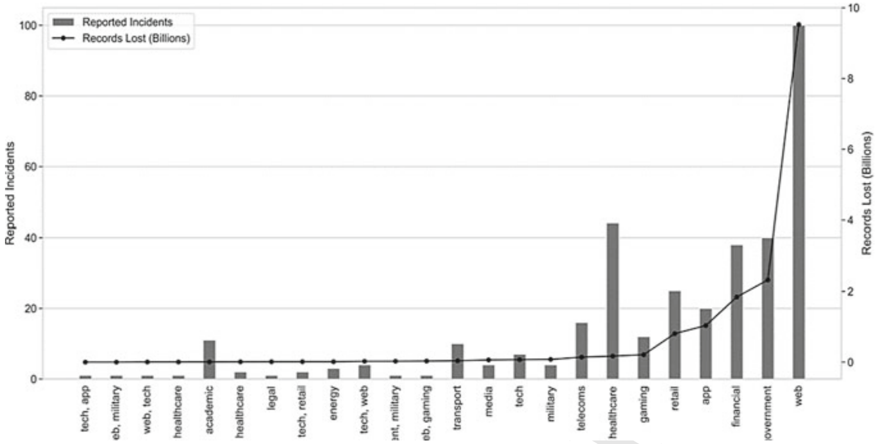

Fig. 2 Incidents and records compromised by economic sector

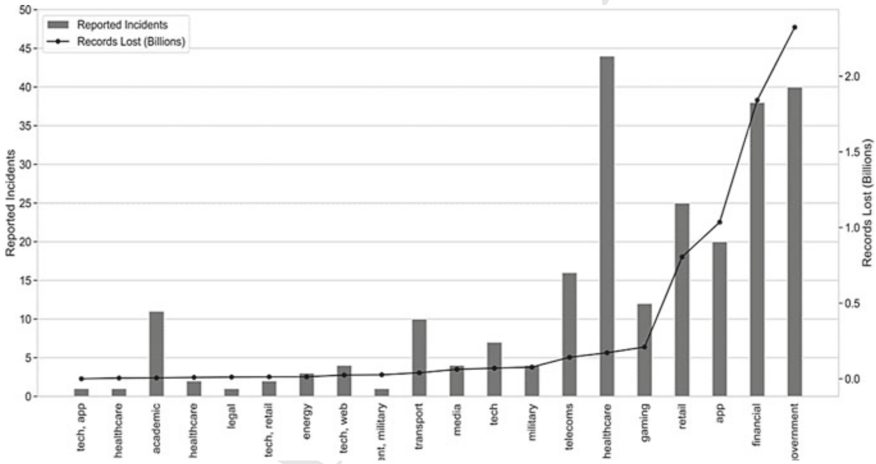

Fig. 3 Incidents and records compromised by the economic sector after removing the web sector

billion records, "inside job" with 2% and 353 million records, and lost device with 1% and 215 million records. This information can be seen in a Pareto chart in Fig. 5.

We grouped the "Oops!", "Inside job" and "lost device" categories into a single category of "insider" that represents 18% of the top offenders. Figure 6 shows the new Pareto after grouping this information.

At this point, it is clear that the "hacked" category represents 53% of data breach problems, "poor security" 29%, and 18% represents the incidents committed by an "insider". An analysis of the "hacked" category was carried out since it is assumed that some of these incidents are due to human oversights.

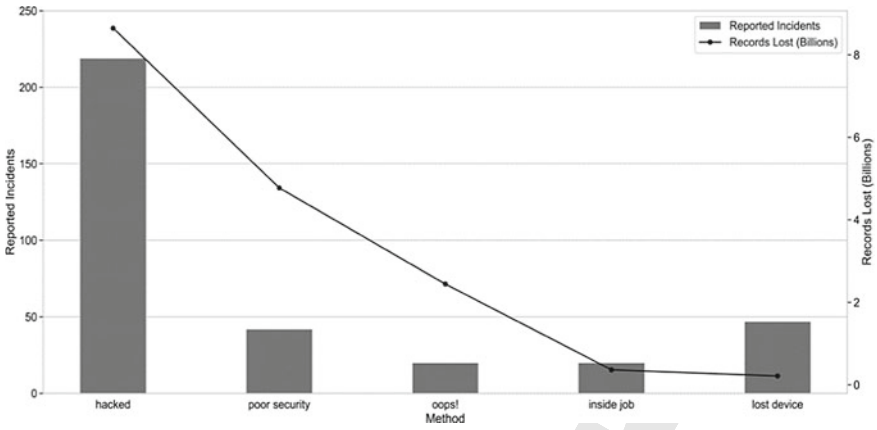

Fig. 4 Reported incidents and total records by the method used that lead to a data breach

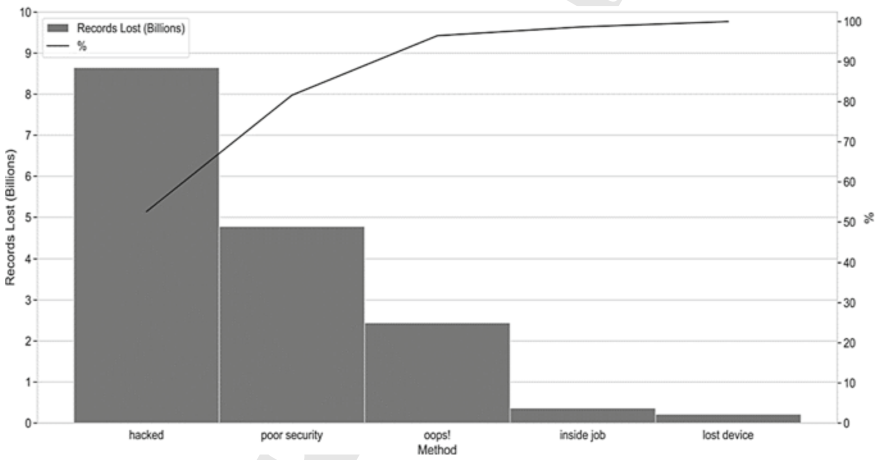

Fig. 5 Pareto chart of total records by the method used that lead to a data breach

By extracting the incidents labeled "hacked" from the previously analyzed dataset, We have a total of 133 such incidents. The calculator in [12] was used to determine a sample of 32 random incidents. These sample of incidents was empirically analyzed, and some subcategories were obtained. Moreover, the root causes that lead to a data breach incident were determined. A summary of subcategories and root causes can be seen in Table 2.

Based on the 32 randomly chosen incidents, 6 incidents were found in misuse accounts, 6 incidents related to improperly secured systems and 4 incidents in phishing attacks were carried out with techniques that did not involve a human factor directly, and 10 (misuse account and phishing attack) were a user was involved that ended up in a data breach.

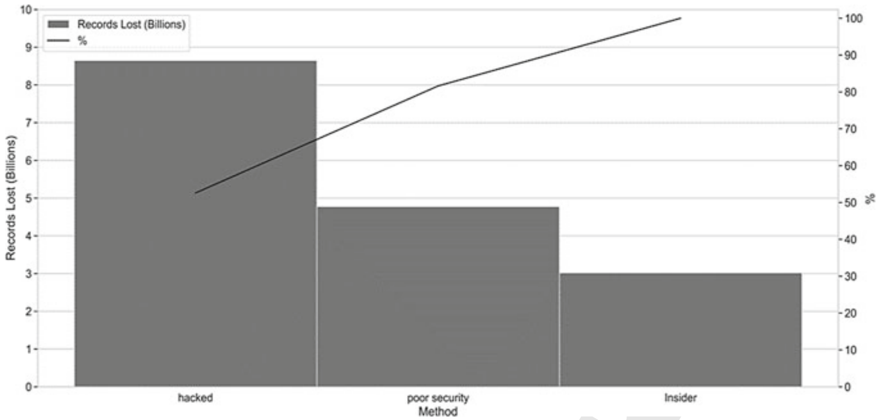

Fig. 6 Pareto chart of total records by the method used that lead to a data breach after grouping "Oops!", "Inside job" and "lost device" categories

Table 2 Description of the 32 randomly chosen incidents of method "hacked"

Subcategory	Root cause	Incidents	Total Records
Hacked	Brute force attack	2	860,083
Hacked	No details	1	270,000
Hacked	Password-guessing attack	1	57,000,000
Hacked	Vulnerability exploitation	12	49,996,000
Insider	Misuse account	6	388,150,000
Insider	Phishing attack	4	14,960,000
Poor security	Improperly secured	6	15,017,000

In Fig. 7, we can see these subcategories from the hacked category.

With the sample of 32 randomly selected incidents, a confidence level of 95% and a confidence interval of ±20, We can conclude that of the 53% that represents the "hacked" category, 77% have been caused firstly by an "insider". With this analysis, it has been concluded that most of the data breach incidents (around 77%) are caused by an insider. An insider could be a compromised user, a careless user, or a malicious user.

2.2 *User Behavior Characterization*

We propose a user behavior characterization and features selection based on a series of public articles and reports found in the dataset previously analyzed [11].

483702_1_En_17_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2021 Pages: 18 Layout: T1-Standard

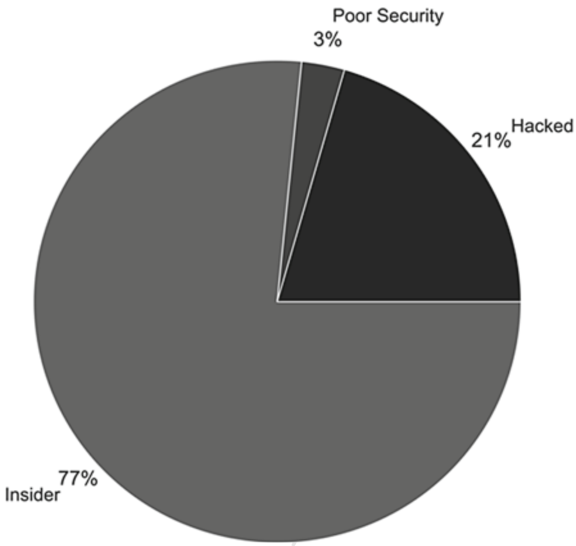

Fig. 7 Percent based pie chart of subcategories of the category hacked

The following unordered list shows examples of compromised users, careless users, and malicious users that lead a data breach in different organizations.

An example of misuse accounts or insiders can be seen in the report in [13]; In this case, around 5.2 million guest records from the Marriot hotels were accessed, apparently with the login credentials of two employees at a franchise property at the end of February 2020. "The company identified that an unexpected amount of guest information might have been accessed", these records included contact details, such as name, mailing address, email address, and phone number.

Desjardins, a financial services company, revealed in June 2019 that "an employee improperly collected information about customers and shared it with a third party outside the financial institution, which is the largest federation of credit unions in North America, with outlets across Quebec and Ontario" [14]. This is a clear example of a data breach inflicted by an insider with access to the information.

Another example of a malicious insider with access to the information occurred in June 2016 [15]. The personal details of 112,000 French police officers "have been uploaded to Google Drive in a security breach … says the details were uploaded by a disgruntled worker … Data includes home addresses."

In 2014, Korea Credit Bureau, a personal credit ratings firm revealed that "an employee has been arrested and accused of stealing the data from customers of three credit card firms while working for them as a temporary consultant" [16]. Certainly, this is another example of how an insider act.

In 2013, a lawsuit against the Vietnamese identity theft service "contends that the theft of up to 3 million records began in 2010 and was orchestrated by Hieu Minh Ngo. Ngo, posing as a private investigator based in Singapore, gained access to a database of consumer information" [17].

In another case, in 2004 [18], the organization AOL released a statement saying that “A former America Online software engineer stole 92 million screen names and e-mail addresses and sold them to spammers who sent out up to 7 billion unsolicited e-mails.”

In August 2007, a job seeker organization called Monster [19] got a trojan by a phishing email. The company said that “A trojan virus stole logins that were used to harvest usernames, e-mail addresses, home addresses, and phone numbers. Soon after, phishing e-mails encouraged users to download a Monster Job Seeker Tool, which was, in fact, a program that encrypted files in their computer and left a ransom note demanding money for their decryption.” This is a clear example of a compromised user that led to a data breach.

The Australian National University [20] was a victim of unauthorized access to information. They said, “We believe there was an unauthorized access to significant amounts of personal staff, student and visitor data extending back 19 years … by a sophisticated operator”.

Medical organizations have also suffered from data breaches. In 2014, St. Vincent Medical Group [21] reported: “The investigation has required electronic and manual review of affected emails to determine the scope of the incident. Through the ongoing investigation of this matter, we determined on March 12, 2015, that the employee email account subject to the phishing contained some personal health information for approximately 760 patients”.

Another company affected by a phishing email that leads to a data breach was JP Morgan [22], “affecting 76 million households and 7 million small businesses, have apparently originated with spear-phishing campaigns that target a small number of employees who have access to data systems and services housing sensitive customer information”.

Based on the reports cited previously, we have identified the potential characteristics that help us to identify possible anomalies in user behavior, for example, login time, active session time, amount of data transfer, accessed directories, among others. It is clear that in many of these scenarios, the users of the organization itself are involved, either through deception, for example, when they are victims of phishing, or by carelessness, for example, users who do not comply with the security policies of their organization. Another possible scenario is when a malicious user, with legitimate access to the organization’s resources, intentionally extracts data.

Table 3 contains the features used to characterize users behavior.

2.3 Scope Definition

The dataset CSE-CIC-IDS2018 [23] was used to extract all the user behavior previously defined in Sect. 2.2 with the features available in the evtx and pcap files; one of the principal characteristics of this dataset is that it has user profiles that contain abstract representations of events and behaviors seen on a network. This dataset

483702_1_En_17_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2021 Pages: 18 Layout: T1-Standard

Table 3 Selected features of users behavior and their description

Feature	Description
Login time	Time in which users gain access to a computer system by identifying and authenticating themselves
Active session time	Time in seconds a user spends with an active valid session
Amount of usual data transfer	Amount of data a user transfer through the network
Data transfer protocol used by a user	Protocols utilized by the user (i.e., HTPPS, FTP, SSH)
Software used	List of software commonly used
Software recurrency	Recurrency of the used software
Software data amount transfer	Amount of data transferred or downloaded by the software
Web pages used	List of commonly visited web pages used by the user
Web pages data transfer	Amount of data transferred through the website
Web pages recurrency	Recurrency of the web pages visited
Accessed directories	List of commonly network directories accessed
Accessed directories data amount transfer	Amount in GB's transferred or downloaded from the directories to a local media
Accessed directories recurrency	Recurrency of access to directories
External media	List of external media connected
External media data amount transfer	Amount of data transferred or downloaded from external media
External media recurrency	Recurrency of connected media

includes an attacking infrastructure with 50 machines and a victim organization with 5 departments that includes 420 machines and 30 servers.

This dataset has *pcap* files containing packets information of the network and *evtx* files containing the list of events logged by Windows from user profiles.

All the events of the machines are saved individually in the *evtx* files in a proprietary binary format that can only be viewed within the Event Viewer program of Windows.

It is necessary to extract all the features available in the evtx files into a plain text file, specifically into a comma-separated values file, in order to process the data and train a machine learning model. To do that, a script with the capacity to extract all the features from these files and save them in comma-separated values format was created. By doing this, we can extract all the features and log information of all the machines and extract features like: date and time of the event created, time a user logged in, time that user kept an active session, programs used, time lasted with an opened program, among others. On the other hand, we extracted all data streams generated by computers on the network from the *pcap* files.

Anomaly detection is a task of finding rare events [24]. Supervised and unsupervised approaches to anomaly detection have been proposed. Some of these approaches include techniques like Bayesian networks, cluster analysis, support vector machines, and neural networks.

In this work, we used a long short-term memory autoencoder neural network to detect anomalies on user behavior. Long short-term memory networks are a type of recurrent neural network capable of learning order dependence to address sequence prediction problems. Firstly, introduced by Hochreiter and Schmidhuber [25] in 1997, and a non-comprehensive contribution of works by Gers [26], Graves and Schmidhuber [27], Wang and Nyberg [28].

There are other techniques to approach time series data such as Markov chains, multilayer perceptron, convolutional neural networks, among others; however we selected long short-term memory autoencoder neural network as in our experience, it is the easiest way to address our particular problem.

In machine learning problems, it is common to have sets of data; these sets are used to train a model and can be seen as an observation of the problem domain. The order of the observations given to the model is not important [29]. On the other hand, when we have a sequence, the order of the observations given to the model is important [30]. Sequence prediction involves predicting the next value given a sequence; for example, given an input sequence of numbers from 1 to 8 to a sequence prediction model, the expected output is 9.

An autoencoder is a type of artificial neural network used to learn features in an unsupervised way. An autoencoder attempts to learn features by training the network to ignore the noise and to force the model to learn representations of the input to assume useful properties.

In order to detect anomalies in user behavior, the autoencoder was prepared as follows:

- The autoencoder is trained on normal sequential data.
- It will be tested taking a new sequence and trying to reconstruct it using the autoencoder.
- If the error for the new sequence is superior to the defined threshold, the given element is labeled as an anomaly.

All the experiments have been done in *Jupyter Notebook,* and the programming language used is *Python*. The python library *Pandas* was used for data manipulation; the *Python* library *TensorFlow* was used to develop and train the machine learning model; the *Python* library *scikit-learn*, that provides useful algorithms for machine learning was also used; finally, the *Python* library *Matplotlib* has been used for all the visualizations presented in the following sections.

2.4 Data Preparation

Based on the information extracted, two features were selected, date and time lasted on the active session.

483702_1_En_17_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2021 Pages: 18 Layout: T1-Standard

The feature "date" is used to express the year, month, and day a user was active, and "actTime" is used to express the active session time in seconds.

Having a look at the selected dataset in Fig. 8, we can see in a linear chart the active time feature for two months.

Before training the model, we need to standardize the dataset. Standardization of a dataset is a common requirement for many machine learning estimators as they might behave poorly or slow down the learning of the model if the individual features do not look like the standard normally distributed data. We were able to accommodate the data with the *scikit-learn* function *StandardScaler*; after that, we had a dataset that looks like Fig. 9.

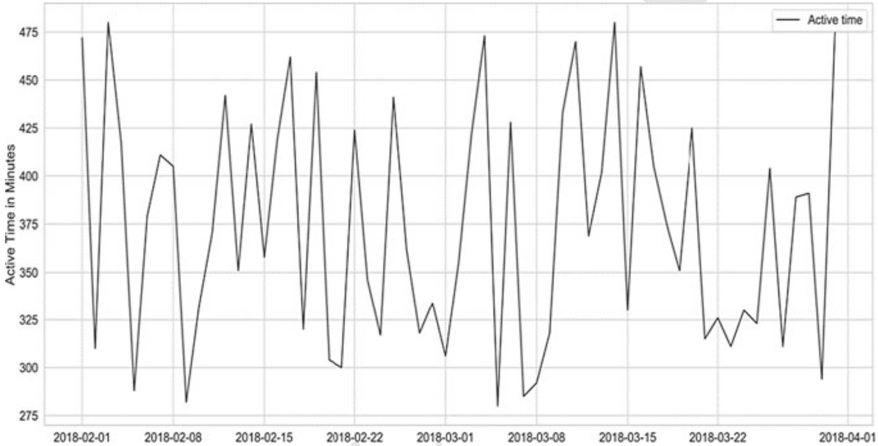

Fig. 8 Lineal representation of the active session time feature

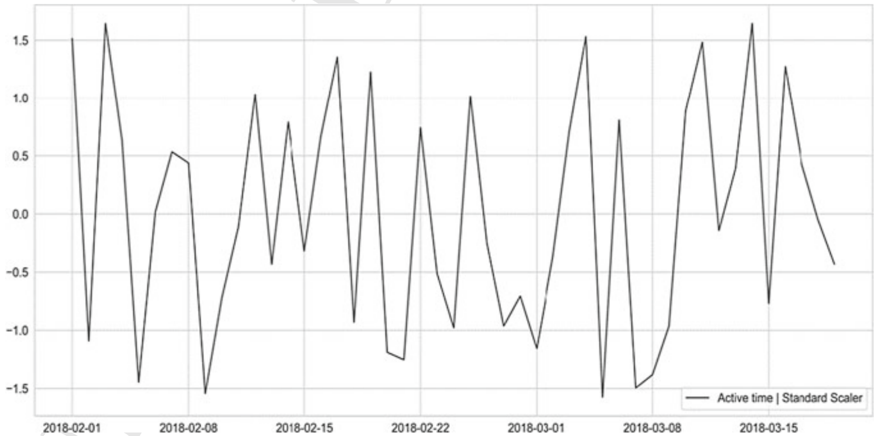

Fig. 9 Lineal representation of the active session in two months period after data rescaling

Table 4 Arguments and values used in the model configuration

Arguments	Value
Dropout rate	0.5
Compile loss	Mean absolute error
Compile optimizer	Adam algorithm

Table 5 Model layer architecture and parameters

Layer (type)	Output shape	Param #
Lstm (LSTM)	(None, 64)	16,896
Dropout (Dropout)	(None, 64)	0
Repeat_vector (RepeatVector)	(None, 2, 64)	0
lstm_1 (LSTM)	(None, 2, 64)	33,024
Dropout_1 (Dropout)	(None, 2, 64)	0
Time_distributed (TimeDistri	(None, 2, 1)	65
Total params: 49,985		
Trainable params: 49,985		
Non-trainable params: 0		

2.5 Model Configuration

The first step is to define a neural network in *Keras*; this network is defined as a sequence of layers contained in a Sequential class. To define a model, an instance of Sequential class is created. Layers are added to this class, and in the end, each layer can be connected. Table 4 presents the arguments and the selected values used for this model. The model was defined as follows:

- Dropout rate. Temporarily remove units from the network to prevent overfitting.
- Compile Loss. Used to judge the performance of the model minimized by the optimization algorithm.
- Compile Optimizer. Optimization algorithm to train the network.

After defining the loss function, the optimizer, and the metrics, the function *Compile* of *Keras* is used to be able to train our model. Table 5 shows the description of the layers with the values of the model.

2.6 Model Training

Once the model is successfully compiled without errors, it needs to be fitted or adapted according to the weights on the training dataset. To accomplish this, the training data needs to be specified with the input and output patterns (X, y). The model is trained using backpropagation through time algorithm, already defined in *Keras*, optimized

Table 6 Arguments and values used for model training

Arguments	Values
Epoch	10
Batch	32

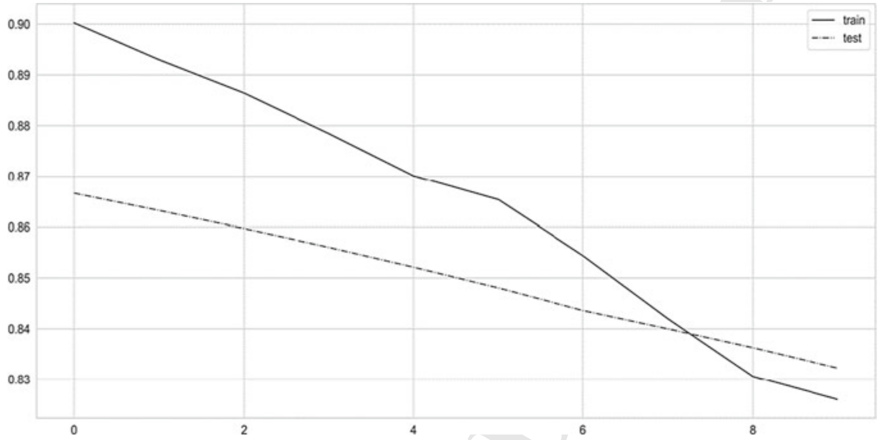

Fig. 10 Performance obtained with 10 epochs

with the Adam algorithm, and for the loss function, the mean absolute error (MAE) was defined in the model configuration. Table 6 presents the arguments used with the selected values. The model was trained with the following parameters:

- Epoch. "Used to separate training into distinct phases, which is useful for logging and periodic evaluation" [31].
- Batch. "Approximates the distribution of the input data better than a single input" [31].

Once fit, an object is returned with the information of the performance during training. We can see the performance returned in Fig. 10.

In Fig. 11, we present the MAE calculated to see the average magnitude of errors in the predictions set on the training data.

A threshold of 0.70 is defined since the loss is not greater than that. If there is an error greater than the established threshold, that element is declared as anomalous behavior. In Fig. 12, we can see the loss and all of the elements above the threshold.

2.7 *Model Predictions*

Once the model is fit, we can make predictions with the model, simply by calling the *Keras* function that performs a prediction with an array of new input patterns. In

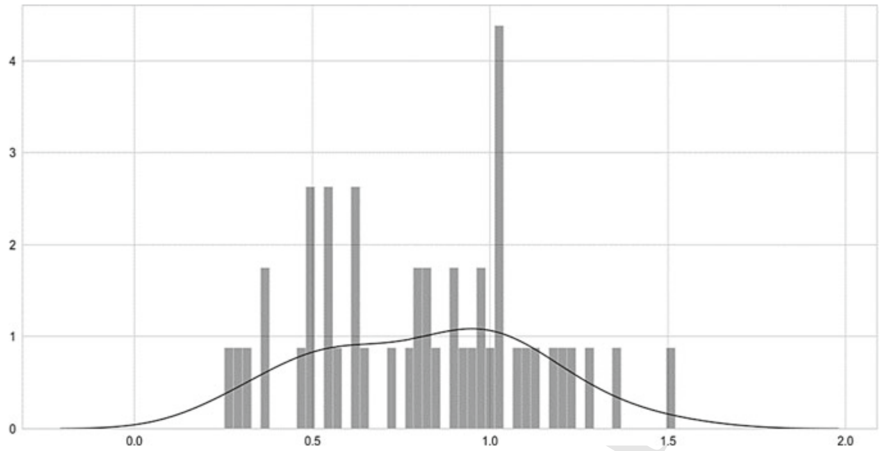

Fig. 11 Mean absolute error of the prediction set

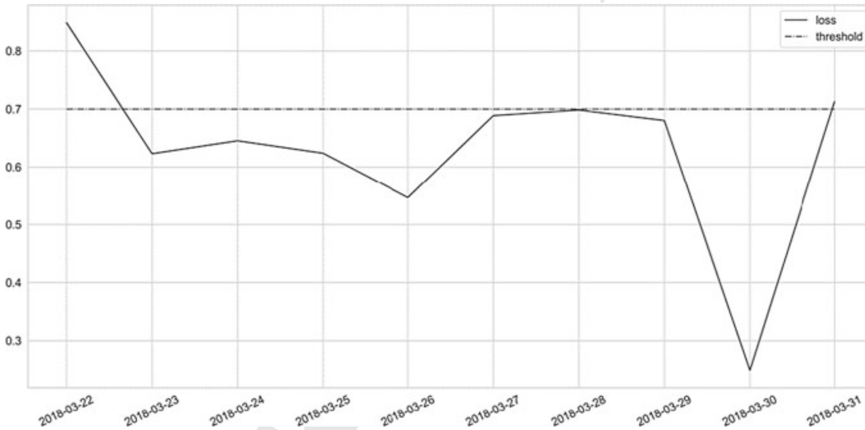

Fig. 12 Threshold and loss from the training dataset

Fig. 13, we can see the anomalies found in the testing data. The dots show the points where there is an abrupt change.

Using two features of the user behavior characterization proposed, we described our data breach anomaly detection. The combination of autoencoders and long short-term memory resulted in a model able to find anomalies on user behavior. The model shows an accuracy of 0.8169, which is considered satisfactory, especially if we take into account that our model was trained without showing a single anomaly.

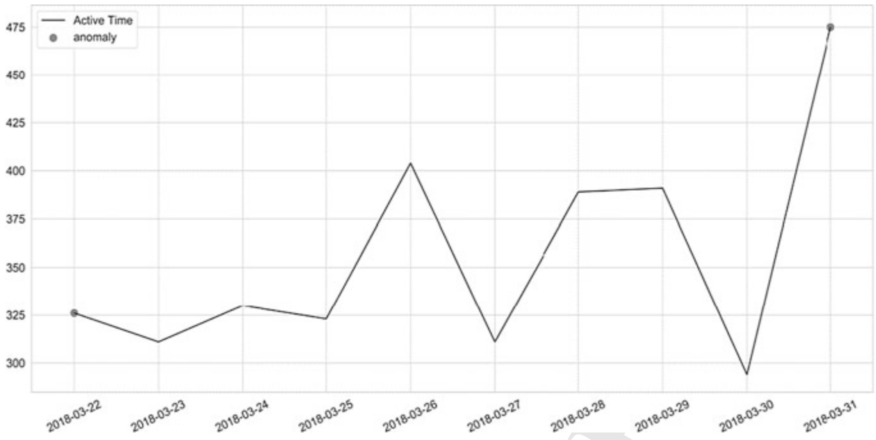

Fig. 13 Test dataset with detected anomalies (dots)

3 Conclusions and Research Directions

There is no doubt that data breach is an ongoing and relevant problem in the information security field as it affects the reputation and finances of organizations. For this reason, organizations must implement systems or mechanisms that allow them to detect and monitor data leakage attempts as part of their business intelligence strategy.

Having carried out an analysis to determine the causes of data breaches in organizations, it is concluded that computer users (insiders) are one of the main causes that lead to a data breach either compromised, careless or malicious. In this sense, characterization of user behavior has been proposed. The proposed user behavior characterization has 16 features, which can be considered as the general characteristics for the majority of users of computer equipment. However, this characterization can be adapted, either reducing or expanding the characteristics according to the needs of each organization.

In this work, a machine learning model and the combination of autoencoders and long short-term memory have been tested. This work has proven that this combination is suitable to detect anomalies in user behavior, based on the characterization proposed. Even though the model was not able to detect all the anomalies, its accuracy was around 80%. However, its accuracy could be improved, either improving the model architecture or diversifying the training data with more parameters and features.

This work can be extended in several ways. For instance, we only used two features from all the proposed characterization. In order to be able to identify a potential data breach, it should be necessary to extend this work using all the features of the characterization. As shown in this work, we can try to tune the model and work with the threshold to get better results.

Another future line of research could be the implementation of other machine learning techniques to the proposed characterization by using a single model or a combination of machine learning models to detect anomalies in user behavior.

Further, additional features can be analyzed to be added to the proposed characterization to understand all the behavior of a computer user that can lead to a data breach by accident or intentionally.

Finally, a combination of different data breach techniques like data content analysis and data context analysis, along with organizational policies such as external devices and external network communications restrictions, as well as procedural measures like user training to identify threats in the form of malicious links or attachments, could be used to have a more complete approach.

In this work, it is estimated that the use of machine learning techniques applied to the detection of a data breach will contribute favorably to the area of information security by exposing an approach to the detection of a data breach through the analysis of user behavior.

References

1. Nieles, M., Dempsey, K., Pillitteri, V.Y.: An introduction to information security. NIST Spec. Publ. **800**(12) (2017). https://doi.org/10.6028/NIST.SP.800-12r1
2. Bement, A.L.: Standards for Security Categorization of Federal In-formation and Information Systems. FIPS, 199 (2004)
3. NIST. csrc.nist.gov/glossary/term/Information_Technology_Laboratory_NIST/ (2019). Accessed 29 Jan 2020
4. The Identity Theft Resource Center. End of Year Data Breach (2019)
5. Layton, R., Watters, P.A.: A methodology for estimating the tangible cost of data breaches. J. Inf. Secur. Appl. **19**(6), 321–330 (2014). https://doi.org/10.1016/j.jisa.2014.10.012
6. The Impact of Data Breaches on Reputation and Share Value. The Ponemon Institute (2017)
7. Cost of a Data Breach Report 2019. Security I., Institute P. (2019)
8. Jaquith, A., Balaouras, S., Crumb, A.: The Forrester WaveTM: Data Leak Prevention Suites, Q4 2010 (2010)
9. Alneyadi, S., Sithirasenan, E., Muthukkumarasamy, V.: A survey on data leakage prevention systems. J. Netw. Comput. Appl. **62**, 137–152 (2016). https://doi.org/10.1016/j.jnca.2016.01.008
10. Petkovic, M., Popovic, M., Basicevic, I., Saric, D.: A host based method for data leak protection by tracking sensitive data flow. In: Proceedings—2012 IEEE 19th International Conference and Workshops on Engineering of Computer-Based Systems, ECBS 2012, pp. 267–274. IEEE, Serbia (2012). https://doi.org/10.1109/ECBS.2012.5
11. McCandless, D., Evans, T., Barton, P., et al.: World's Biggest Data Breaches. www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/ (2020). Accessed 31 Jul 2020
12. Creative Research Systems. Sample Size Calculator. www.surveysystem.com/sscalc.htm (2012). Accessed 31 Jul 2020
13. Marriott International. Marriott International Notifies Guests of Property System Incident. news.marriott.com/news/2020/03/31/marriott-international-notifies-guests-of-property-system-incident (2020). Accessed 19 Aug 2020
14. MacFarlane, J.: 4.2 million Desjardins members affected by data breach, credit un-ion now says. www.cbc.ca/news/canada/montreal/desjardins-data-breach-1.5344216 (2019). Accessed 19 Aug 2020

15. BBC News. French police hit by security breach as data put online. www.bbc.com/news/world-europe-36645519 (2016). Accessed 12 Aug 2020
16. The Straits Times. 20 million people in South Korea fall victim to latest data leak. www.straitstimes.com/asia/20-million-people-in-south-korea-fall-victim-to-latest-data-leak (2014). Accessed 19 Aug 2020
17. Krebs, B.: Experian Sold Consumer Data to ID Theft Service. krebsonsecurity.com/2013/10/experian-sold-consumer-data-to-id-theft-service/ (2013). Accessed 19 Aug 2020
18. Wired. AOL Worker Sells 92 Million Names. www.wired.com/2004/06/aol-worker-sells-92-million-names/ (2004). Accessed 12 Aug 2020
19. BBC News. Monster attack steals user data. news.bbc.co.uk/2/hi/6956349.stm (2007). Accessed 19 Aug 2020
20. The Guardian. Australian National University hit by huge data breach. www.theguardian.com/australia-news/2019/jun/04/australian-national-university-hit-by-huge-data-breach (2019). Accessed 19 Aug 2020
21. Doe, D.: IN: St. Vincent Medical Group notifies patients after successful phishing attempt compromises PHI. www.databreaches.net/in-st-vincent-medical-group-notifies-patients-after-successful-phishing-attempt-compromises-phi/ (2015). Accessed 19 Aug 2020
22. Roman, J.: Chase Breach: Prosecutors Demand Details. www.bankinfosecurity.com/chase-breach-prosecutors-demand-details-a-7798 (2015). Accessed 19 Aug 2020
23. University of New Brunswick, Canadian Institute for Cybersecurity. A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018). registry.opendata.aws/cse-cic-ids2018/ (2019). Accessed 31 Jan 2020
24. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3) (2009). https://doi.org/10.1145/1541880.1541882
25. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
26. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. Neural Comput. **12**(10), 2451–2471 (2000). https://doi.org/10.1162/089976600300015015
27. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005). https://doi.org/10.1016/j.neunet.2005.06.042
28. Wang, D., Nyberg, E.: A long short-term memory model for answer sentence selection in question answering. In: ACL-IJCNLP 2015—53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, pp. 707–712. Association for Computational Linguistics (ACL), Beijing (2015). https://doi.org/10.3115/v1/P15-2116
29. Burkov, A.: The Hundred-Page Machine Learning, illustrate. Andriy Burkov (2019)
30. Brownlee, J.: Long short-term memory networks with python. Mach. Learn. Mastery Python **1**, 228 (2017)
31. Keras.: Keras FAQ. keras.io/getting_started/faq/#what-do-sample-batch-epoch-mean (2020). Accessed 2 Feb 2020

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	• • • under matter to remain	
Insert in text the matter indicated in the margin		New matter followed by or
Delete	through single character, rule or underline or through all characters to be deleted	or
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	,
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	