Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Deep learning for insider threat detection: Review, challenges and opportunities

Shuhan Yuan[a,*], Xintao Wu[b]

[a] Utah State University, Logan, UT, USA
[b] University of Arkansas, Fayetteville, AR, USA

ARTICLE INFO

Article history:
Received 12 August 2020
Revised 23 December 2020
Accepted 7 February 2021
Available online 13 February 2021

Keywords:
Deep learning
Insider threats
Insiders
Cybersecurity

ABSTRACT

Insider threats, as one type of the most challenging threats in cyberspace, usually cause significant loss to organizations. While the problem of insider threat detection has been studied for a long time in both security and data mining communities, the traditional machine learning based detection approaches, which heavily rely on feature engineering, are hard to accurately capture the behavior difference between insiders and normal users due to various challenges related to the characteristics of underlying data, such as high-dimensionality, complexity, heterogeneity, sparsity, lack of labeled insider threats, and the subtle and adaptive nature of insider threats. Advanced deep learning techniques provide a new paradigm to learn end-to-end models from complex data. In this brief survey, we first introduce commonly-used datasets for insider threat detection and review the recent literature about deep learning for such research. The existing studies show that compared with traditional machine learning algorithms, deep learning models can improve the performance of insider threat detection. However, applying deep learning to further advance the insider threat detection task still faces several limitations, such as lack of labeled data, adaptive attacks. We discuss such challenges and suggest future research directions that have the potential to address challenges and further boost the performance of deep learning for insider threat detection.

1. Introduction

Insider threats are malicious threats from people within the organization, which usually involve intentional fraud, the theft of confidential or commercially valuable information, or the sabotage of computer systems. The subtle and dynamic nature of insider threats makes detection extremely difficult. The 2018 U.S. State of Cybercrime Survey indicates that 25% of the cyberattacks are committed by insiders, and 30% of respondents indicate incidents caused by insider attacks are more costly or damaging than outsider attacks (CSO et al., 2018).

According to the latest technical report (Costa et al., 2016) from the CERT Coordination Center (CERT/CC), a malicious insider is defined as "a current or former employee, contractor, or business partner who has or had authorized access to an organization's network, system, or data, and has intentionally exceeded or intentionally used that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems."

Compared to the external attacks whose footprints are difficult to hide, the attacks from insiders are hard to detect because the malicious insiders already have the authorized power to access the internal information systems. Insider threat detection has attracted significant attentions over

[*] Corresponding author.
E-mail addresses: shuhan.yuan@usu.edu (S. Yuan), xintaowu@uark.edu (X. Wu).

https://doi.org/10.1016/j.cose.2021.102221
0167-4048/© 2021 Elsevier Ltd. All rights reserved.

the last decade, where various insider threat detection approaches have been proposed (Eldardiry et al., 2013; Le and Zincir-Heywood, 2018; Rashid et al., 2016; Salem et al., 2008; Sanzgiri and Dasgupta, 2016; Tuor et al., 2017). Most of existing approaches achieve the insider threat detection by analyzing the users' behaviors via the audit data, such as host-based data that record activities of users on their own computers, network-based data that are recorded by network equipments, and context data that record users' profile information. A recent survey (Sanzgiri and Dasgupta, 2016) further categorizes the insider threat detection techniques into 9 classes based on strategies and features used in detection: (1) anomaly based approaches, (2) role based access control, (3) scenario based techniques, (4) decoy documents and honeypot techniques, (5) risk analysis using psychological factors, (6) risk analysis using workflow, (7) improving defense of the network, (8) improving defense by access control, and (9) process control to dissuade insiders.

Although existing approaches demonstrate good performance on insider threat detection, the traditional shallow machine learning models are unable to make full use of the user behavior data due to their high-dimensionality, complexity, heterogeneity, and sparsity. On the other hand, deep learning as a representation learning algorithm, which is able to learn multiple levels of hidden representations from the complicated data based on its deep structure (Bengio, 2013; LeCun et al., 2015), can be used as a powerful tool to analyze the user behavior in an organization to identify the potential malicious activities from insiders.

Recently, by leveraging the deep feedforward neural network, convolutional neural network (CNN), recurrent neural network (RNN), and graph neural network (GNN), several approaches have been proposed for insider threat detection (Hu et al., 2019; Jiang et al., 2019; Liu et al., 2018b; Lu and Wong, 2019). For example, some RNN-based models are proposed to analyze the user sequential data to identify the malicious activities (Lu and Wong, 2019; Yuan et al., 2019; Zhang et al., 2018a), while a GNN-based model is investigated to detect the insiders based on the user structural data in an organization (Jiang et al., 2019). However, using deep learning for insider threat detection still faces various challenges related to the characteristics of insider threat detection data, such as extremely small number of malicious activities and adaptive attacks. Hence, developing advanced deep learning models that can improve the performance of insider threat detection is still under-explored.

Currently, there is no existing review on the topic of deep learning for insider threat detection. We do not aim to provide a comprehensive survey on the domain of insider threat (see Homoliak et al., 2019; Liu et al., 2018c for details) or a general review on deep learning (see LeCun et al., 2015). In this work, we focus on reviewing the current progresses and pointing out potential future directions of deep learning for insider threat detection. We first briefly review the deep learning and its application on anomaly detection in Section 2. In Section 3 we introduce commonly-used datasets used in insider threat detection, explain why deep learning is needed for insider threat detection, and review a few recent research works of deep learning based insider threat detection. In Section 4, we identify and categorize challenges into extremely unbalanced data, subtle attacks, temporal information in attacks, heterogeneous data fusion, adaptive threats, fine-grained detection, early detection, explainability, lack of testbed, and lack of practical metrics. In Section 5, we point out research opportunities of insider threat detection based on few-shot learning, self-supervised learning, deep marked temporal point process, multi-modal learning, deep survival analysis, deep Bayesian nonparametric learning, deep reinforcement learning, explainable deep learning in addition to testbed development. Finally we conclude this survey in Section 6.

The main contributions of this survey are as follows. First, to the best of our knowledge, this is the first survey about using deep learning techniques to tackle the challenges of insider threat detection, where we summarize the state-of-the-art deep learning models for insider threat detection. Second, we summarize ten existing challenges based on the characteristics of insiders and insider threats. Third, we point out ten future directions to improve the performance of deep learning models for insider threat detection.

2. Deep learning and its application on anomaly detection

2.1. *Deep learning*

With great achievement in various domains, such as computer vision (He et al., 2016), natural language processing (Liu et al., 2020), and speech recognition (Deng et al., 2013), deep learning has dominated the machine learning community in the past few years (LeCun et al., 2015). Compared with traditional machine learning models that heavily rely on hand engineering to identify useful features to represent raw data, deep learning models are able to learn semantic representations from the raw data with minimal human efforts. The deep learning models as representation learning models adopt a multi-layer structure to learn data representation, where the lower layers capture the low-level features of data, while the higher layers extract the high-level abstract.

Deep learning models can be broadly categorized into four groups based on their architectures: (1) deep feedforward neural network (DFNN), which includes a number of deep learning models consisting of multiple layers, such as deep belief network (Hinton et al., 2006), deep Boltzmann machine (Salakhutdinov and Hinton, 2009) and deep autoencoder (Vincent et al., 2008); (2) convolutional neural network (CNN), which leverages the convolutional and pooling layers to achieve the shift-invariant property; (3) recursive neural network (RvNN), which takes a recursive data structure of variable sizes and makes predictions in a hierarchical structure; (4) recurrent neural network (RNN), which maintains an internal hidden state to capture the sequential information. Since the deep learning field is growing so fast, many new types of deep structures are proposed each year. The readers can refer to recent surveys, e.g., Arulkumaran et al. (2017); Chalapathy and Chawla (2019); Pouyanfar et al., 2018; Zhang et al. (2018b,c), to learn more information about deep learning and its applications in various domains.

2.2. Deep learning for anomaly detection

Anomaly detection is to identify instances that are dissimilar to all others, which is an important problem with multiple applications, such as fraud detection, intrusion detection, and video surveillance (Chandola et al., 2009). Anomalies are referred to as abnormalities, deviants, or outliers in the data mining and statistics literature and roughly speaking insider threats can be treated as one type of anomalies. A variety of machine learning and deep learning based anomaly detection approaches have been developed, however, they are not necessarily applicable for insider threat detection due to the characteristics of insider threat as will be shown in Section 3. A recent survey (Chalapathy and Chawla, 2019) categorizes the deep learning-based anomaly detection into three groups based on the availability of labels, i.e., supervised, semi-supervised, and unsupervised deep anomaly detection. In some cases, when both normal and anomalous data are available, supervised deep anomaly detection approaches are proposed for binary or multi-class classification (Chalapathy et al., 2016a; 2016b). A more common scenario is that it is easy to collect many normal samples while only a small number of anomalous samples is available, so that semi-supervised deep anomaly detection can be adopted by leveraging the normal samples to separate outliers (Akcay et al., 2018; Song et al., 2017; Zheng et al., 2019). When no labeled data are available, unsupervised deep anomaly detection is applied to detect anomalies based on intrinsic properties of data samples (Hendrycks et al., 2018; Su et al., 2019; Tuor et al., 2017).

Although many deep learning-based approaches for anomaly detection have been proposed, the performance improvement of deep learning for anomaly detection may not be as significant as deep learning for other domains, such as computer vision and natural language process. The main potential reason is that deep learning models consisting of millions of parameters require a large amount of labeled data for training properly. However, for anomaly detection, it is very difficult, if not impossible, to collect a large number of labeled anomalies in the training data.

3. Literature review

Insider threats as one of the most challenging attacks to address in cyberspace have attracted much attention for a long time. Various insider threat detection approaches, such as hidden Markov model (HMM) and support vector machine (SVM), have been proposed in literature (Homoliak et al., 2019; Sanzgiri and Dasgupta, 2016). However, leveraging the deep learning models for insider threat detection is not well-explored in literature as only a few papers are available. It is natural to question the need of deep learning models for insider threat detection. In this section, we first describe the insider threats that have been reported in literature. Then, we introduce some well-known insider threat datasets and further focus on a widely used insider threat dataset to illustrate the challenges of insider threat detection by describing its characteristics. After that, we present why deep learning based insider threat detection models are needed. At the end of this section, we review the existing deep learning based insider threat detection papers in literature and categorize them based on the adopted deep learning architectures.

3.1. Insiders and insider threats

Taxonomy of insiders. An insider usually indicates “a person with legitimate access to an organization’s computers and networks” (Pfleeger et al., 2010). In general, there are three types of insiders: traitors, masqueraders, and unintentional perpetrators (Liu et al., 2018c). *Traitors* are insiders who misuse their privileges to commit malicious activities for financial or personal gains. For example, an employee who foresees the layoff copies confidential information from the employer and sells it to a competitor. *Masqueraders* are insiders who conduct illegal actions on behalf of legitimate employees of an institute. For example, attackers leverage some technical methods or social engineering to obtain the login authority and then access confidential assets of an organization. *Unintentional perpetrators* are insiders who unintentionally make mistakes and expose confidential information to outsiders. For example, an employee does not pay attention to the security policy of an organization and uses an infected USB drive on a workstation, which enables the attackers to compromise the internal system.

Taxonomy of insider threats. Insider threats indicate the “threats with malicious intent directed toward organizations” by insiders (Gavai et al., 2015). Based on the malicious activities conducted by the insiders, the insider threats can also be categorized into three types: IT sabotage, theft of intellectual property, and fraud (Homoliak et al., 2019). *IT sabotage* indicates directly using IT to make harm to an organization, which is usually conducted by insiders with technical skills. *Theft of intellectual property* indicates stealing crucial information from the institute, such as customer information or source code, which can be conducted by technical staff or non-technical staff. *Fraud* indicates unauthorized modification, addition, or deletion of data. The motivation is usually for financial gain.

3.2. Datasets

Datasets are critical for research on insider threat detection. Currently, there is no comprehensive real-world dataset publicly available for insider threat detection. The most existing public datasets usually generate synthetically executed attack sessions to simulate the scenarios of insider threats. A recent survey (Homoliak et al., 2019), present existing datasets gathered from laboratory experiments and the real world and further group commonly used eleven datasets into five categories: masquerader-based, traitor-based, substituted masqueraders, identification/authentication-based, and miscellaneous malicious. We first briefly review the important datasets in each category. Table 1 summarizes the basic information of the datasets.

- **RUU dataset** (Salem and Stolfo, 2011) is a masquerader-based dataset that consists of host-based events from 34 normal users and 14 masqueraders. 14 volunteers serve as masqueraders to find data with financial value.
- **Enron dataset** (Cohen, 2009) is a traitor-based dataset that consists of half-million real-world emails from 150

Table 1 – Widely-used datasets for insider threat detection.

Dataset	Category	Statistics
RUU	Masquerader	34 normal users and 14 masqueraders
Enron	Traitor	Half million emails from 150 emplyees
Schonlau	Substituted Masquerader	Unix shell commands from 50 users
Greenberg	Authentication	Full Unix shell commands from 168 users
TWOS	Miscellaneous Malicious	24 users; 12 masquerader and 5 traitor sessions
CERT	Miscellaneous Malicious	3995 normal users and 5 insiders

employees in Enron Corporation. This dataset does not have the ground-truth information, but can be used for insider threat detection based on text and social network analysis.

- **TWOS dataset** (Harilal et al., 2017) is a miscellaneous dataset including both masqueraders and traitors. The dataset contains activities from 24 users in 5 days collected based on a multi-player game, where 12 masquerader and 5 traitor sessions are simulated. The dataset consists of logs from various data sources such as mouse, keystroke, network, and host monitor logs of system calls.
- **CERT dataset** (Glasser and Lindauer, 2013) is a synthetic dataset that includes system logs with labeled insider threat activities. Most of the recent deep learning-based studies adopt the CERT dataset to evaluate their proposed approaches. We will introduce the detailed information of this dataset in the following subsection.
- **Schonlau dataset** (Schonlau et al., 2001) is a substituted masquerader dataset that consists of 50 users. Each user generates 15,000 Unix shell commands. The commands in masquerade sessions are randomly injected from unknown users.
- **Greenberg's dataset** (Greenberg, 1988) is an authentication-based dataset that contains full commands of Unix C shell from 168 different users. Different from the Schonlau dataset, Greenberg's dataset includes arguments and timestamps in command entries. When using this dataset for insider threat detection, several users are randomly selected to serve as the source of masqueraders (Maxion, 2003).

3.2.1. *CERT insider threat dataset*

Most of the recent deep learning-based studies adopt the CMU CERT dataset to evaluate their proposed approaches. In this survey, we introduce the CERT dataset to illustrate challenges for insider threat detection.

The CERT division of Software Engineering Institute at Carnegie Mellon University maintains a database of more than 1000 real case studies of insider threat and has generated a collection of synthetic insider threat datasets using scenarios containing traitor instances and masquerade activities.

CERT dataset consists of five log files that record the computer-based activities for all employees in a simulated organization, including **logon.csv** that records the logon and logoff operations of all employees, **email.csv** that records all the email operations (send or receive), **http.csv** that records all the web browsing (visit, download, or upload) operations, **file.csv** that records activities (open, write, copy or delete) involving a removable media device, and **decive.csv** that records the usage of a thumb drive (connect or disconnect). Table 2 lists the activity types in each log file. For each activity, it also contains related descriptions. For example, the activity type "Send Internal Email" includes time, sender, receivers, subject, and content information. Besides the employees' activity data on computers, the CERT dataset also provides the psychometric score for each employee, known as "Big Five personality traits", in **psychometric.csv**. The data generation process reflects practical constraints and employs a number of different model types, e.g., graph models of the population's social structure, topic models for generating content, psychometric models for dictating behavior and preferences, models workplace behaviors.

Table 2 – Activity types in log files.

Files	Operation types
logon.csv	Weekday Logon (employee logs on a computer on a weekday at work hours)
	Afterhour Weekday Logon (employee logs on a computer on a weekday after work hours)
	Weekend Logon (employees logs on at weekends)
	Logoff (employee logs off a computer)
email.csv	Send Internal Email (employee sends an internal email)
	Send External Email (employee sends an external email)
	View Internal Email (employee views an internal email)
	View external Email (employee views an external email)
http.csv	WWW Visit (employee visits a website)
	WWW Download (employee downloads files from a website)
	WWW Upload (employee uploads files to a website)
device.csv	Weekday Device Connect (employee connects a device on a weekday at work hours)
	Afterhour Weekday Device Connect (employee connects a device on a weekday after hours)
	Weekend Device Connect (employee connects a device at weekends)
	Disconnect Device (employee disconnects a device)
file.csv	Open doc/jpg/txt/zip File (employee opens a doc/jpg/txt/zip file)
	Copy doc/jpg/txt/zip File (employee copies a doc/jpg/txt/zip file)
	Write doc/jpg/txt/zip File (employee writes a doc/jpg/txt/zip file)
	Delete doc/jpg/txt/zip File (employee deletes a doc/jpg/txt/zip file)

There are several versions of datasets according to when the datasets were created. The most widely-used versions are

Table 3 – The statistics of CERT datasets r4.2 and r6.2.

	# employees	# insiders	# activities	# malicious activities
r4.2	1000	70	32,770,227	7323
r6.2	4000	5	135,117,169	470

Table 4 – The numbers of activities, malicious activities, sessions, and malicious sessions for each insider.

	ACM2278	CDE1846	CMP2946	MBG3183	PLJ1771
Activity #	31,370	37,754	61,989	42,438	20,964
Malicious activity #	22	134	242	4	18
Session #	316	374	627	679	770
Malicious session #	2	9	53	1	3

r4.2 and r6.2. Table 3 shows the statistics of these two datasets. In short, r4.2 is a “dense” dataset that contains many insiders and malicious activities, while r6.2 is a “sparse” dataset that contains 5 insiders and 3995 normal users. For each user, it records activities from January 2010 to June 2011. On average, the number of activities for each employee is around 40000. Specifically, the CERT in r6.2 dataset simulates the following five scenarios of attacks from insiders.

- User ACM2278 who did not previously use removable drives or work after hours begins logging in after hours, use a removable drive, and upload data to wikileaks.org.
- User CMP2946 begins surfing job websites and soliciting employment from a competitor. Before leaving the company, they use a thumb drive to steal data.
- System administrator PLJ1771 becomes disgruntled. Downloads a keylogger and uses a thumb drive to transfer it to his supervisor’s machine. The next day, he uses the collected keylogs to log in as his supervisor and send out an alarming mass email, causing panic in the organization. He leaves the organization immediately.
- User CDE1846 logs into another user’s machine and searches for interesting files, emailing to their home email.
- User MBG3183, as a member of a group decimated by layoffs, uploads documents to Dropbox, planning to use them for personal gain.

Table 4 shows the statistics of each of the above five insiders, including the numbers of activity, malicious activity, session, and malicious sessions. In sum, insider threat detection is a task like looking for a needle in a haystack, thus it is generally infeasible to manually define features or use shallow machine learning models to detect insider threats.

3.3. *Why deep learning for insider threat detection?*

Among the many attractive properties of deep learning models, the potential advantages of deep learning for insider threat detection can be summarized as follows.

- **Representation Learning**. The most significant advantage of deep learning models is based on the capability to automatically discover the features needed for detection. The user behavior in cyberspace is complicated and non-linear. Manually designed features are hard and inefficient to capture user behavior information. Meanwhile, the learning models with shallow structures, such as HMM and SVM, are relatively simple structures with only one layer for transforming the raw feature into a high-level abstract that can be used for detection. These shallow models are effective for solving many well-constrained problems, but shallow models with limited capability are hard to model complicated user behavior data. Comparatively, deep learning models are able to leverage deep non-linear modules to learn the representation by using a general-purpose learning procedure. Hence, it is a natural fit to use deep learning models to capture complex user behavior and precisely detect user’s intentions, especially those malicious ones.
- **Sequence Modeling.** Deep learning models, such as recurrent neural network (RNN) and the newly-proposed Transformer, have shown promising performance in modeling sequential data, like video, text, and speech (Graves, 2013; Vaswani et al., 2017). Since it is natural to represent the user activities recorded in audit data as sequential data, leveraging RNN or Transformer to capture the salient information of complicated user behavior has the great potential to boost the performance of insider threat detection.
- **Heterogeneous Data Composition.** Deep learning models also have achieved great performance on tasks that fuse heterogeneous data, such as image captioning (Cheng et al., 2017; Huang et al., 2019). For insider threat detection, besides modeling the user activity data as sequences, other information, such as user profile information and user structure information in an organization, is also critical. Combining all the useful data for insider threat detection is expected to achieve better performance than only using a single type of data. Compared with traditional machine learning methods, deep learning models are more powerful to combine the heterogeneous data for detection.

3.4. *Deep learning for insider threat detection*

In this subsection, we review the main literature and categorize deep learning based insider threat detection papers based on the adopted deep learning architectures. Table 5 summarizes all the papers discussed in this section, and

Table 5 – Categorization of deep learning based insider threat detection papers discussed in this section.

Model	Paper	Training	Granularity		
			Insider	Session	Activity
Deep Feed-forward Neural Network	Liu et al. (2018b)	Unsupervised		√	
	Lin et al. (2017)	One-class		√	
Recurrent Neural Network	Lu and Wong (2019)	Unsupervised		√	
	Tuor et al. (2017)	Unsupervised		√	
	Zhang et al. (2018a)	Unsupervised		√	
	Yuan et al. (2019)	Unsupervised		√	
	Yuan et al. (2018)	Supervised		√	
Convolutional Neural Network	Hu et al. (2019)	Supervised		√	
Graph Neural Network	Jiang et al. (2019)	Supervised	√		
	Liu et al. (2019)	Unsupervised			√

Table 6 – Advantages and limitations of each type deep learning model for insider threat detection.

Model	Advantage	Limitation
DFNN	The idea of using deep autoencoder for anomaly detection is intuitive	Cannot capture the temporal information
RNN	Capture the temporal information of user activity sequences	Could face high false alert if users change the daily pattern instead of conducting malicious activities
CNN	High accuracy if the user activity data can be represented as images	The data that are suitable for CNN are limited in the insider threat detection area (only images data is suitable)
GNN	Powerful to model the graph data, such as organization information networks (social network, emails)	When graph data is not available, it requires a lot of manual work to build a graph

Table 6 further summarizes the advantages and disadvantages of each type of deep learning models for insider threat detection.

As shown in Table 5, due to the extremely unbalanced nature of the dataset, most of the proposed approaches adopt the unsupervised learning paradigm for insider threat detection. For the detection granularity, most of the papers focus on detecting malicious subsequence (e.g., activities in 24 h) or malicious session. Here, a session indicates a sequence of activities among “Logon” and “Logoff”. If there are malicious activities in a session (subsequence), the session (subsequence) will be labeled as a malicious session (subsequence). Due to the limited information that can be leveraged, detecting malicious activities is difficult. Currently, there is only one work focusing on activity level detection.

3.4.1. *Deep feedforward neural network*

Deep feedforward neural network (FNN) is a classical type of deep learning model. Various deep learning models are feedforward neural networks, such as deep autoencoder, deep belief network, and deep Boltzmann machine (Pouyanfar et al., 2018). These deep neural networks are able to learn different levels of representations from the input data based on the multi-layer structures.

Several studies have proposed to use the deep feadforward neural network for insider threat detection. Liu et al. (2018b) uses deep autoencoder to detect the insider threat. Deep autoencoder consists of an encoder and a decoder, where the encoder encodes the input data to hidden representations while the decoder aims to reconstruct the input data based on the hidden representations. The objective of the deep autoencoder is to make the reconstructed input close to the original input. Because the majority of activities in an organization are benign, the input with insider threats should have relatively high reconstruction errors. As a result, the reconstruction error of the deep autoencoder can be used as an anomalous score to identify the insider threat. Another idea of leveraging the autoencoder structure is that after learning the hidden representations based on the reconstruction error, a one-class classifier, such as one-class SVM, is applied on the learned hidden representations to identify the insider threats (Lin et al., 2017).

The advantage of deep feedforward neural networks is that the idea of using deep feedforward neural networks, such as deep autoencoder, to derive anomalous scores for insider threat detection is straightforward and easy to implement. On the other hand, the major disadvantage of deep feedforward neural networks is that they cannot capture the temporal information of user activities.

3.4.2. *Recurrent neural network*

Recurrent neural network (RNN) is mainly used for modeling the sequential data, which maintains a hidden state with a self-loop connection to encode the information from a sequence (Graves, 2013). The standard RNN is difficult to train over long sequences due to the vanishing or exploding gradient problem (Bengio et al., 1994). Currently, two variants of the standard RNN, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2015), are widely-used to model the long sequences and capture the long-time dependence by incorporating gate mechanisms.

The user activities on a computer can be naturally modeled as sequential data. As a result, many RNN based approaches have been proposed to model the user activities (Lu and Wong, 2019; Tuor et al., 2017; Yuan et al., 2018; 2019) for insider threat detection. The basic idea is to train an RNN model to predict the user's next activity or period of activities. As long as the prediction results and the user's real activities do not have significant differences, we consider the user follows the normal behavior. Otherwise, user activities are suspicious. Tuor et al. (2017) proposes a stacked LSTM structure to capture the user activities in a day and adopts negative log-likelihoods of user activities as the anomalous scores to identify malicious sessions. Rather than only using the activity type, e.g., web visiting or file uploading, for insider threat detection, Yuan et al. (2019) proposes a hierarchical neural temporal point processes model to capture both activity types and time information in a user session and then derives an anomalous score based on the differences between the predicted results and real activities in terms of types and time.

The major advantage of recurrent neural networks is that RNN can capture the temporal information from the user activity sequences based on its recurrence equation. Hence, RNN is more suitable for insider threat detection with sequential data as inputs. However, in order to train the recurrent neural network, the objective function is to predict future activities instead of detecting anomalous sequences. As a result, the RNN model could face the issue of high false alerts if users change their activity patterns instead of conducting malicious activities.

3.4.3. *Convolutional neural network*

Convolutional Neural Network (CNN) has achieved great success in computer vision. A typical CNN structure consists of a convolutional layer followed by a pooling layer and a fully connected layer for prediction. The convolutional and pooling layer ensures that the extracted features from inputs are rotational and positional invariant, which is a very useful property for image classification. The modern CNNs are extremely deep with tens of convolutional and pooling layers, which can achieve promising performance for image classification (He et al., 2016; Krizhevsky et al., 2012).

A recent study on insider threat detection proposes a CNN-based user authentication method by analyzing mouse bio-behavioral characteristics (Hu et al., 2019). The proposed approach represents the user mouse behaviors on a computer as an image. If an ID theft attack occurs, the user mouse behaviors will be inconsistent with the legal user. Hence, a CNN model is applied on images generated based on the mouse behavior to identify potential insider threats.

The major advantage of CNN for insider threat detection is that it is very powerful for modeling image data. Hence, once we have the data, such as mouse bio-behavioral characteristics, that can be represented as images, CNN can be used for capturing the information. However, since it is not easy to collect the data that are suitable for CNN in the field of insider threat detection, the application scenarios of CNN are relatively limited.

3.4.4. *Graph neural network*

Graph neural network (GNN), which is able to model the relationships between nodes, has gained increasing popularity for graph analysis (Scarselli et al., 2009; Wu et al., 2019). A wildly-used GNN model is a graph convolutional network (GCN) that uses graph convolutional layer to extract node information. The graph convolutional layer has similar properties of a typical convolutional layer, such as local connections and shared weights, which are suitable for graph analysis. For example, the nodes in graphs are usually locally connected so that the convolutional layer is able to aggregate the feature information of a node from its neighbors. Besides GCN, graph embedding, which aims at learning low-dimensional latent representation of nodes in a network, has also attracted a lot of attention. The learned node representations can be used as features for various graph analysis tasks, such as classification, clustering, link prediction, and visualization (Chen et al., 2018).

Recent work (Jiang et al., 2019) adopts a GCN model to detect insiders. Since users in an organization often make connections to each other via email or operation on the same devices, it is natural to use a graph structure to capture the inter-dependencies among users. Besides taking the adjacency matrix of structural information as input, GCN also incorporates the rich profile information about the users as the feature vectors of nodes. After applying the convolutional layers for information propagation based on the graph structure, GCN adopts the cross-entropy as the objective function to predict malicious nodes (users) in a graph.

Inspired by graph embedding methods, research in Liu et al. (2019) proposes log2vec to detect malicious activities. Log2vec first constructs a heterogeneous graph by representing various activities in audit data as nodes and rich relationships among activities as edges and then train node embeddings that can encode activity relationships. Finally, by applying clustering algorithms on the node embeddings, log2vec is able to separate malicious and benign activities into different clusters and identify malicious ones.

The advantage of GCN is that it can capture the hidden patterns of graph data. Since employees in an organization can be naturally represented as a graph, GCN provides a new aspect to detect insiders in the organization. However, if the data, such as user activity sequences, cannot be explicitly modeled as a graph structure, it requires a lot of manual engineering to construct a graph (Liu et al., 2019).

4. Challenges

Although some progress has been achieved using deep learning models for insider threat detection, there are many unsolved challenges from the perspectives of characteristics of underlying data, insider threat, user expectations of detection algorithms, testbed and evaluation metrics development. In the following, we highlight the identified ten key challenges:

- **Extremely Unbalanced Data.** Compared with the benign activities, the malicious activities from insiders are extremely rare in real-world scenarios. Therefore, the insider threat dataset is an unbalanced dataset, which is a big challenge to train deep learning models. In general, deep learning models, which consist of tons of parameters, require large amounts of labeled data to train properly. However, it is infeasible to collect a large number of malicious insiders in reality. How to leverage the existing small samples to properly train the deep learning models is crucial to the insider threat detection task.

→ time information

- **Temporal Information in Attacks.** Most of the existing approaches for insider threat detection only focus on the activity type information, such as copying files to a removable disk or browsing a Web page. However, it is insufficient to detect attacks simply based on activity types conducted by users as the same activity could be either benign or malicious. A simple case is that copying files in working hours looks normal, but copying files at mid-night is suspicious. The temporal information plays an important role in analyzing user behavior to identify those malicious threats, and how to incorporate such temporal information is challenging.
- **Heterogeneous Data Fusion.** Besides the temporal information, leveraging various data sources and fusing such heterogeneous data are also critical to improve the insider threat detection. For example, a user who copies files in a daily routine foresees his potential layoff and has activities of copying credential files to the removable disk in purpose. In such scenarios, considering the user profile (i.e., psychometric score) or user interaction data could help to identify potential insider threats.
- **Subtle Attacks.** Currently, most of the existing work consider the insider threat detection task as the anomaly detection task, which usually models anomalous samples as out-of-distribution samples. The existing models are usually trained on samples from benign users and then applied to identify insiders that are dissimilar to observed benign samples. A threshold or anomalous score is derived to quantify the dissimilarity between insiders and benign users. However, in reality, we cannot expect insiders have a significant pattern change to conduct malicious activities. In order to evade detection, insider threats are subtle and hard to notice, which means that insiders and benign users are close in the feature space. The traditional anomaly detection approaches cannot detect insiders that are close to benign users. Moreover, recently, in order to bypass the authentication models, the adversarial attacks that are intentionally designed to mislead the model to make a mistake also cause big security issues in cyberspace. For example, recent work (Marcus Tan et al., 2019) proposed adversarial attack strategies on remote user authentication which can be used by insiders to conduct malicious activities.
- **Adaptive Threats.** The insiders always improve attacking strategies to evade detection. However, the learning-based models are unable to detect new types of attacks after training. It is inefficient to train the models from scratch again when new types of attacks are observed. First, it usually needs some time to collect enough samples to train the model. More importantly, the re-training strategy cannot ensure in-time detection and prevention. Designing a model that can adaptively improve the performance of insider threat detection is an important and challenging task.
- **Fine-grained Detection.** The existing deep learning based approaches usually detect malicious sessions that contain malicious activities. However, users usually conduct a large number of activities in a session. Such coarse-grained detection faces the problem that it is hard to achieve in time detection. Hence, how to identify the fine-grained malicious subsequence or the exact malicious activity is important for insider threat detection. It is also a very challenging task. This is because the information we can leverage from each activity is very limited, i.e., we only observe when and what activity is conducted by a user. Without enough information, it is hard to achieve fine-grained insider threat detection.
- **Early Detection.** Current approaches focus on insider threat detection, which means malicious activities already occur and the significant loss is already caused to organizations. Hence, an emerging topic is how to achieve the insider threat early detection, i.e., detecting potential malicious activities ahead of they actually happen. Several approaches are proposed to defend the insider threat by using general IT security mechanisms (Alneyadi et al., 2016; Shabtai et al., 2012), but there is no learning-based approach to achieve early detection. Proactively identifying users who have high chances to conduct malicious activities in the near future is critical so that the organizations can conduct the intervention ahead to prevent or reduce the loss.
- **Interpretability.** Deep learning models are usually considered as black boxes. Although deep learning can achieve promising performance in many domains, the reason why the models work is still under-exploited. When an employee is detected as an insider, it is critical to understand the reason why the model makes such predictions since employees are usually the most valuable asset in an organization. Especially, deep learning models cannot achieve 100% of accuracy on insider threat detection. The false positive cases (misclassifying benign users as insiders) can seriously affect the loyalty of employees to the organization. Hence, the model interpretability is a key to provide the insight of the model to domain-expert so that further actions can be conducted with high confidence.
- **Lack of Testbed.** Currently, there is no real-world dataset that is publicly-available for researchers. Although the CERT dataset tries to provide comprehensive information that is close to the high level of human realism, there is still a gap between the synthetic data and real-world scenario.
  - *Data Complexity.* Since the CERT dataset is a synthetic dataset, most of the activities are randomly generated with limited complexity. For example, the websites that are accessed by employees are very limited. As a result, some insider threats, such as visiting wikileak.org, can be easily identified. Meanwhile, the fine-grained user activities are randomly generated, so there is no daily routine pattern in this dataset. Furthermore, most of the activity time in the dataset are randomly generated. As a result, it is hard to leverage the temporal information to detect insider threats based on this dataset.

Table 7 – Advantages and limitations of potential research topics.

Research topic	Advantage	Limitation
Few-shot Learning	Achieve insider threat detection with limited data	Hard to detect new type of attack that is significantly different from the observed ones
Self-supervised Learning	Achieve insider threat detection without using any labeled information	Require hand-crafted rules that are tailored to each dataset
Deep Marked Temporal Point Process	Capture the temporal information in terms of time	Require a large amount of samples for training
Multi-model Learning	Capture the user information from multiple perspective	Hard to obtain multi modality data
Deep Survival Analysis	Achieve the insider threat early detection	Require a number of event samples
Deep Bayesian Nonparametric Model	Capture fine-grained user activity patterns	High time complexity
Deep Reinforcement Learning	Keep improving the capability to identify insider threats via the reward function	Hard to design a proper reward function

  - *Insider Threat Complexity*. The insider threat scenarios simulated in the datasets are also narrow compared with the various insider threats conducted in the real-world. The latest version of CERT dataset only consists of five scenarios. Consequently, the proposed approaches, which can achieve reasonable performance on this dataset, may not be able to achieve good performance in practice. Meanwhile, even for the five insider threat scenarios, the difficulty of identifying these insider threats are different. As shown by the ROC curves in most of the existing papers (Lin et al., 2017; Liu et al., 2018b; Lu and Wong, 2019; Yuan et al., 2019), many studies can achieve 80% of the true positive rate with a low false negative rate. However, the false negative rate increases significantly when the true positive rate keeps increasing. It means 80% of the insider threats in this dataset can be well detected while the rest of 20% insider threats are hard to detect.

- **Lack of Practical Evaluation Metrics.** The commonly-used classification metrics, such as *true positive rate* (*TPR*), false positive rate (FPR), precision, and recall, are adopted to evaluate performance of insider threat detection. Based on the TPR and FPR, an receiver operating characteristic (ROC) curve can be drawn by setting FPR and TPR as x and y axes, respectively, which indicates the trade-off between true positive and false positive. Ideally, we expect an insider threat detection algorithm can achieve TPR to be 1 and FPR to be 0. Currently, in literature, the ROC area under curve (AUC) score is widely-used to compare the performance of different detection algorithms (Lin et al., 2017; Liu et al., 2019; 2018b; Lu and Wong, 2019; Yuan et al., 2019). Another metric is the precision-recall (PR) curve, which is a plot of recall and precision as x and y axes and adopted in evaluating the unbalanced data classification. Compared to ROC-AUC, PR-AUC may be more useful to evaluate the algorithms for insider threat detection because the PR curve more focuses on the performance of classifiers on the minority class. However, due to the extremely small number of insiders and the corresponding malicious activities, it is unclear whether ROC-AUC or PR-AUC is practical to evaluate insider threat detection. For example, the ROC-AUC values from different detection algorithms are often close (Liu et al., 2018b; Lu and Wong, 2019; Yuan et al., 2019), which means that it is hard to identify a better model based on ROC-AUC values.

5. Future directions

The above challenges lead to several opportunities and future research directions to improve the performance of deep learning models for insider threat detection. We point out the following topics, which we believe would be promising for future research. Table 7 summarizes the key advantages and limitations of some potential research topics.

- **Few-shot Learning based Insider Threat Detection.** Few-shot learning aims at classifying samples from unknown classes given only a few labeled samples (Wang et al., 2020). Few-shot learning can further extend to a more rigorous setting, one-shot learning (Fei-Fei et al., 2006) or zero-shot learning (Wang et al., 2019a), where only one or totally no labeled sample is available. Consider the extremely small number of insiders, few-shot learning is a natural fit for insider threat detection. To tackle the challenge of a few labeled samples, few-shot learning leverage the prior knowledge. Based on how to use the prior knowledge, the existing few-shot learning algorithms can be categorized into three groups, the data based approaches which augment training data by prior knowledge, the model based approaches which use the prior knowledge to constrain hypothesis space, and the algorithm based approaches which alter search strategy in hypothesis space by prior knowledge (Wang et al., 2020).
  The advantage of few-shot learning is that it can achieve insider threat detection by using only limited samples based on prior knowledge. The limitation is that the current few-shot learning assumes a fixed task distribution. Once a new type of attack, which is significantly different from the observed once, is conducted, the few-shot learning model may not be able to detect such attacks.
- **Self-supervised Learning based Insider Threat Detection.** Self-supervised learning aims at training a model using labels that can be easily derived from the input data rather than requiring human efforts to label the data (Arandjelovic and Zisserman, 2017; Asano et al., 2019;

Wang et al., 2019b). Self-supervised learning has achieved great success in computer vision and natural language processing. A typical self-supervised learning task in natural language processing is to pretrain a deep learning model by a language model, which is trained to predict the next word of a sentence. The task we use to pretrain the deep learning model is called “pretext task”. After pretraining, the model can achieve great performance on the “downstream tasks”, such as sentiment analysis, by further fine-tuning on very little data. The success of self-supervised learning is that via pretraining on the pretext tasks, the deep learning model is able to learn the salient information about the input data. In order to tackle the challenge of detecting the subtle insider threat, a potential research direction is to design the proper self-supervised tasks that can capture the difference between insiders and benign users.

The advantage of self-supervised learning for insider threat detection is that it has the potential to identify insiders without using any labeled information. However, the self-supervised tasks usually require hand-crafted rules that are tailored to each dataset.

- **Deep Marked Temporal Point Process based Insider Threat Detection.** Marked temporal point process is a powerful mathematical tool to model the observed random event patterns along time (Du et al., 2016; Rasmussen, 2018). Since temporal dynamics is an important aspect of user behavior, marked temporal point process is a suitable tool to analyze the user behavior in terms of activity types and time. Recently, several deep learning based marked temporal point process models have been proposed, which usually adopt the recurrent neural network to characterize the conditional intensity function in temporal point process (Du et al., 2016; Li et al., 2018; Xiao et al., 2018). Hence, using deep marked temporal point process models has the potential to improve the performance of insider threat detection by combining the user activity types and time information.

  The advantage of the deep marked temporal point process is that it can capture the temporal information of user activities in terms of time. In general, by incorporating more information, we can expect better detection results. However, the traditional temporal point process models usually make assumptions about temporal data distribution by predefined intensity functions, which may not be followed by real user activity data. On the other hand, although the deep temporal point process models do not make any assumptions, they usually need a large number of samples for training due to a large number of parameters, which is infeasible for insider threat detection. Combining the idea of few-shot learning with the deep temporal point process is a direction that is worth exploration.

- **Multi-model Learning based Insider Threat Detection.** Because the same activity could be either benign or malicious, besides the user activity data derived from the log files, leveraging other sources is also important to improve the performance of insider threat detection. In literature, several studies investigated the performance of insider threat detection via the users’ psychological data (Almehmadi and El-Khatib, 2014; Greitzer et al., 2013; Hashem et al., 2015), while some studies constructed user graph based on the organization hierarchy or email communication to identify the outliers (Gamachchi et al., 2018; Jiang et al., 2019; Moriano et al., 2017). However, how to combine the user activity data with the user profile data as well as the user relationship data is under-exploited and worth to explore.

  By incorporating data with multiple modalities, we can capture the user patterns from different perspectives, which leads to high detection accuracy. However, for insider threat detection, obtaining multi-modality data, such as users’ psychological data, is challenging due to the privacy concern.

- **Deep Survival Analysis based Insider Threat Early Detection.** Survival analysis is to model the data where the outcome is the time until the occurrence of an event of interest (Wang et al., 2019). Survival analysis is originally used in health data analysis (Liu et al., 2018a; Luck et al., 2017) and has been applied to many applications, such as predicting student dropout time (Ameri et al., 2016) or web user return time (Jing and Smola, 2017). If we consider the time when an insider conducts a malicious activity as the event of interest, we can use the survival analysis to predict when the event (conducting a malicious activity) occurs. As a result, organizations can have early alerts about potential attacks from insiders. Recently, deep learning models are adopted to model the complex survival distributions (Chapfuwa et al., 2018; Katzman et al., 2018; Luck et al., 2017).

  Leveraging deep survival analysis models has a great potential to capture the user activity time information and thus achieve insider threat early detection. The challenge of applying a deep survival analysis model is that it usually requires a large amount of event data for training, while it is hard to collect many event samples.

- **Deep Bayesian Nonparametric Model for Fine-grained Insider Threat Detection.** In order to achieve the fine-grained insider threat detection, one potential solution is to consider activities from one user in audit data as an activity stream and apply a clustering algorithm on the stream to identify the potential malicious clusters of activities. Bayesian nonparametric models, such as Dirichlet processes, are often used for data clustering and able to generate unbounded clusters (Paisley et al., 2010). The infinite nature of these models is suitable to model complicated user behavior. Recently, several Bayesian nonparametric deep generative models, which are proposed to combine the deep structure with Bayesian nonparametric (Goyal et al., 2017; Nalisnick and Smyth, 2017; Zhang and Paisley, 2018), are effective to learn rich representation based on neural networks with Bayesian methods. Leveraging the deep Bayesian nonparametric models has the potential to achieve fine-grained insider threat detection.

  The advantage of a deep Bayesian nonparametric model is that the size of the model is able to grow with the data. Since the user activities can model as a stream, a deep Bayesian nonparametric model is suitable for modeling such stream data. However, the challenge of applying a deep Bayesian nonparametric model is that it is nontrivial to design an efficient formulation with a reasonable time complexity (Wang and Yeung, 2020).

- **Deep Reinforcement Learning based Insider Threat Detection.** Deep reinforcement learning is able to learn optimal policies for sophisticated agents in a complex environment (Arulkumaran et al., 2017). The advantage of deep reinforcement learning is that the policy consistently enhances its performance through reward signals. In the insider threat detection task, the insider detector can be considered as an agent in the deep reinforcement learning framework. With a properly designed reward function, the insider detector is able to keep improving the capability to identify insider attacks including the adaptive attacks. One challenge of applying the deep reinforcement learning for insider threat detection is that due to the complicated of malicious attacks, sometimes it is hard to design a good reward function. In such scenario, inverse reinforcement learning framework, whose goal is to identify a reward function automatically based on the behavior of insiders, can be further considered (Ng and Russell, 2000; Oh and Iyengar, 2019). Another challenge is that deep reinforcement learning usually requires large amounts of training data that are unavailable in the insider detection task. To tackle this challenge, other machine learning paradigms, such as meta-learning or imitation learning can be further combined with the deep reinforcement learning in practice (Duan et al., 2017; Wang et al., 2017a). Overall, although facing several challenges, deep reinforcement learning as a powerful framework has the opportunity to make a breakthrough in insider threat detection.
- **Interpretable Deep Learning for Insider Threat Detection.** Unlike some online anomaly detection tasks, such as bot detection on social media, which do not have an impact on real human beings, the insider threat detection is to identify malicious individuals, which is a high-stakes decision. Consequently, it is critical for insider threat detection models to have proper interpretability even the models can achieve superior performance. Hence, how to make prediction results understandable to human is key toward a trustworthy and reliable insider threat detection model. One recent work leverages the attention mechanism for interpretable log anomaly detection (Brown et al., 2018). Meanwhile, several interpretable sequence learning models are proposed in literature (Ming et al., 2019; Ribeiro et al., 2018). However, most of the existing studies focus on supervised training tasks, while for insider threat detection, it is usually infeasible to train a supervised model. Another advantage of developing explainable deep learning models is that such models have the potential to achieve fine-grained malicious activity detection. For example, if we consider the user activity sequence in a day as a data point and each activity in the sequence as a feature, the counterfactual explanation model (Molnar, 2019) which finds a similar data point by changing some of the features for which the predicted outcome changes in a relevant way, has the potential capability to identify malicious activities from an insider activity sequence.
- **Testbed Development.** To achieve insider threat detection, human actions within the monitored environment should be used as the analytical data. However, due to the privacy and confidentiality issues, the publicly available datasets in literature are very limited. Most of the recent work (Glasser and Lindauer, 2013; Lin et al., 2017; Liu et al., 2018b; Tuor et al., 2017) adopt the CERT dataset. However, as a synthetic dataset, the user activities in CERT dataset are not complicated enough. Consequently, developing a comprehensive testbed for insider threat detection evaluation is greatly needed.
- **Practical Evaluation Metrics.** Due to the extremely small number of insiders and the corresponding malicious activities, the commonly-used classification metrics, such as *accuracy, F1, ROC-AUC*, and *PR-AUC* are not sufficient for evaluating the performance of insider threat detection. It is an open question which metrics are more practical and whether some new metrics need to be developed. A recent study proposes a recall based metric, called *cumulative recall (CR-k)*, to evaluate the performance of algorithms on insider threat detection (Tuor et al., 2017). Cumulative recall assumes that there is a daily budget k to exam the top-k samples with the highest malicious scores derived from the algorithms. Then, the CR-k is defined as the sum of the recalls for all budgets up to k. For example, if we define $R(i)$ to be the recall with a budget of i, CR-k is calculated as $R(25) + R(50) + \ldots + R(k)$. CR-k can be considered as an approximation to an area under the recall curve. Because for insider threat detection tasks, the cost of a missed detection is substantially higher than the cost of a false positive, a recall based metric may be a suitable metric.

6. Conclusion

In this brief survey paper, we have reviewed various approaches in deep learning-based insider threat detection and categorized the existing approaches based on the adopted deep learning architectures. Although some progress has been achieved, the topic of using deep learning for insider threat detection is not well-exploited due to various challenges. We have discussed the challenges in this task and proposed several research directions that have the potential to advance insider threat detection based on deep learning techniques. Overall, deep learning for insider threat detection is an underexplored research topic. This survey could be extended and updated in the future as more advanced approaches are proposed.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by NSF grant 1564250.

REFERENCES

Akcay S., Atapour-Abarghouei A., Breckon T.P.. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. 2018. arXiv:1805.06725

Almehmadi A, El-Khatib K. On the possibility of insider threat detection using physiological signal monitoring. In: Proceedings of the 7th International Conference on Security of Information and Networks (SIN '14). Glasgow, Scotland, UK: Association for Computing Machinery; 2014. p. 223–30. doi:10.1145/2659651.2659654. ISBN 978-1-4503-3033-6

Alneyadi S, Sithirasenan E, Muthukkumarasamy V. A survey on data leakage prevention systems. J. Netw. Comput. Appl. 2016;62(C):137–52.

Ameri S, Fard MJ, Chinnam RB, Reddy CK. Survival analysis based framework for early prediction of student dropouts. In: CIKM; 2016. p. 903–12.

Arandjelovic R, Zisserman A. Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. p. 609–17.

Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep Reinforcement Learning: A Brief Survey. IEEE Signal Process. Mag. 2017;34(6):26–38. doi:10.1109/MSP.2017.2743240. 2017, ISSN: 1558-0792

Asano YM, Rupprecht C, Vedaldi A. In: ICLR. A critical analysis of self-supervision, or what we can learn from a single image; 2019.

Bengio Y. Deep learning of representations: looking forward. In: Dediu AH, Martín-Vide C, Mitkov R, Truthe B, editors. In: Statistical Language and Speech Processing. Springer Berlin Heidelberg; 2013. p. 1–37. Number 7978 in Lecture Notes in Computer Science, 978-3-642-39592-5 978-3-642-39593-2

Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. Neural Netw. IEEE Trans. 1994;5(2):157–66. doi:10.1109/72.279181. 1994

Brown A, Tuor A, Hutchinson B, Nichols N. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In: MLCS'18. New York, NY, USA: Association for Computing Machinery; 2018. p. 8. doi:10.1145/3217871.3217872. 9781450358651.

Chalapathy R., Borzeshi E.Z., Piccardi M.. Bidirectional LSTM-CRF for Clinical Concept Extraction. arXiv:161005858 2016a.

Chalapathy R., Borzeshi E.Z., Piccardi M.. An Investigation of Recurrent Neural Architectures for Drug Name Recognition. arXiv:160907585 2016b.

Chalapathy R., Chawla S.. Deep Learning for Anomaly Detection: A Survey. arXiv:190103407 2019.

Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. Assoc. Comput. Mach. 2009;41(3):1–58.

Chapfuwa P, Tao C, Li C, Page C, Goldstein B, Carin L, Henao R. In: ICML. Adversarial time-to-event modeling; 2018 arXiv:1804.03184.

Chen H., Perozzi B., Al-Rfou R., Skiena S.. A tutorial on network embeddings. 2018; arXiv:1808.02590.

Cheng Y, Huang F, Zhou L, Jin C, Zhang Y, Zhang T. A hierarchical multimodal attention-based neural network for image captioning. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17). New York, NY, USA: ACM; 2017. p. 889–92. doi:10.1145/3077136.3080671. 978-1-4503-5022-8.

Chung J, Gulcehre C, Cho K, Bengio Y. In: ICML. Gated feedback recurrent neural networks; 2015 arXiv:1502.02367.

Cohen W.W.. Enron email dataset. 2009.

Costa D, Albrethsen M, Collins M, Perl S, Silowash G, Spooner D. In: Technical Report CMU/SEI-2016-TR-007. An insider threat indicator ontology. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA; 2016.

CSOCERT Division of SRI-CMU, and ForcePoint. In: Technical Report. 2018 U.S. State of Cybercrime; 2018.

Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications: an overview. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013.

Du N, Dai H, Trivedi R, Upadhyay U, Gomez-Rodriguez M, Song L. Recurrent marked temporal point processes: embedding event history to vector. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). New York, NY, USA: ACM; 2016. p. 1555–64. doi:10.1145/2939672.2939875. 978-1-4503-4232-2.

Duan Y, Andrychowicz M, Stadie B, Ho OJ, Schneider J, Sutskever I, Abbeel P, Zaremba W, Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R. One-shot imitation learning. In: NIPS. Curran Associates, Inc.; 2017. p. 1087–98.

Eldardiry H, Bart E, Liu J, Hanley J, Price B, Brdiczka O. Multi-domain information fusion for insider threat detection. In: 2013 IEEE Security and Privacy Workshops; 2013. p. 45–51. doi:10.1109/SPW.2013.14.

Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. 2006;28(4):594–611. doi:10.1109/TPAMI.2006.79. ISSN1939-3539.

Gamachchi A., Sun L., Boztas S.. A Graph Based Framework for Malicious Insider Threat Detection. arXiv:180900141 2018.

Gavai G, Sricharan K, Gunning D, Hanley J, Singhal M, Rolleston R. Supervised and unsupervised methods to detect insider threat from enterprise social and online activity data. J. Wirel. Mob Netw. Ubiquitous Comput. Dependable Appl. 2015;6(4):47–63.

Glasser J, Lindauer B. Bridging the gap: a pragmatic approach to generating insider threat data. In: 2013 IEEE Security and Privacy Workshops; 2013. p. 98–104. doi:10.1109/SPW.2013.37.

Goyal P, Hu Z, Liang X, Wang C, Xing E. In: ICCV. Nonparametric variational auto-encoders for hierarchical representation learning; 2017 arXiv:1703.07027.

Graves A.. Generating Sequences With Recurrent Neural Networks. arXiv:13080850 2013).

Greenberg S.. Using unix: Collected traces of 168 users. 1988.

Greitzer FL, Kangas LJ, Noonan CF, Brown CR, Ferryman T. Psychosocial modeling of insider threat risk based on behavioral and word use analysis. e-Service J. 2013;9(1):106–38. doi:10.2979/eservicej.9.1.106.

Harilal A, Toffalini F, Castellanos J, Guarnizo J, Homoliak I, Ochoa M. Twos: a dataset of malicious insider threat behavior based on a gamified competition. In: Proceedings of the 2017 International Workshop on Managing Insider Security Threats; 2017. p. 45–56.

Hashem Y, Takabi H, GhasemiGol M, Dantu R. Towards insider threat detection using psychophysiological signals. In: Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats (MIST '15). Denver, Colorado, USA: Association for Computing Machinery; 2015. p. 71–4. 978-1-4503-3824-0.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: CVPR; 2016. p. 770–8. arXiv:1512.03385.

Hendrycks D., Mazeika M., Dietterich T.G.. Deep Anomaly Detection with Outlier Exposure. arXiv:1812046062018;

Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–54. doi:10.1162/neco.2006.18.7.1527. ISSN0899-7667.

Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.

Homoliak I, Toffalini F, Guarnizo J, Elovici Y, Ochoa M. Insight into insiders and IT: a survey of insider threat taxonomies. Anal. Model. Countermeas. 2019;52(2):40.

Hu T, Niu W, Zhang X, Liu X, Lu J, Liu Y. An insider threat detection approach based on mouse dynamics and deep learning. Secur. Commun. Netw. 2019;2019:3898951. doi:10.1155/2019/3898951. ISSN1939-0114.

Huang C, Zhang C, Zhao J, Wu X, Yin D, Chawla N. MiST: a multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting. In: The World Wide Web Conference (WWW '19). San Francisco, CA, USA: ACM; 2019. p. 717–28. doi:10.1145/3308558.3313730. 978-1-4503-6674-8.

Jiang J, Chen J, Gu T, Choo KKR, Liu C, Yu M, Huang W, Mohapatra P. Anomaly detection with graph convolutional networks for insider threat and fraud detection. In: MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM); 2019. p. 109–14. doi:10.1109/MILCOM47813.2019.9020760. ISSN2155-7586.

Jing H, Smola AJ. Neural survival recommender. In: CIKM; 2017. p. 515–24. doi:10.1145/3018661.3018719. 978-1-4503-4675-7.

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med. Res. Methodol. 2018;18(1):24. ISSN1471-2288.

Krizhevsky A, Sutskever I, Hinton GE, Pereira F, Burges CJC, Bottou L, Weinberger KQ. In: NIPS. ImageNet classification with deep convolutional neural networks; 2012.

Le DC, Zincir-Heywood AN. Evaluating insider threat detection workflow using supervised and unsupervised learning. In: 2018 IEEE Security and Privacy Workshops (SPW); 2018. p. 270–5.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44. ISSN0028-0836.

Li S, Xiao S, Zhu S, Du N, Xie Y, Song L, Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R. Learning temporal point processes via reinforcement learning. In: NIPS. Curran Associates, Inc.; 2018. p. 10804–14.

Lin L, Zhong S, Jia C, Chen K. Insider threat detection based on deep belief network feature representation. In: 2017 International Conference on Green Informatics (ICGI); 2017. p. 54–9. doi:10.1109/ICGI.2017.37.

Liu B, Li Y, Sun Z, Ghosh S, Ng K. In: AAAI. Early prediction of diabetes complications from electronic health records: a multi-task survival analysis approach; 2018a.

Liu F, Wen Y, Zhang D, Jiang X, Xing X, Meng D. Log2vec: A heterogeneous graph embedding based approach for detecting cyber threats within enterprise. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (London, United Kingdom) (CCS '19). New York, NY, USA: Association for Computing Machinery; 2019. p. 1777–94. 9781450367479.

Liu L, De Vel O, Chen C, Zhang J, Xiang Y. Anomaly-based insider threat detection using deep autoencoders. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW); 2018b. p. 39–48. doi:10.1109/ICDMW.2018.00014. ISSN2375-9259.

Liu L, De Vel O, Han Q, Zhang J, Xiang Y. Detecting and preventing cyber insider threats: a survey. IEEE Commun. Surv. Tutor. 2018c;20(2):1397–417. ISSN1553-877X.

Liu Q., Kusner M.J., Blunsom P.. A Survey on Contextual Embeddings. arXiv:2003072782020.

Lu J, Wong RK. Insider threat detection with long short-term memory. In: Proceedings of the Australasian Computer Science Week Multiconference (ACSW 2019). Sydney, NSW, Australia: Association for Computing Machinery; 2019. p. 1–10. doi:10.1145/3290688.3290692. 978-1-4503-6603-8.

Luck M., Sylvain T., Cardinal H., Lodi A., Bengio Y.. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. arXiv:17051024520l7;

Marcus Tan YX, Iacovazzi A, Homoliak I, Elovici Y, Binder A. Adversarial attacks on remote user authentication using behavioural mouse dynamics. In: 2019 International Joint Conference on Neural Networks (IJCNN); 2019. p. 1–10. doi:10.1109/IJCNN.2019.8852414.

Maxion RA. Masquerade detection using enriched command lines. In: 2003 International Conference on Dependable Systems and Networks, 2003. Proceedings. IEEE; 2003. p. 5–14.

Ming Y, Xu P, Qu H, Ren L. Interpretable and steerable sequence learning via prototypes. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Anchorage, AK, USA) (KDD '19). New York, NY, USA: Association for Computing Machinery; 2019. p. 903–13. doi:10.1145/3292500.3330908. 9781450362016.

Molnar Christoph. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book; 2019.

Moriano P, Pendleton J, Rich S, Camp LJ. Insider threat event detection in user-system interactions. In: Proceedings of the 2017 International Workshop on Managing Insider Security Threats (MIST '17). Dallas, Texas, USA: Association for Computing Machinery; 2017. p. 1–12. 978-1-4503-5177-5.

Nalisnick E, Smyth P. In: ICLR. Stick-breaking variational autoencoders; 2017 arXiv:1605.06197.

Ng AY, Russell SJ. Algorithms for inverse reinforcement learning. In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2000. p. 663–70. 978-1-55860-707-1.

Oh Mh, Iyengar G. Sequential anomaly detection using inverse reinforcement learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Anchorage, AK, USA: Association for Computing Machinery; 2019. p. 1480–90. 978-1-4503-6201-6.

Paisley J, Carin L, Lu RW. Machine Learning with Dirichlet and Beta Process Priors : Theory and Applications. Duke University; 2010. Ph.D. Dissertation.

Pfleeger SL, Predd JB, Hunker J, Bulford C. Insiders behaving badly: addressing bad actors and their actions. IEEE Trans. Inf. Forensics Secur. 2010;5(1):169–79. doi:10.1109/TIFS.2009.2039591.

Rashid T, Agrafiotis I, Nurse JRC. A new take on detecting insider threats: exploring the use of hidden Markov models. In: Proceedings of the 8th ACM CCS International Workshop on Managing Insider Security Threats (MIST '16). Vienna, Austria: ACM; 2016. p. 47–56. 978-1-4503-4571-2.

Rasmussen J.G.. Lecture Notes: Temporal Point Processes and the Conditional Intensity Function. arXiv:1806002212018.

Ribeiro MT, Singh S, Guestrin C. Anchors: high-precision model-agnostic explanations, volume 18; 2018. p. 1527–35.

Salakhutdinov R, Hinton G. Deep Boltzmann Machines. Proc. Int. Conf. Artif. Intell. Stat. 2009;5(2):448–56.

Salem MB, Hershkop S, Stolfo SJ, Stolfo SJ, Bellovin SM, Keromytis AD, Hershkop S, Smith SW, Sinclair S. A survey of insider attack detection research. In: Insider Attack and Cyber Security: Beyond the Hacker. Boston, MA: Springer US; 2008. p. 69–90. 978-0-387-77322-3.

Salem MB, Stolfo SJ. Modeling user search behavior for masquerade detection. In: International Workshop on Recent Advances in Intrusion Detection. Springer; 2011. p. 181–200.

Sanzgiri A, Dasgupta D. Classification of insider threat detection techniques. In: Proceedings of the 11th Annual Cyber and Information Security Research Conference (CISRC '16). New York, NY, USA: ACM; 2016. p. 25:1–25:4. doi:10.1145/2897795.2897799. 978-1-4503-3752-6.

Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans. Neural Netw.

2009;20(1):61–80. doi:10.1109/TNN.2008.2005605. ISSN1941-0093.

Schonlau M, DuMouchel W, Ju WH, Karr AF, Theus M, Vardi Y. Computer intrusion: detecting masquerades. Stat. Sci. 2001;16(1):58–74.

Shabtai A, Elovici Y, Rokach L. A Survey of Data Leakage Detection and Prevention Solutions. Springer Science & Business Media; 2012.

Song H, Jiang Z, Men A, Yang B. A hybrid semi-supervised anomaly detection model for high-dimensional data. Comput. Intell. Neurosci. 2017;2017:8501683. ISSN1687-5265.

Su Y, Zhao Y, Niu C, Liu R, Sun W, Pei D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Anchorage, AK, USA: Association for Computing Machinery; 2019. p. 2828–37. 978-1-4503-6201-6.

Tuor A, Kaplan S, Hutchinson B, Nichols N, Robinson S. In: AI for Cyber Security Workshop at AAAI 2017. Deep learning for unsupervised insider threat detection in structured cybersecurity data streams; 2017 arXiv:1710.00811.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. In: NIPS. Attention is all you need; 2017 arXiv:1706.03762.

Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning - ICML '08. New York, New York, USA: ACM Press; 2008. p. 1096–103. doi:10.1145/1390156.1390294. 978-1-60558-205-4.

Wang P, Li Y, Reddy C K. Machine learning for survival analysis: a survey. ACM Computing Surveys (CSUR) 2019;51(6):1–36.

Wang H, Yeung DY. A survey on bayesian deep learning. ACM Comput. Surv. 2020;53(5):37. doi:10.1145/3409383. ISSN0360-0300.

Wang J.X., Kurth-Nelson Z., Tirumala D., Soyer H., Leibo J.Z., Munos R., Blundell C., Kumaran D., Botvinick M.. Learning to reinforcement learn. arXiv:1611057632017a.

Wang W, Zheng VW, Yu H, Miao C. A survey of zero-shot learning: settings, methods, and applications. ACM Trans. Intell. Syst. Technol. 2019a;10(2):13:1–13:37. doi:10.1145/3293318. ISSN2157-6904.

Wang X, Jabri A, Efros AA. In: CVPR. Learning correspondence from the cycle-consistency of time; 2019b arXiv:1903.07593.

Wang Y., Yao Q., Kwok J., Ni L.M.. Generalizing from a Few Examples: A Survey on Few-Shot Learning. arXiv:1904050462020;

Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P.S.. A Comprehensive Survey on Graph Neural Networks. arXiv:1901005962019.

Xiao S, Xu H, Yan J, Farajtabar M, Yang X, Song L, Zha H. Learning conditional generative models for temporal point processes. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

Yuan F, Cao Y, Shang Y, Liu Y, Tan J, Fang B, Shi Y, Fu H, Tian Y, Krzhizhanovskaya VV, Lees MH, Dongarra J, Sloot PMA. Insider threat detection with deep neural network. In: Computational Science – ICCS 2018 (Lecture Notes in Computer Science). Cham: Springer International Publishing; 2018. p. 43–54. 978-3-319-93698-7.

Yuan S, Zheng P, Wu X, Li Q. Insider threat detection via hierarchical neural temporal point processes. In: 2019 IEEE International Conference on Big Data (Big Data); 2019. p. 1343–50.

Zhang A, Paisley J. Deep Bayesian nonparametric tracking. In: International Conference on Machine Learning; 2018. p. 5833–41.

Zhang D, Zheng Y, Wen Y, Xu Y, Wang J, Yu Y, Meng D. Role-based log analysis applying deep learning for insider threat detection. In: Proceedings of the 1st Workshop on Security-Oriented Designs of Computer Architectures and Processors (SecArch'18). Toronto, Canada: Association for Computing Machinery; 2018a. p. 18–20. doi:10.1145/3267494.3267495. ISBN: 978-1-4503-5991-7.

Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Shyu ML, Chen SC, Iyengar SS. A survey on deep learning: algorithms, techniques, and applications. ACM Computing Survey 2018;51(5):1–36.

Zhang L., Wang S., Liu B.. Deep Learning for Sentiment Analysis : A Survey. arXiv:1801078832018b.

Zhang Z., Cui P., Zhu W.. Deep Learning on Graphs: A Survey. arXiv:1812042022018c;.

Zheng P, Yuan S, Wu X, Li J, Lu A. One-class adversarial nets for fraud detection. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019; 2019. p. 1286–93.

Shuhan Yuan received the BE degree in network engineering in 2009 and the ME degree in computer engineering in 2012 from Huaqiao University, China, and the Ph.D. degree in computer science from Tongji University, China, in 2017. He is an assistant professor in the Computer Science Department, Utah State University. His major research interests include data mining, deep learning and its applications on anomaly detection.

Xintao Wu received the BS degree in information science from the University of Science and Technology of China, in 1994, the ME degree in computer engineering from the Chinese Academy of Space Technology, in 1997, and the Ph.D. degree in information technology from George Mason University, in 2001. He is a professor in the Department of Computer Science and Computer Engineering, University of Arkansas. His major research interests include data mining and knowledge discovery, data privacy and security, and fairness aware learning.