Intelligent Communication, Computing and Energy Lab.

Deep learning for insider threat detection Review, challenges and opportunities

서울과학기술대학교
컴퓨터공학과
Intelligent Communication, Computing, and Energy lab
Heejae Park

Contents

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.

- Compared to the external attacks whose footprints are difficult to hide, the attacks from insiders are hard to detect.

- Malicious insiders already have the authorized power to access the internal information systems.

1. Introduction

- The traditional shallow machine learning models are unable to make full use of the user behavior data.

High-dimensionality, Complexity, Heterogeneity, and Sparsity.

- Deep learning can be used as a powerful tool to analyze the user behavior in an organization to identify the potential malicious activities from insiders.

1. Introduction

- Main Contribution of the paper

1. This is the first survey about using deep learning techniques to tackle the challenges of insider threat detection.
2. This paper summarizes ten existing challenges based on the characteristics of insiders and insider threats.
3. This paper points out ten future directions to improve the performance of deep learning models for insider threat detection.

2.1 Deep learning

- Compared with traditional machine learning models, deep learning models are able to learn semantic representations from the raw data with minimal human efforts.

- 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.
  2. Convolutional neural network (CNN), which leverages the convolutional and pooling layers.
  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.

2.2 Deep learning for anomaly detection

- Anomaly detection is to identify instances that are dissimilar to others.

- A recent survey categorizes the deep learning-based anomaly detection into three groups based on the availability of labels: [Deep Learning for Anomaly Detection: A Survey]
  1. When both normal and anomalous data are available - Supervised
  2. When many normal samples are available while only a small number of anomalous samples is available - Semi-supervised
  3. When no labeled data are available - Unsupervised

- However, for anomaly detection, it is very difficult, if not impossible, to collect a large number of labeled anomalies in the training data.

3.1 Insiders and insider threats

- Taxonomy of insiders
- Insider usually indicates "a person with legitimate access to an organization's computers and networks"
- In general, there are three types of insiders: traitors, masqueraders, and unintentional perpetrators
  1. Traitors are insiders who misuse their privileges to commit malicious activities for financial or personal gains.
  2. Masqueraders are insiders who conduct illegal actions on behalf of legitimate employees of an institute.
  3. Unintentional perpetrators are insiders who unintentionally make mistakes and expose confidential information to outsiders.

3.1 Insiders and insider threats

- Taxonomy of insider threats
- Insider threats indicate the "threats with malicious intent directed toward organizations"
- 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
    1. IT sabotage indicates directly using IT to make harm to an organization, which is usually conducted by insiders.
    2. 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.
    3. Fraud indicates unauthorized modification, addition, or deletion of data.

Financial Gain

3.2 Datasets

- Datasets are critical for research on insider threat detection.
- However, there is no comprehensive real-world dataset publicly available for insider threat detection.

- Simulation result (11 datasets)

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

3.2 Datasets

- Most of the recent deep learning-based studies adopt the CMU CERT (Computer Emergency Response Time) dataset to evaluate their proposed approaches.
- CERT dataset consists of five log files that record the computer-based activities for all employees in a simulated organization.

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)

3.2 Datasets

- There are several versions of datasets according to when the datasets were created.
- The most widely-used versions are r4.2 and r6.2.

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

- r4.2 is a "dense" dataset that contains many insiders and malicious activities.
- r6.2 is a "sparse" dataset that contains 5 insiders and 3995 normal users.

3.2 Datasets

- Specifically, the CERT in r6.2 dataset simulates the following five scenarios of attacks from insiders.

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

- User ACM2278 use a removable drive, and upload data to wikileaks.org.
- User CDE1846 logs into another user's machine and searches for interesting files.
- User CMP2946 begins surfing job websites and soliciting employment from a competitor and uses a thumb drive to steal data.
- User MBG3183, who decimated by layoffs, uploads documents to Dropbox.
- User PLJ1771 downloads a keylogger by thumb drive and uses the collected keylogs to log in as his supervisor.

3.3 Why deep learning for insider threat detection?

- Potential advantages of deep learning for insider threat detection.
  1. Representation Learning - User behavior in cyberspace is complicated and non-linear. By using deep non-linear model, it is natural to use deep learning models to capture complex user behavior.
  2. Sequence Modeling - Since we can represent the user activities recorded in audit data as sequential data, leveraging RNN can boost the performance of insider threat detection.
  3. Heterogeneous Data Composition - Combining all the useful data for insider threat detection is expected to achieve better performance than only using a single type of data. Deep learning models are more powerful to combine the heterogeneous data for detection.

3.4 Deep learning for insider threat deteciton

- Due to the extremely unbalanced nature of the dataset, most of the proposed approaches adopt the unsupervised learning model.
- Most of the papers focus on detecting malicious subsequence (e.g., activities in 24 h) or malicious session.

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			✓

3.4.1 Deep feedforward neural network (FNN)

- FNN is a classical type of deep learning model.
- Liu et al. [2018 ICDMW]
  - ✓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.
- Insider threats should have relatively high reconstruction errors.

Anomalous Score

- The idea of using deep autoencoder for anomaly detection is intuitive.
- Cannot capture the temporal information.

3.4.2 Recurrent neural network (RNN)

- RNN is mainly used for modeling the sequential data, which maintains a hidden state with a self-loop connection.
- Lu and Wong. [2019 ACSW]
  - ✓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.

- Capture the temporal information of user activity sequences.
- Could face high false alert if users change the daily pattern instead of conducting malicious activities.

3.4.3 Convolutional neural network (CNN)

- A typical CNN structure consists of a convolutional layer followed by a pooling layer and a fully connected layer for prediction.
- Hu et al. [2019 *Security and communication networks*]
  - ✓Proposed CNN-based user authentication method by analyzing mouse biobehavioral characteristics.
  - ✓If an ID theft attack occurs, the user mouse behaviors will be inconsistent with the legal user.

CNN Model

- 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.

3.4.4 Graph neural network (GNN)

- GNN is able to model the relationships between nodes.
- A widely used GNN model is a graph convolutional network (GCN) that uses graph convolutional layer to extract node information.

- Jiang et al. [2019 MILCOM]
  - ✓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.

Graph Sturcture

- 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.

3.4.4 Graph neural network (GNN)

Intelligent Communication, Computing and Energy Lab.

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
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

4. Challenges

- Extremely Unbalanced Data
- Compared with the benign activities, the malicious activities from insiders are extremely rare in real-world scenarios.
- 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.

- Temporal Information in Attacks
- Existing approaches 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.
- The temporal information plays an important role.

4. Challenges

- Heterogeneous Data Fusion
- Leveraging various data sources and fusing such heterogeneous data are also critical to improve the insider threat detection.
- For example, considering the user profile (i.e., psychometric score) or user interaction data could help to identify potential insider threats.

- Subtle Attacks
- In reality, we cannot expect insiders have a significant pattern change to conduct malicious activities.
- Insider threats are subtle and hard to notice, which means that insiders and benign users are close in the feature space.

4. Challenges

- Adaptive Threats
- Learning-based models are unable to detect new types of attacks after training.
- It is inefficient to train the models
  - ✓Need time to collect enough samples
  - ✓Cannot ensure in-time detection and prevention
- Designing a model that can adaptively improve the performance is critical.

- Fine-grained Detection
- Users usually conduct a large number of activities in a session.

Hard to achieve in time detection

- How to identify the fine-grained malicious subsequence or the exact malicious activity is important.

4. Challenges

- Early Detection
- Current approaches focus on insider threat detection, which means malicious activities already occur and the significant loss is already caused to organizations.

Early Detection is needed

- Interpretability
- Deep learning models are usually considered as black boxes.
- It is critical to understand the reason why the model makes such predictions since employees are usually the most valuable asset.

Reduce misclassify
Provide insight of the model

4. Challenges

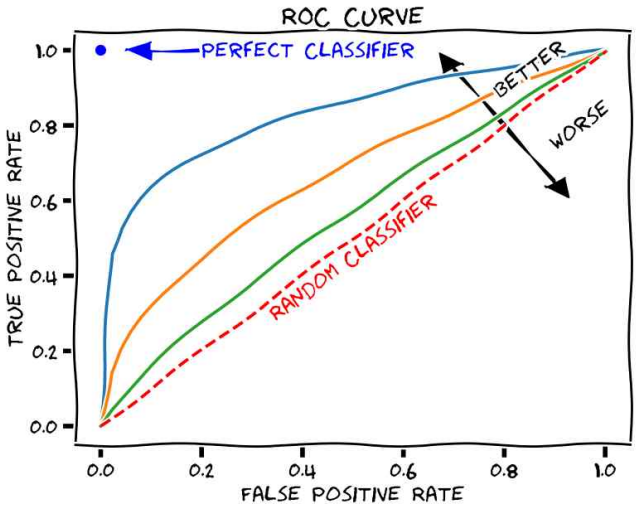
- Lack of Testbed
- There is no real-world dataset & CERT dataset has gap between the synthetic data and real-world scenario.

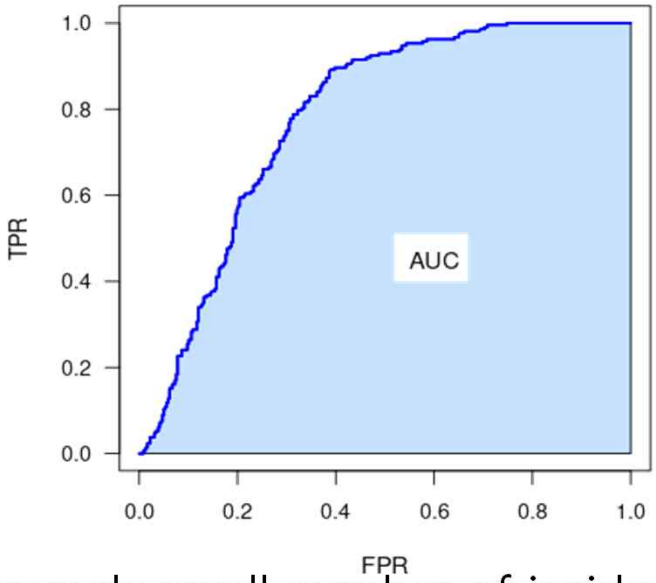
- Since the CERT dataset is a synthetic dataset, most of the activities are randomly generated with limited complexity.
  - ✓No daily routine pattern in this dataset.
- The latest version of CERT dataset only consists of five scenarios.
  - ✓Datasets are too narrow

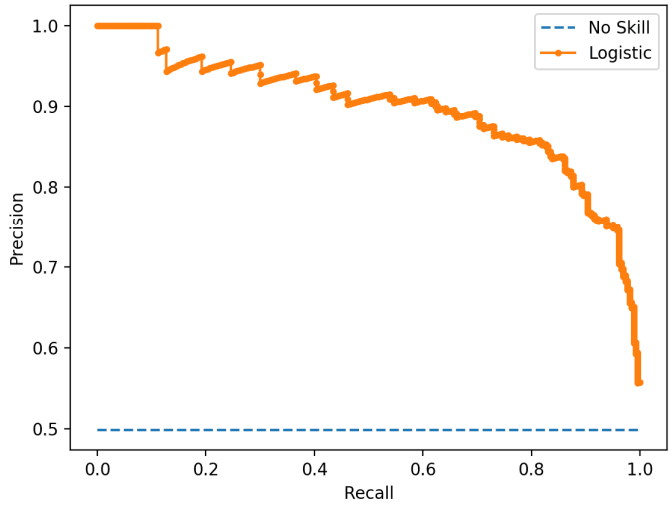
Poor performance

4. Challenges

- Lack of Practical Evaluation Metrics
- The commonly-used classification metrics, such as true positive rate (TPR), false positive rate (FPR), precision, and recall, can be adopted to evaluate performance.







- However, due to the extremely small number of insiders and the corresponding malicious activities.

Unclear that above metrics are practical

5. Future direction

- Few-shot Learning based Insider Threat Detection
- Aims at classifying samples from unknown classes given only a few labeled samples.
- It can extend to where only one or totally no labeled sample is available.

Fit to insider threat detection

- Wang et al. [2020 TIST]
  1. Data based approaches which augment training data by prior knowledge
  2. Model based approaches which use the prior knowledge to constrain hypothesis space
  3. Algorithm based approaches which alter search strategy in hypothesis space by prior knowledge
- Achieve insider threat detection with limited data
- Hard to detect new type of attack that is significantly different from the observed ones

5. Future direction

- Self-supervised Learning based Insider Threat Detection
- 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.

- The task we use to pretrain the deep learning model is called "pretext task".
- 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.

Easily detect the subtle insider threat

- Achieve insider threat detection without using any labeled information
- Require hand-crafted rules that are tailored to each dataset

5. Future direction

- Deep Marked Temporal Point Process based Insider Threat Detection
- Marked temporal point process is a powerful mathematical tool to model the observed random event.
- 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.
- Du et al. [2016 SIN]
  - Adopted RNN with marked temporal point process.

- Potential to improve the performance of insider threat detection by combining the user activity types and time information.

- Capture the temporal information in terms of time
- Require a large amount of samples for training which is infeasible for insider threat detection.

5. Future direction

- 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 (ex. Psychological study, email communication).
- 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.

- Capture the user information from multiple perspective
- Hard to obtain multi modality data

5. Future direction

- 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.
- 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 (malicious activity) occurs.

Early Alerts

- Achieve the insider threat early detection
- Require a number of event samples

5. Future direction

Intelligent Communication, Computing and Energy Lab.

- Deep Bayesian Nonparametric Model for Fine-grained Insider Threat Detection
- One potential solution is to make user data as an activity stream and apply a clustering algorithm to identify the potential malicious activities.
- Bayesian nonparametric models are often used for data clustering and able to generate unbounded clusters.

Suitable to model complicated user behavior

- Capture fine-grained user activity patterns
- High time complexity

5. Future direction

- Deep Reinforcement Learning based Insider Threat Detection
- Deep reinforcement learning is able to learn optimal policies for sophisticated agents in a complex environment.
- In the insider threat detection task, the insider detector can be considered as an agent in the deep reinforcement learning framework.

- Keep improving the capability to identify insider threats via the reward function
- Due to the complicated of malicious attacks, hard to design a proper reward function
- Requires large amounts of training data

Combine with other ML/DL models

5. Future direction

- Interpretable Deep Learning for Insider Threat Detection
- How to make prediction results understandable to human is key.

Enhance the performance

- Most of the existing studies focus on supervised training tasks, while for insider threat detection, it is usually infeasible to train a supervised model.

- Potential to achieve fine-grained malicious activity detection.

5. Future direction

- 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 adopt the CERT dataset.

- Consequently, developing a comprehensive testbed for insider threat detection evaluation is greatly needed.

5. Future direction

- Practical Evaluation Metrics
- Commonly-used classification metrics, such as accuracy, F1, ROC-AUC, and PR-AUC are not sufficient.
- It is an open question.

- Tuor et al. [2017 AAAI]
  - ✓A recent study proposes cumulative recall (CR-k), to evaluate the performance of algorithms.
  - ✓Cumulative recall assumes that there is a daily budget k to exam the top k samples.
  - ✓For example, if we define R(k) to be the recall with a budget of k, CR-k is calculated as R(25) + R(50) + . . . + R(k).
  - ✓CR-k can be considered as an approximation to an area under the recall curve.

5. Future direction

Intelligent Communication, Computing and Energy Lab.

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

6. Conclusion

1. Reviewed various approaches in deep learning-based insider threat detection and categorized the existing approaches based on the adopted deep learning architectures.
2. Discussed the challenges and proposed several research directions that have the potential to advance insider threat detection based on deep learning techniques.
3. Deep learning for insider threat detection is an underexplored research topic. Therefore, it can be extended.